Appendix

In this appendix, we provide extended discussions and additional results that complement the main paper. Sec. A presents examples of the images generated by the Intrinsic Simulation Module. Sec. B shows examples of the intrinsic aware text descriptions generated by the Intrinsic Encoder. Sec. C summarizes the hyperparameter settings and implementation details. Sec. D reports the results for multi dataset training. Sec. E provides detailed results on the KITTI dataset. Sec. F provides detailed results on the Waymo dataset. Sec. G provides detailed results on the nuScenes dataset. Sec. H includes additional ablation studies. Sec. I provides further qualitative visualizations. Sec. K discusses the limitations of MonoIA.

A. Generated Images

We first present the synthetic images produced by the Intrinsic Simulation Module. Given an original image and its associated ground truth focal length, the module renders a new view that corresponds to a target focal length. This is achieved by adjusting the field of view according to the target focal and then rescaling the transformed image back to the original resolution to maintain a consistent input size for the detector. This process allows us to systematically vary the intrinsic parameters while preserving the scene content, enabling controlled studies of intrinsic sensitivity and robustness. Representative examples of the rendered images are shown in Fig. A1, where changes in effective perspective, object scale, and scene geometry become visually evident as the focal length varies.

B. Generated Intrinsic Texts

In the Intrinsic Encoder, we take the simulated images as visual references and use them to guide the generation of intrinsic aware text descriptions. Specifically, for each pair of original and focal length transformed images, we provide both views to a large language model and ask it to articulate the visual effects introduced by the intrinsic change. The model describes how the modified focal length alters object scale, perceived depth, foreground background separation, and overall scene perspective. These descriptions are phrased in natural language and capture the perceptual consequences of intrinsic variation rather than numeric changes alone. They serve as semantically rich prompts that enable the detector to associate visual cues with their underlying intrinsic causes. We provide representative examples of the generated text prompts in Fig. A2, which illustrate how the language model explains perspective changes, scale distortion, and depth variation induced by different focal configurations.

Item	Value
optimizer	AdamW
learning rate	2e-4
weight decay	1e-4
number of feature scales	4
hidden dim	256
nheads	8
number of encoder layers	3
number of decoder layers	3
encoder npoints	4
decoder npoints	4
number of group	11
α in class loss	0.25
class loss weight	2
bbox loss weight	5
GIoU loss weight	2
3D center loss weight	10
dim loss weight	1
depth loss weight	1
depth map loss weight	1
scheduler	Step
decay rate	0.5
decay list	[85,125,165,205]
dropout	0.1
number of queries	50
feedforward dim	256
class cost weight	2
bbox cost weight	5
GIoU cost weight	2
3D center cost weight	10

Table A1. **Main hyperparameters of MonoIA.**

C. Implementation details

MonoIA is built upon the MonoDGP [56] and MonoCoP [89] frameworks. For each intrinsic configuration, we use ChatGPT 4o [51] to generate a collection of 24 diverse and semantically informative prompts that describe the intrinsic properties in natural language. These descriptions are encoded using the CLIP ViT H/14 [78] text encoder, which offers strong image text alignment and provides a stable foundation for intrinsic aware representation learning. We summarize training hyperparameters in Tab. A1. We adopt the AdamW optimizer with a learning rate of 2×10^{-4} and a weight decay of 10^{-4}. The model is trained for 250 epochs with a batch size of 16.

D. Multi-dataset Training Results

Due to its intrinsic awareness, MonoIA naturally supports multi-dataset training, a setting where images originate from cameras with vastly different intrinsic parameters such as

