Towards Intrinsic-Aware Monocular 3D Object Detection

Zhihao Zhang[1] Abhinav Kumar[1] Xiaoming Liu[1,2]
[1]Michigan State University [2]University of North Carolina at Chapel Hill
zhan2365@msu.edu abhinav3663@gmail.com liuxm@cs.unc.edu

Abstract

Monocular 3D object detection (Mono3D) aims to infer object locations and dimensions in 3D space from a single RGB image. Despite recent progress, existing methods remain highly sensitive to camera intrinsics and struggle to generalize across diverse settings, since intrinsic governs how 3D scenes are projected onto the image plane. We propose MonoIA, a unified intrinsic-aware framework that models and adapts to intrinsic variation through a language-grounded representation. The key insight is that intrinsic variation is not a numeric difference but a perceptual transformation that alters apparent scale, perspective, and spatial geometry. To capture this effect, MonoIA employs large language models and vision–language models to generate intrinsic embeddings that encode the visual and geometric implications of camera parameters. These embeddings are hierarchically integrated into the detection network via an Intrinsic Adaptation Module, allowing the model to modulate its feature representations according to camera-specific configurations and maintain consistent 3D detection across intrinsics. This shifts intrinsic modeling from numeric conditioning to semantic representation, enabling robust and unified perception across cameras. Extensive experiments show that MonoIA achieves new state-of-the-art results on standard benchmarks including KITTI, Waymo, and nuScenes (e.g., $+1.18\%$ *on the KITTI leaderboard), and further improves performance under multi-dataset training* (e.g., $+4.46\%$ *on KITTI Val). Code and models are publicly available at https://github.com/alanzhangcs/MonoIA.*

1. Introduction

Monocular 3D object detection (Mono3D) aims to estimate the 3D locations and dimensions of objects from a single RGB image, offering a cost-effective alternative to LiDAR-based approaches [54, 62, 81]. Due to its low hardware requirements, Mono3D has attracted increasing attention in autonomous driving [65] and robotics [48].

In real-world scenarios, cameras exhibit diverse intrinsic parameters, making robustness to intrinsic variation essential for practical deployment. However, existing state-of-the-art (SoTA) detectors [5, 31, 56, 79, 89] typically assume fixed intrinsics during both training and inference, which limits their generalization. For instance, Fig. 1 shows that MonoDGP [84], MonoDGP [56] and MonoCoP [89] perform well when evaluated under the *seen* intrinsic, but exhibit significant degradation under unseen intrinsics. As a result, models trained under one intrinsic often fail to generalize to unseen configurations, and adapting to new cameras usually requires full retraining.

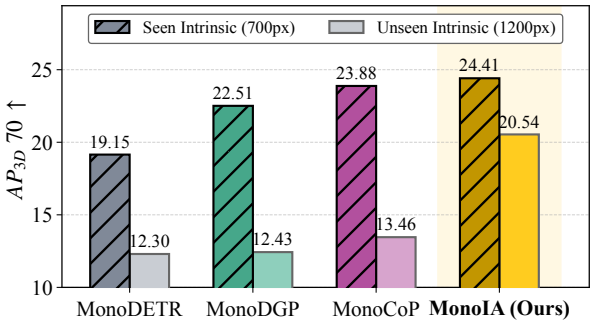


Figure 1. **MonoIA enables intrinsic awareness in Mono3D on KITTI Val.** Existing Mono3D detectors [56, 84, 89] lack intrinsic awareness and thus generalize poorly to images with unseen intrinsics. In contrast, our intrinsic-aware MonoIA achieves superior performance under seen intrinsics and demonstrates strong generalization to the unseen one.