Figure A1. **Examples of synthetic images.**

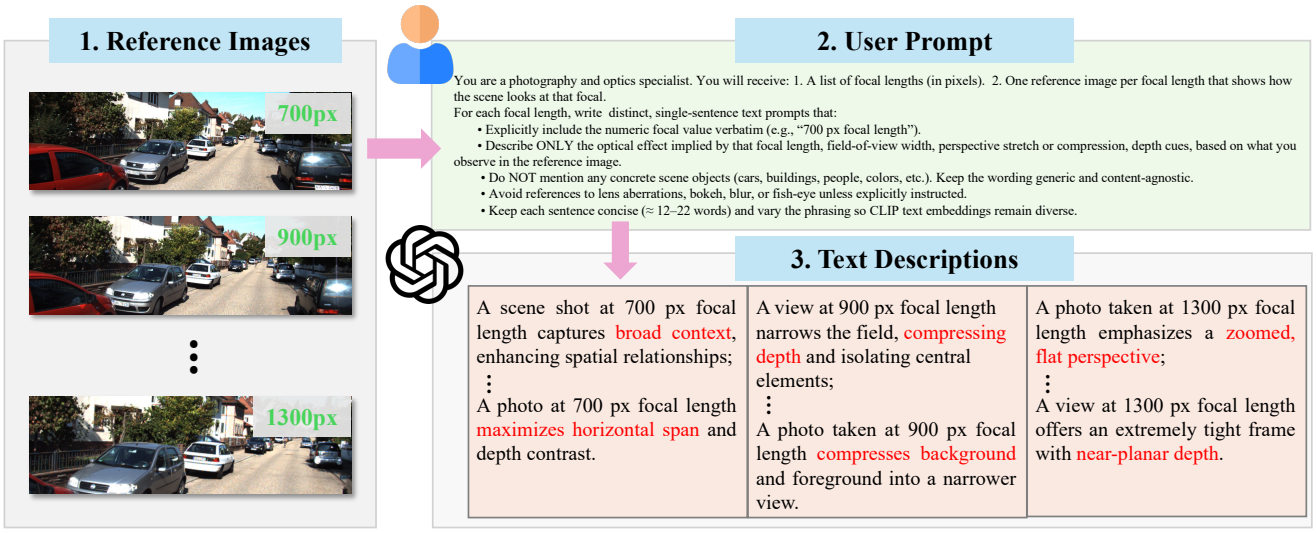


Figure A2. **Overiew of Generation of Intrinsic Texts.**

focal lengths, sensor sizes, and principal point offsets. Conventional monocular 3D detectors struggle in this regime because they implicitly assume a fixed projection geometry. When trained on heterogeneous datasets, their learned depth–feature relationships become inconsistent, leading to degraded or unstable performance. In contrast, MonoIA explicitly conditions the detection pipeline on intrinsic representations, allowing it to correctly interpret geometric variations across datasets and maintain consistent depth reasoning.

To examine this capability in depth, we evaluate several multi-dataset training strategies that vary category coverage and input resolutions. As shown in Tab. A2, MonoIA exhibits consistent improvements across all configurations, demonstrating its ability to absorb complementary information from different datasets while aligning their geometric discrepancies through intrinsic-aware modeling. Specifically, when trained on all categories using both KITTI and nuScenes, the KITTI Val performance improves from 24.40% to 26.54%, while the nuScenes performance increases from 7.09% to 8.19%. This indicates that the model not only benefits from the additional visual diversity but also properly handles the large intrinsic gap between these two datasets. Furthermore, when Waymo is included in the joint training set, the KITTI performance further rises from 26.54% to 28.91%. Notably, this gain persists despite the fact that Waymo has yet another distinct imaging pipeline and scene distribution, which typically destabilizes conventional monocular 3D detectors.

These results together reveal two key findings. First, multi-dataset training with MonoIA yields a model that surpasses all individually trained models, suggesting that intrinsic-aware design enables effective consolidation of knowledge from different domains. Second, the unified model is not only more accurate but also more robust, highlighting that intrinsic conditioning allows MonoIA to generalize across datasets, these experiments confirm that intrinsic-aware modeling provides a principled mechanism for exploiting visual diversity while preserving coherent depth representations, which is crucial for building scalable monocular 3D detectors in real-world multi-camera environments.

E. Detailed KITTI Results

While Tab. 3 in the main paper provides a simplified overview due to space constraints, we include a more comprehensive version here. Tab. A3 presents detailed comparisons of monocular 3D object detection methods on the KITTI Val set under the challenging IoU ≥ 0.7 setting. We report both AP_{3D} and AP_{BEV} across Easy, Moderate, and Hard difficulty levels. MonoIA achieves SoTA performance

Method	Training Cates	Resolution	Training Datasets KITTI	nuScenes	Waymo	AP_{3D}^{KIT}	AP_{3D}^{NU}	AP_{3D}^{Way}
MonoDETR [84]	Car	1280 × 384	✓			20.61		
MonoDGP [56]			✓			22.49		
MonoCoP [89]	ALL	1280 × 384	✓			23.64		
MonoIA			✓			24.40		
MonoIA	ALL	896 × 512		✓			7.09	
MonoIA	ALL	768 × 512			✓			8.94
MonoIA	ALL	1280 × 384	✓	✓		25.09	8.19	
			✓	✓	✓	26.28	10.45	7.77
MonoIA	ALL	1280 × 512	✓	✓		26.54	9.81	
	ALL		✓	✓	✓	28.91	11.48	10.19
	Car		✓	✓	✓	29.31	12.70	11.84