We first expose detectors to a broader range of focal lengths by varying the field of view (FoV) of input images. However, empirical results show that such diversity alone yields only marginal gains. This finding reveals that the key challenge lies not in data diversity but in how detectors represent intrinsic cues. As illustrated in Fig. 2, changes in focal length reshape how a 3D scene is projected onto the image plane: the same object appears larger and the background more compressed under a longer focal length, even though its 3D position remains unchanged. Such variations alter apparent scale, perspective, and spatial geometry, which are fundamental to reliable Mono3D. Yet, conventional detectors [5, 56, 89] treat intrinsics as raw numeric inputs, forcing the network to infer their perceptual effects from limited supervision. Consequently, models tend to either ignore

Figure 2. **Impact of intrinsic variation on image appearance. Left:** The two images show the same object in the same 3D position but captured with different intrinsics. As the focal length increases, the object appears larger and the FoV is smaller. **Right:** Schematic illustration of how intrinsic variations affect object appearance.

intrinsic cues or overfit to a few discrete training values, resulting in poor generalization to unseen configurations.

To address this gap, we propose MonoIA, a unified intrinsic-aware framework that explicitly models and integrates intrinsic information throughout the detection process. MonoIA introduces two key components. The *Intrinsic Encoder* transforms numeric intrinsics into *language-grounded* representations. For each intrinsic configuration, a large language model (LLM) [51] generates textual descriptions that capture its perceptual and geometric effects, such as changes in field of view, perspective distortion, and depth compression. These descriptions are then encoded using a CLIP Text Encoder [69] to form semantically structured intrinsic embeddings. Unlike raw numbers, these embeddings capture how intrinsic variations manifest visually, yielding a perceptually continuous and geometrically organized representation space. This language-grounded encoding provides a strong inductive bias for intrinsic-aware feature learning and supports robust generalization to unseen focal lengths.

While the Intrinsic Encoder captures the perceptual meaning of intrinsics, the resulting embeddings remain external to the detection process. To fully leverage this knowledge, we introduce an *Intrinsic Adaptation Module* that consists of a lightweight Connector and a hierarchical fusion mechanism. The fixed intrinsic embeddings are preserved to maintain their semantic consistency, while the Connector maps them into a learnable latent space for interaction with visual features. Through hierarchical fusion, these adapted intrinsic features are integrated at multiple network stages, allowing intrinsic cues to guide both low-level representation learning and high-level object reasoning. This design ensures that the semantic understanding of camera intrinsics becomes an integral part of the detection pipeline.

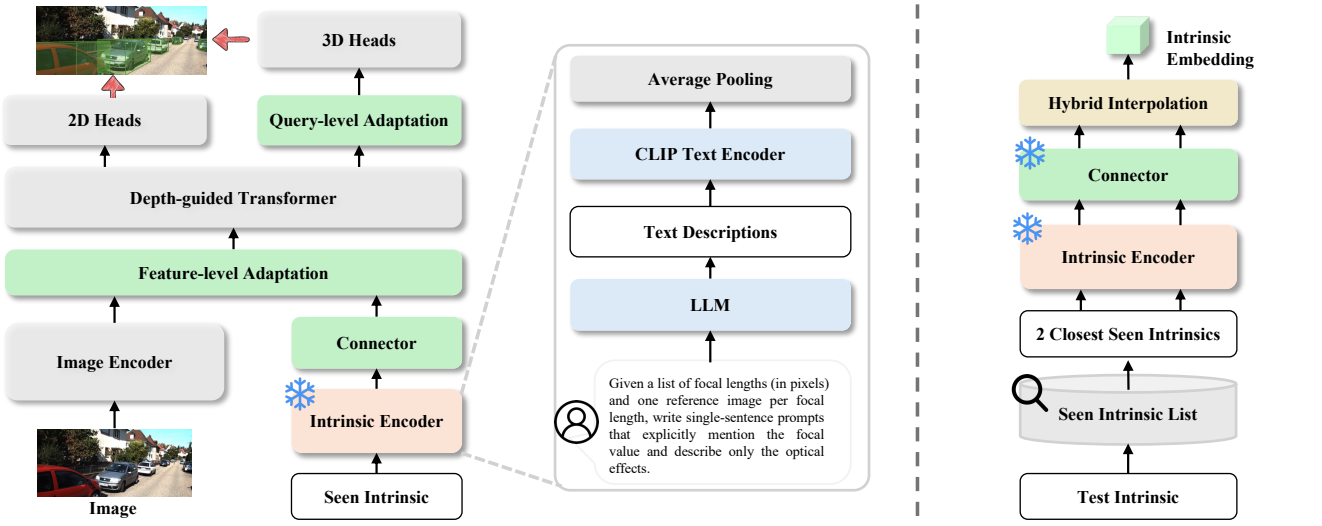
Overall, MonoIA shifts intrinsic modeling from numeric conditioning to semantic representation. This design brings three key advantages. It improves zero shot generalization to unseen focal lengths, enables natural compatibility with multi dataset training, and delivers stronger performance on standard 3D benchmarks, providing a unified intrinsic aware solution for robust Mono3D.