Table A2. **Detailed results on Multi-Dataset Training.** [Key: KIT: KITTI, NU: nuScenes, Way: Waymo]

Method	Extra Data	Venue	AP_{3D} (↑) Easy	Mod.	Hard	AP_{BEV} (↑) Easy	Mod.	Hard
OccupancyM3D [54]	LiDAR	CVPR 24	26.87	19.96	17.15	35.72	26.60	23.68
OPA-3D [66]	Depth	ICRA 23	24.97	19.40	16.59	33.80	25.51	22.13
MonoFlex [85]	None	CVPR 21	23.64	17.51	14.83	–	–	–
GUP Net [46]		CVPR 21	22.76	16.46	13.72	31.07	22.94	19.75
DEVIANT [31]		ECCV 22	24.63	16.54	14.52	32.60	23.04	19.99
MonoCon [75]		AAAI 22	26.33	19.01	15.98	–	–	–
MonoUNI [25]		NeurIPS 23	24.51	17.18	14.01	–	–	–
MonoDETR [84]		ICCV 23	28.84	20.61	16.38	37.86	26.95	22.80
MonoCD [79]		CVPR 24	26.45	19.37	16.38	34.60	24.96	21.51
FD3D [74]		AAAI 24	28.22	20.23	17.04	36.98	26.77	23.16
MonoMAE [26]		NeurIPS 24	30.29	20.90	17.61	40.26	27.08	23.14
MonoDGP [56]		CVPR 25	30.76	22.34	19.02	39.40	28.20	24.42
MonoCoP [89]		CVPR 26	<u>32.06</u>	<u>23.98</u>	<u>20.64</u>	<u>42.20</u>	<u>31.29</u>	<u>27.58</u>
MonoIA (Ours)		CVPR 26	**33.61**	**24.40**	**20.80**	**44.69**	**32.17**	**27.93**

Table A3. **KITTI Val results at $IoU_{3D} \geq 0.7$.** MonoIA achieves SoTA performance across all metrics. [Key: **First**, Second]

across all metrics and difficulty levels, surpassing prior methods that either rely on external signals (*e.g.*, depth or LiDAR) or use strong supervision across all categories. Notably, our method maintains its superiority under both Car-only and All-category training settings. This highlights the robustness and generalizability of the proposed intrinsic-aware design, especially under data-scarce monocular settings. Additionally, while many competing methods incorporate external depth or LiDAR signals, MonoIA achieves superior results with image-only input, demonstrating its efficiency and practicality. Interestingly, we observe that the most significant performance gain occurs at the Easy level. This aligns with the fact that enriching the dataset with diverse intrinsics increases the number of Easy objects, further validating the effectiveness of our Intrinsic Awareness in handling varying camera intrinsics.

F. Detailed Waymo Results

Table A4 presents a comprehensive comparison on the Waymo Val set under varying IoU thresholds (0.5 and 0.7) and difficulty levels (Level 1 and Level 2). MonoIA consistently outperforms all baselines across both AP_{3D} and APH_{3D}, achieving top or second-best performance in every metric. Notably, our method demonstrates strong generalization across distances, particularly excelling in the mid-range $(0, 30\text{m})$ and long-range $(30, 50\text{m})$ detection, where prior approaches typically degrade. These results highlight the