In summary, our main contributions are as follows:

- We reveal existing Mono3D methods are highly sensitive to intrinsic variations and generalize poorly to unseen intrinsics.
- We identify that intrinsic variation is not a simple numeric difference but a *perceptual transformation* that alters apparent scale, perspective, and spatial geometry, redefining how 3D scenes are visually perceived.
- We introduce **MonoIA**, a unified intrinsic-aware framework that first transforms numeric intrinsics into *language-grounded representations* capturing their perceptual and geometric effects, and then integrates them *hierarchically* into the detector for intrinsic-aware feature learning.
- Extensive experiments across multiple benchmarks demonstrate that MonoIA achieves (1) superior zero-shot generalization to unseen focal lengths, (2) natural compatibility with multi-dataset training, and (3) significant accuracy gains under standard 3D settings, validating its effectiveness and broad generalization capability.

2. Related Work

Mono3D. Monocular 3D object detection (Mono3D) relies solely on a single RGB image as input, posing significant challenges due to the inherent ambiguity in recovering depth from 2D projections. Early methods addressed this task using hand-crafted features [53], but recent advances predominantly leverage deep neural networks [3, 4]. A broad spectrum of techniques has been explored to enhance performance, including architectural improvements [22, 76], equivariant representations [10, 31], loss function design [2, 12], uncertainty modeling [29, 46], and explicit depth estimation [33, 50, 56, 74, 79, 85]. Several works incorporate additional signals during training, such as non-maximum suppression (NMS) [30, 41, 94], corrected extrinsics [90], CAD models [9, 34, 43], or even LiDAR supervision [23, 44, 45, 59]. Others propose innovations like pseudo-LiDAR representations [23, 47, 71], diffusion-based generation [38, 58], or BEV (bird's-eye view) encoding [27, 37, 86]. Transformer-based approaches [87] have also gained traction [7], with modifications including positional encoding [21, 63, 70], learned queries [15, 24, 36, 83], and query denoising [39]. Additional techniques include knowledge distillation [28, 40, 72, 80], stereo input [35, 73], and advanced loss functions [32, 42]. For a broader overview of the field, we refer readers to recent surveys [48, 49]. Our MonoIA focuses on intrinsic-aware Mono3D, improving performance across unseen intrinsics.



(a) Training Stage (b) Intrinsic Encoding for Testing

Figure 3. **Overview of MonoIA.** (a) **Training Stage:** MonoIA is a unified intrinsic-aware detection framework built upon two designs. The **Intrinsic Encoder** leverages the knowledge of LLM and CLIP to convert numeric intrinsics into semantically meaningful embeddings that capture their perceptual and geometric effects, providing a strong prior for generalization across cameras. The **Intrinsic Adaptation Module** bridges this semantic knowledge with visual perception through a lightweight Connector and hierarchical fusion, enabling the detector to interpret visual features in an intrinsic-aware manner and maintain consistent 3D detection under diverse camera settings. (b) **Testing Stage:** For each test intrinsic, we retrieve its two nearest seen intrinsics together with their embeddings, and then apply a **Hybrid Interpolation Strategy** that adaptively switches between nearest-neighbor selection and linear interpolation. If the intrinsic gap is ≤ 32 px, the nearest seen embedding is reused; otherwise, the two nearest embeddings are linearly interpolated to synthesize the test intrinsic embedding.

Foundation Models in 3D Tasks. Foundation models [52] such as Large Language Models (LLMs) [1] and vision-language models like CLIP [57, 69] have demonstrated remarkable semantic understanding and cross-modal alignment, driven by large-scale pretraining on massive text and 2D image datasets [60]. However, due to the larger search space in 3D and the limited availability of large-scale 3D datasets [17], analogous foundation models for 3D tasks are still lacking. Recent efforts [55] aim to bridge this gap by leveraging existing 2D or language foundation models to enhance 3D performance. For instance, some [78, 88] pretrain 3D encoders under CLIP supervision, while others [20, 77] utilize LLMs for reasoning in complex 3D scenes. Meanwhile, CLIP Text Encoder has been extensively used in vision and generation tasks [16, 61, 82], demonstrating its ability to encode fine-grained textual semantics, numeric descriptions, and geometric attributes. Distinct from these prior efforts, our work is the first to leverage LLM and CLIP for encoding *camera intrinsics*, transforming numeric parameters into semantically meaningful embeddings that enhance robustness and generalization in Mono3D.

3. Approach

Overview. As Fig. 3 shows, MonoIA comprises three components that jointly simulate, represent, and integrate camera intrinsics within the detection framework. The *Intrinsic Simulation Module* (see Sec. 3.1) generates images with diverse focal lengths while preserving geometric consistency, enriching the training distribution. The *Intrinsic Encoder Module* (see Sec. 3.2) leverages large language and vision–language models to transform numeric intrinsics into semantic representations that capture their perceptual and geometric effects. Finally, the *Intrinsic Adaptation Module* (see Sec. 3.3) injects these embeddings into the detector through a lightweight Connector and hierarchical fusion, enabling consistent 3D understanding across cameras with varying intrinsics.

3.1. Intrinsic Simulation Module

Prior works handle intrinsic variation by normalizing all images to a canonical focal length [5] or by applying heuristic 2D augmentations such as random cropping and scaling [31]. These strategies either eliminate intrinsic diversity or distort the geometric relationship between focal length and field of view (FoV). To address this, we design an *Intrinsic Simulation Module* that performs *FoV-based image approximation* to emulate diverse focal lengths while preserving geometric plausibility. Given an image and its intrinsic matrix $\mathbf{K}_{\text{orig}}$, we randomly sample a target focal length $f_i \in [700, 1300]$ and compute the corresponding FoV:

$$\theta = 2 \arctan\left(\frac{w}{2f_i}\right), \tag{1}$$

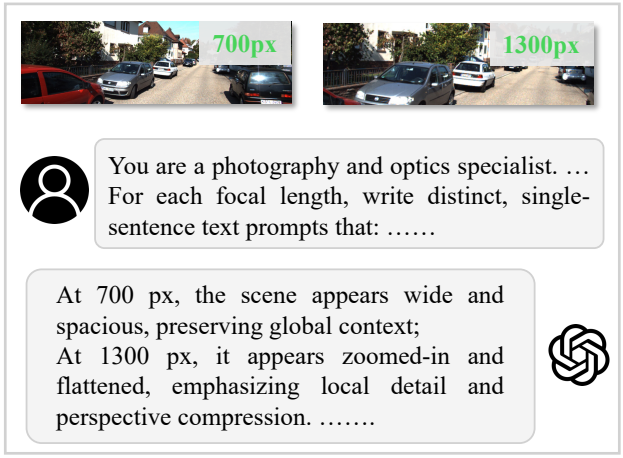


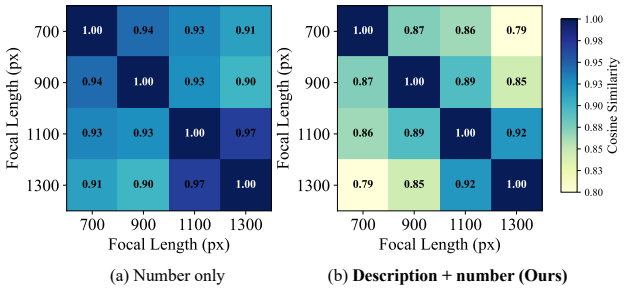
Figure 4. **LLM-Guided Description Generation.** Images rendered with diverse camera intrinsics are fed into an LLM, which generates concise descriptions linking each intrinsic's perceptual and geometric effects with its numeric focal value, forming semantic intrinsic descriptions.

where w denotes the image width. A smaller f_i yields a wider FoV (zoom-out effect), while a larger f_i produces a narrower FoV (zoom-in effect). The simulated image is obtained by resizing the original according to the new FoV, effectively mimicking different camera perspectives without any 3D re-rendering or depth supervision. Although approximate, this lightweight transformation efficiently increases intrinsic diversity while maintaining geometric consistency, allowing the detector to experience a wide spectrum of camera configurations and preparing it for intrinsic-aware learning. We provide simulated image samples in Appendix A.

3.2. Intrinsic Encoder

While the Intrinsic Simulation Module exposes detectors to diverse focal lengths, data diversity alone is insufficient for achieving intrinsic awareness. Empirically, directly training detectors such as MonoCoP [89] on simulated images yields marginal gains, indicating that raw numeric intrinsics (*e.g.*, focal length) provide weak inductive bias. These values do not convey how intrinsic changes alter perceived geometry, scale, or perspective, which are essential cues for intrinsic-aware reasoning. To bridge this gap, we introduce an *Intrinsic Encoder* (see Fig. 3a) that maps numeric intrinsics into *language-grounded* representations, enabling the detector to interpret intrinsics through their perceptual and geometric implications.

LLM-Guided Description Generation. As shown in Fig. 4, for each focal length f_i, an LLM receives its numeric value and a simulated image from the Intrinsic Simulation Module, and generates N concise, content-independent descriptions that capture the optical effects of this intrinsic setting (*e.g.*, changes in field of view, perspective distortion, and depth compression). This prompting design explicitly ties quantitative intrinsics to perceptual outcomes, allowing the LLM to express how focal variation reshapes visual appearance. For example, a shorter focal length yields a wide and spacious view emphasizing global context, whereas a longer focal length compresses perspective and magnifies distant objects. We provide additional details of the prompts and the generated text descriptions in the Appendix B.



(a) Number only (b) **Description + number (Ours)**

Figure 5. **Cosine similarity of intrinsic embeddings under different encoding strategies** (a) Numeric-only encoding produces uniformly high similarity, showing that CLIP text embeddings of raw focal values lack discriminative structure. (b) Our Intrinsic Encoder, which integrates LLM-generated perceptual descriptions with numeric grounding, yields a smooth and ordered similarity pattern, indicating a structured and geometry-aware intrinsic space.

Text Encoding and Embedding Formation. The generated intrinsic descriptions $\{p_i\}_{i=1}^{N}$ are encoded into text embeddings by CLIP Text Encoder [57]:

$$\mathbf{t}_i = \text{CLIP}_{\text{Text}}(p_i), \quad \mathbf{t}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{t}_i. \tag{2}$$

Averaging across descriptions yields an intrinsic embedding $\mathbf{t}_{\text{avg}}$ that captures the shared perceptual meaning of each focal length. CLIP encodes the LLM-generated descriptions into a semantic space where numerically close focal lengths map to similar embeddings, forming a perceptually continuous and geometry-aware representation.