IoU_{3D}	Difficulty	Method	APH_{3D} [%](↑) All	0-30	30-50	50-∞	AP_{3D} [%](↑) All	0-30	30-50	50-∞
0.7	Level 1	GUP Net [46] in [31]	2.27	6.11	0.80	0.03	2.28	6.15	0.81	0.03
		DEVIANT [31]	2.67	6.90	0.98	0.02	2.69	6.95	0.99	0.02
		MonoDETR [84] in [89]	2.10	5.94	0.73	0.12	2.11	5.99	0.73	0.12
		MonoDGP [56] in [89]	2.39	6.62	0.84	0.12	2.41	6.67	0.84	0.12
		MonoCoP [89]	2.70	7.38	1.06	**0.16**	2.72	7.44	1.07	**0.16**
		MonoIA (Ours)	**3.05**	**8.43**	**1.11**	0.13	**3.07**	**8.50**	**1.12**	0.13
	Level 2	GUP Net [46] in [31]	2.12	6.08	0.77	0.02	2.14	6.13	0.78	0.02
		DEVIANT [31]	2.50	6.87	0.94	0.02	2.52	6.93	0.95	0.02
		MonoDETR [84] in [89]	1.97	5.92	0.70	0.10	1.98	5.96	0.71	0.10
		MonoDGP [56] in [89]	2.24	6.59	0.81	0.10	2.26	6.65	0.81	0.10
		MonoCoP [89]	2.53	7.35	1.02	**0.14**	2.55	7.41	1.03	**0.14**
		MonoIA (Ours)	**2.86**	**8.40**	**1.07**	0.12	**2.88**	**8.47**	**1.08**	0.12
0.5	Level 1	GUP Net [46] in [31]	9.94	24.59	4.78	0.22	10.02	24.78	4.84	0.22
		DEVIANT [31]	10.89	26.64	5.08	0.18	10.98	26.85	5.13	0.18
		MonoDETR [84] in [89]	9.60	23.58	4.67	0.99	9.68	23.78	4.72	1.00
		MonoDGP [56] in [89]	9.84	23.73	5.01	0.98	10.06	24.01	5.06	0.99
		MonoCoP [89]	11.65	27.35	5.97	**1.46**	11.76	27.59	6.03	**1.48**
		MonoIA (Ours)	**12.44**	**29.52**	**6.51**	1.18	**12.54**	**29.75**	**6.57**	1.19
	Level 2	GUP Net [46] in [31]	9.31	24.50	4.62	0.19	9.39	24.69	4.67	0.19
		DEVIANT [31]	10.20	26.54	4.90	0.16	10.29	26.75	4.95	0.16
		MonoDETR [84] in [89]	9.00	23.49	4.51	0.86	9.08	23.70	4.55	0.87
		MonoDGP [56] in [89]	9.32	23.65	4.84	0.85	9.43	23.92	4.88	0.86
		MonoCoP [89]	10.93	27.25	5.76	**1.27**	11.03	27.49	5.82	**1.29**
		MonoIA (Ours)	**11.66**	**29.40**	**6.29**	1.03	**11.76**	**29.64**	**6.34**	1.04

Table A4. **Waymo Val Vehicle results.** MonoIA outperforms all methods on most metrics across both difficulty (Level 1 and Level 2) and IoU threshold (0.5 and 0.7). [Key: **First**, Second]

effectiveness of our intrinsic-aware design in enabling robust and accurate monocular 3D object detection across challenging real-world scenarios.

G. Detailed nuScenes Results

Tab. A5 summarizes our detection results on the nuScenes Val split. Compared to existing methods, MonoIA delivers strong improvements across both 3D detection (AP_{3D}) and BEV detection (AP_{BEV}) under multiple IoU thresholds. These gains are consistent across all difficulty levels, indicating that intrinsic-aware modeling provides more stable geometric reasoning in complex urban driving scenes.

Specifically, under the more challenging IoU ≥ 0.7 setting, MonoIA surpasses the strongest prior baseline, MonoCoP, by $+1.03\%$ on AP_{3D} (Moderate) and $+1.47\%$ on AP_{BEV} (Moderate). Notably, nuScenes includes diverse camera intrinsics and varying viewpoints, making high-IoU improvements particularly meaningful. Even under the relatively easier IoU ≥ 0.5 condition, MonoIA continues to outperform all baselines, achieving the highest scores across every category and metric. These results demonstrate that MonoIA not only improves precise 3D localization but also enhances BEV spatial alignment, confirming its robustness in large-scale, multi-camera environments.

H. More Ablations

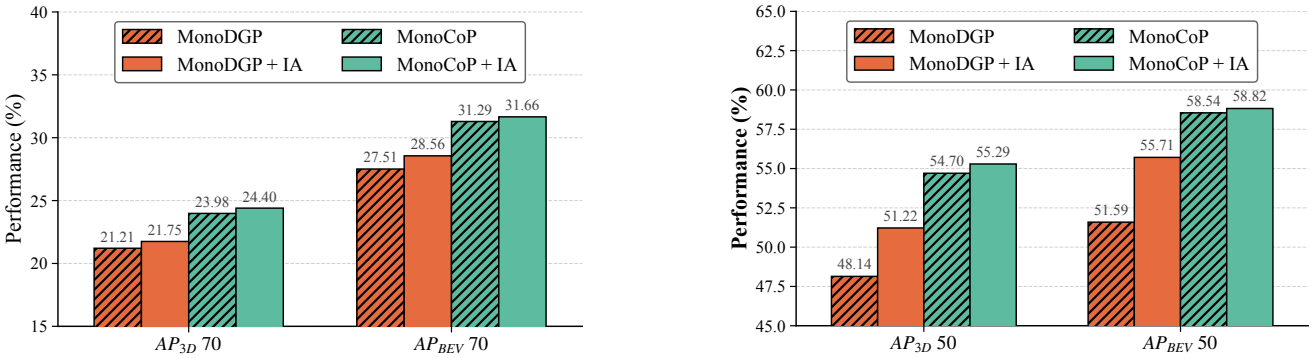
We provide additional ablation studies to further assess the effectiveness of the proposed MonoIA.

H.1. Support Multi baselines

In the main paper, we primarily evaluate MonoIA on the MonoCoP [89] framework. To further assess the generalizability and plug-and-play nature of our Intrinsic Awareness (IA) module, we additionally integrate it into MonoDGP [56], a recent monocular 3D detector accepted to CVPR 2025. As shown in Fig. A3, IA consistently improves both AP_{3D} and AP_{BEV} across the two architectures. For example, MonoDGP improves from 48.14% to 51.22% on $AP_{3D}50$ and from 51.59% to 55.71% on $AP_{BEV}50$. Similarly, MonoCoP improves from 54.70% to 55.29% on $AP_{3D}50$ and

Method	$IoU_{3D} \geq 0.7$				$IoU_{3D} \geq 0.5$			
	AP_{3D}		AP_{BEV}		AP_{3D}		AP_{BEV}	
	Easy	Mod.	Easy	Mod.	Easy	Mod.	Easy	Mod.
DEVIANT [31]	9.69	8.33	16.28	14.36	31.47	28.22	35.61	31.93
MonoDETR [84]	9.53	8.19	16.39	14.41	31.81	28.35	35.70	31.96
MonoDGP [56]	10.04	8.78	16.55	14.53	29.56	26.17	32.67	29.44
MonoCoP [89]	<u>10.85</u>	<u>9.71</u>	<u>17.83</u>	<u>15.86</u>	<u>33.70</u>	<u>29.91</u>	<u>37.44</u>	<u>34.01</u>
MonoIA (Ours)	**12.33**	**10.74**	**19.56**	**17.33**	**35.01**	**31.07**	**38.57**	**34.13**

Table A5. **nuScenes Val Results.** MonoIA achieves SoTA performance on 3D detection and BEV detection. [Key: **First**, <u>Second</u>]



(a) AP_{3D} 70. performance.

(b) AP_{3D} 50 performance.

Figure A3. Generalizability of Intrinsic Awareness (IA) on **different baseline methods**.

from 58.54% to 58.82% on AP_{BEV}50 after adopting IA.

These consistent gains demonstrate two key properties of IA. First, IA enhances performance across detectors with very different designs. Second, the simultaneous improvements in both depth-sensitive metrics (AP_{3D}) and BEV spatial metrics (AP_{BEV}) indicate that IA improves not only depth estimation but also the geometric coherence of the predicted 3D layout. Together, these results show that IA is a model-agnostic, easily pluggable module that can reliably strengthen a wide range of monocular 3D detection frameworks.

H.2. Support Multi-backbones.

Furthermore, we evaluate the robustness of MonoIA across various image backbones, including ResNet-18, ResNet-34, ResNet-50, and ResNet-101. As shown in Fig. A4, MonoIA consistently surpasses both MonoCoP and MonoDGP across all backbones and difficulty levels. Notably, with a lightweight ResNet-18, MonoIA achieves a significant gain of +2.13% in AP_{3D} (0.7) Moderate over MonoDGP. With deeper backbones like ResNet-50 and ResNet-101, MonoIA maintains top performance, achieving 24.40% and 23.92% in Moderate settings, respectively. Interestingly, the slightly lower performance with ResNet-101 suggests that deeper networks do not always yield better results in Mono3D tasks. These consistent improvements demonstrate that our intrinsic-awareness design not only enhances performance but also generalizes effectively across architectures.

H.3. Number of Intrinsic Texts

We further study how the number of LLM generated intrinsic texts affects MonoIA. In our design, these texts are encoded using the CLIP Text Encoder and then aggregated through average pooling to form the intrinsic text embedding. As shown in Tab. A6, increasing the number of intrinsic texts consistently improves performance, indicating that richer textual descriptions provide more stable intrinsic representations. We select 24 texts as our default configuration, as it yields the best overall performance.

H.4. Number of Intrinsic Tokens

We further evaluate the impact of the learned intrinsic tokens by training MonoIA under different settings of intrinsics. As shown in Tab. A7, increasing the number of intrinsic tokens initially improves performance. The model achieves peak performance when using four tokens, corresponding to a balanced representation of intrinsic variations. However, introducing more tokens beyond this point leads to a performance drop, due to the over-fragmentation of the intrinsic space.

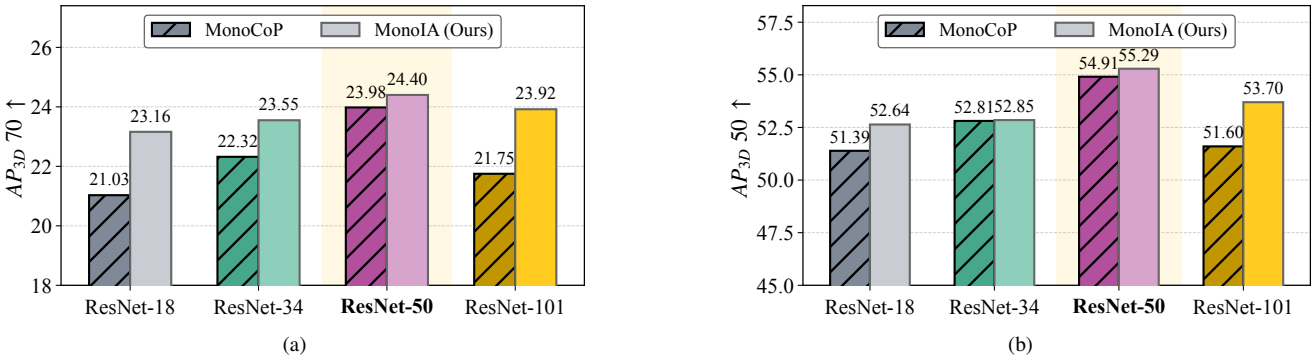


Figure A4. Generalizability of Intrinsic Awareness (IA) on **different image backbones**.

Number of Intrinsic Texts	AP_{3D} 70 (%) (↑) Easy	Mod	Hard
0	29.80	22.16	17.76
1	32.53	23.69	20.09
12	**33.96**	24.17	20.55
24	33.61	**24.40**	20.80
36	33.64	23.41	**20.83**

Table A6. **Impact of Number of Intrinsic Texts.**

Learned Tokens	AP_{3D} 70 (%) (↑) Easy	Mod	Hard
[700]	31.78	23.26	19.82
[700, 900]	32.32	23.54	20.14
[700, 900, 1100]	33.68	24.13	20.72
[700, 900, 1100, 1300]	**33.61**	**24.40**	**20.80**
[700, 900, 1100, 1300, 1500]	32.36	23.41	19.93

Table A7. **Impact of Number of Intrinsic Embeddings.**

I. Visualization

In this section, we present visualizations of detection results on KITTI (Fig. A5), nuScenes (Fig. A6), and Waymo (Fig. A7). Predictions by the baseline method MonoCoP are highlighted in orange, while those by our MonoIA are highlighted in green.

J. Why mostly focusing on focal length variation.

We mainly focus on focal length for the following three reasons:

1) Geometric motivation. Focal length is the dominant intrinsic component in Mono3D, as it directly controls the depth–scale mapping in monocular projection, while principal point shifts mainly induce image-plane translations. Such translation effects can be largely compensated by modern CNN- or Transformer-based detectors and therefore have a much smaller impact on Mono3D performance.

2) Experimental evidence. To support our choice, we separately evaluate principal point and focal length shifts on KITTI. a 200-pixel principal point shift causes only marginal performance degradation, while an equivalent focal length shift leads to a much larger drop, indicating the dominant impact of focal length in Mono3D.

3) Cross-dataset practice. Across common Mono3D benchmarks, the principal point is typically close to the image center and varies little, whereas intrinsic differences across datasets are mainly reflected in focal length. Therefore, focal length dominates intrinsic variation in practice.

K. MonoIA Limitation and Future Work

MonoIA achieves intrinsic awareness by learning dedicated intrinsic embeddings. While it demonstrates strong generalization to unseen intrinsics, it is not an intrinsic-invariant network. Future work could explore intrinsic-invariant architectures that can naturally handle diverse camera settings without relying on explicit embedding learning. Moreover, recent advances in multimodal learning and recognition systems have shown strong capabilities in visual reasoning and multimodal fusion [11, 13, 14, 18, 67, 91–93]. However, these models are not designed for 3D perception tasks. Bridging vision-language models with 3D object detection, especially in terms of geometry-aware reasoning and spatial understanding, remains an important direction for future research.

Figure A5. **Qualitative results on KITTI.** [Key: MonoIA, MonoCoP]

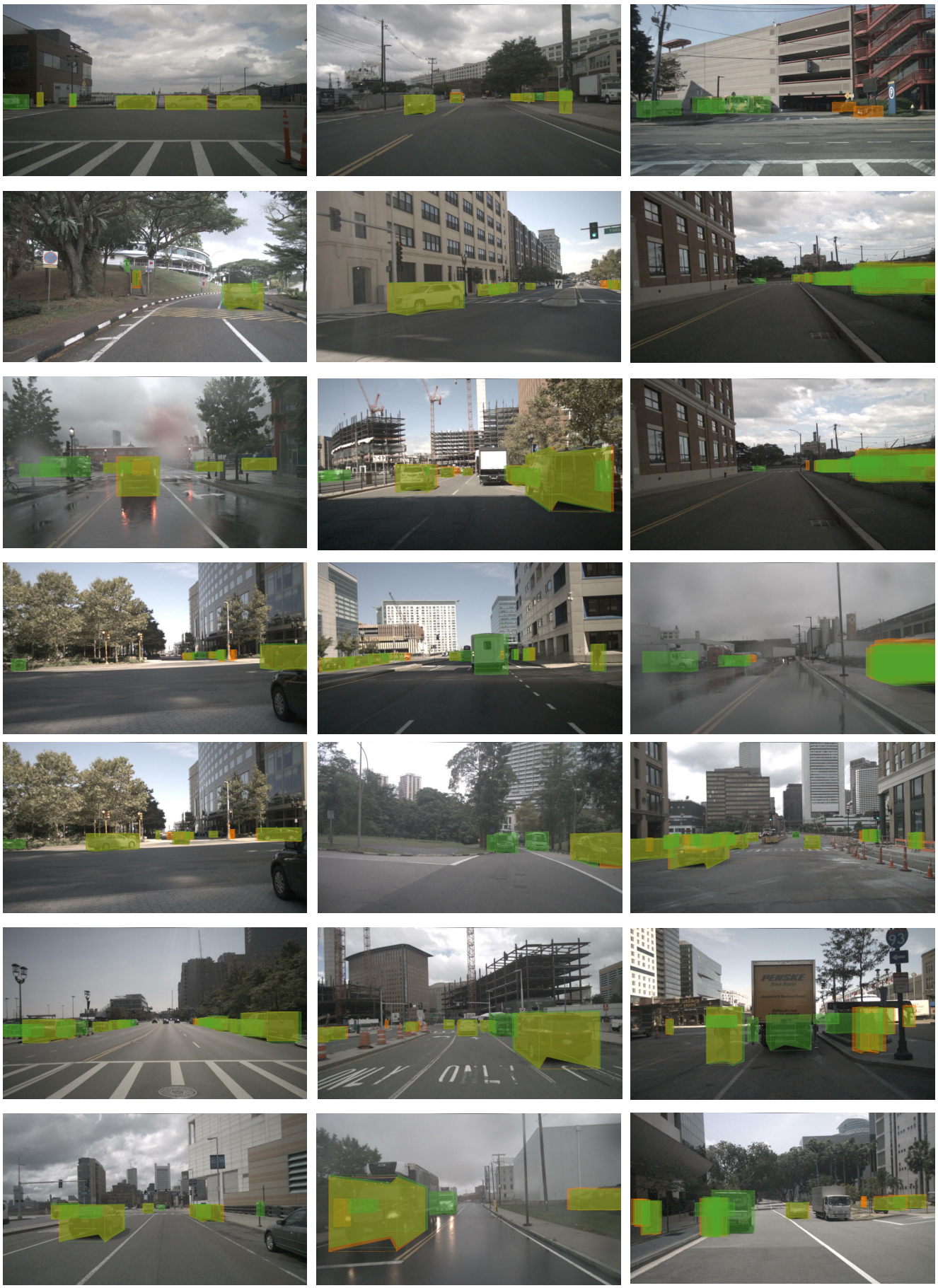


Figure A6. **Qualitative results on nuScenes.**[Key: MonoIA, MonoCoP]

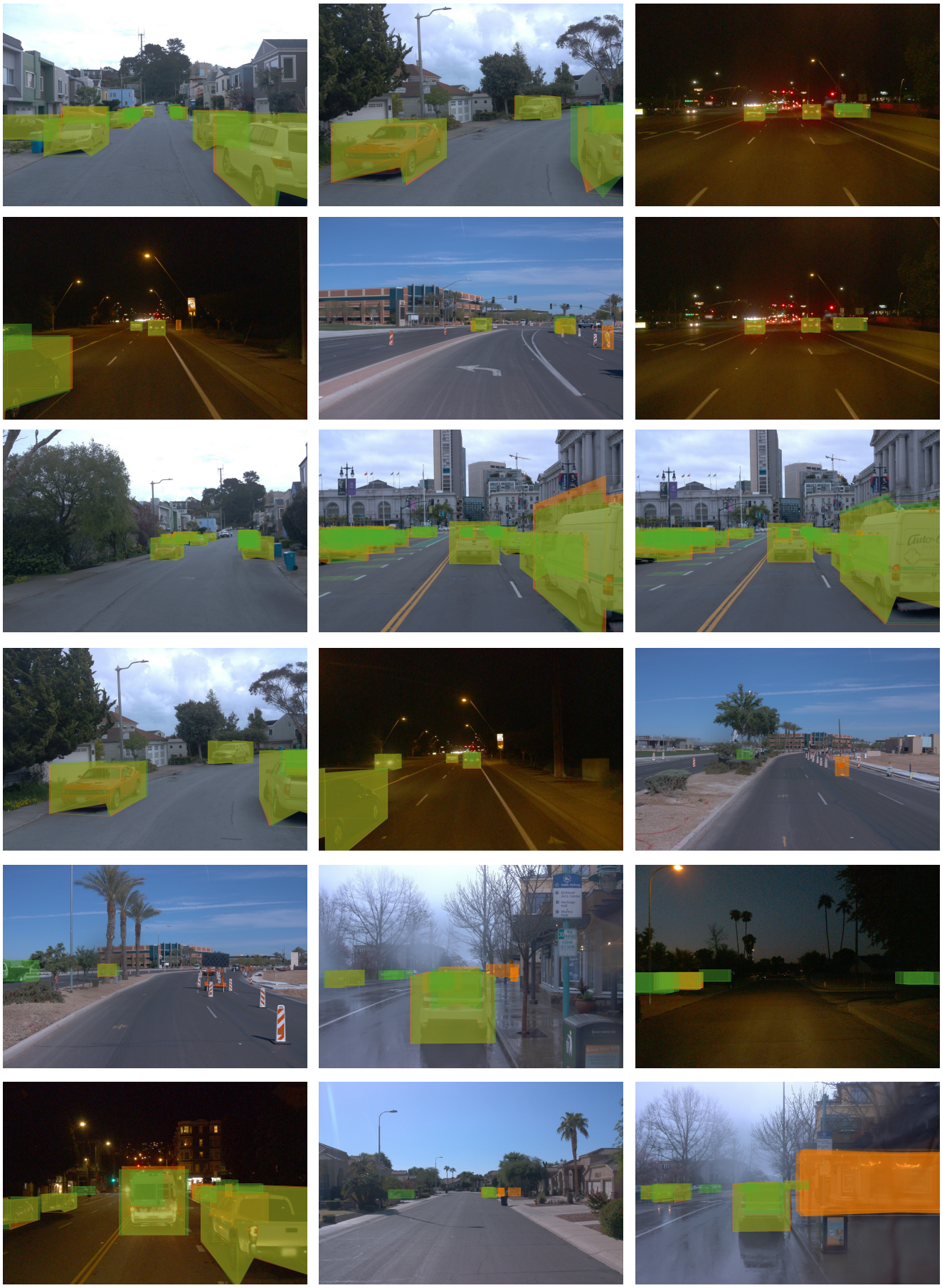


Figure A7. **Qualitative results on Waymo.** [Key: MonoIA, MonoCoP]