Embedding Analysis. We visualize pairwise cosine similarities among embeddings for focal lengths between 700–1300. As shown in Fig. 5, numeric-only encodings produce uniformly high similarity, indicating a lack of geometric structure. In contrast, our language-guided CLIP embeddings exhibit an ordered pattern where neighboring focal lengths are more correlated, demonstrating that the Intrinsic Encoder successfully models focal variation.

3.3. Intrinsic Adaptation Module

While the Intrinsic Encoder produces semantically rich embeddings that capture the perceptual and geometric meaning of camera intrinsics, these embeddings remain external to the detection process. To make intrinsic awareness actionable, we introduce an *Intrinsic Adaptation Module* that integrates intrinsic embeddings into the Mono3D through a lightweight

Connector and a hierarchical fusion mechanism. The Connector bridges the frozen semantic space and the learnable visual space, while the hierarchical fusion injects intrinsic cues into both feature maps and transformer queries, enabling the detector to adapt its representations according to camera geometry.

Bridging Semantic and Visual Spaces. Intrinsic embeddings from the Intrinsic Encoder reside in a high-level language-aligned space. To preserve their semantic priors while allowing task-specific adaptation, the Connector projects these frozen embeddings into a trainable, vision-aligned space using a two-layer MLP with GELU activation [19]. This projection serves as an interface between semantic priors and visual features, ensuring that intrinsic cues can modulate the detection process without disrupting their original structure and semantic meaning.

Hierarchical Intrinsic Fusion. We then hierarchically inject the transformed intrinsic embedding $\mathbf{t}_{\text{intr}}$ into the detector at both the feature and query levels.

(a) Feature-Level Adaptation. At early stages, intrinsic information conditions the multi-scale backbone features on camera geometry. Given feature maps $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$, each is projected to a shared dimension d' via a 1×1 convolution. The intrinsic embedding is broadcast and added to each spatial position:

$$\widetilde{\mathbf{F}}_i(x,y) = \mathbf{F}'_i(x,y) + \mathbf{t}_{\text{intr}}, \quad i = 1,2,3. \tag{3}$$

This conditioning injects camera awareness into the feature hierarchy, allowing the backbone to maintain geometric consistency across different intrinsics.

(b) Query-Level Adaptation. To propagate intrinsic context into object-level prediction, we modulate object queries used for 3D prediction as:

$$\widetilde{\mathbf{q}}_j = \mathbf{q}_j + \mathbf{t}_{\text{intr}}, \quad j = 1,2,\ldots,N_q. \tag{4}$$

Each query $\mathbf{q}_j$ corresponds to a potential object hypothesis whose appearance and projection depend on the camera intrinsics. This fusion enables the decoder to interpret visual evidence under different focal configurations, producing more stable depth estimation and consistent 3D localization across cameras. Overall, the Intrinsic Adaptation Module links semantic understanding of camera intrinsics with 3D understanding, effectively turning intrinsic knowledge into Mono3D detection.

3.4. Loss Function and Inference

Training. During training, the Intrinsic Encoder is frozen to maintain its pre-trained semantic space, while the Intrinsic Adaptation Module is trained jointly with the detector. Following DETR-based approaches [8, 89], MonoIA uses the Hungarian algorithm to match predictions with ground-truth annotations. The overall training loss is defined as:

$$\mathcal{L}_{\text{overall}} = \frac{1}{N_{gt}} \sum_{n=1}^{N_{gt}} \left(\mathcal{L}_{2D} + \mathcal{L}_{3D} + \mathcal{L}_{\text{dmap}}\right), \tag{5}$$

where N_{gt} is the number of ground-truth objects. $\mathcal{L}_{2D}$ denotes the 2D bounding box loss, $\mathcal{L}_{3D}$ supervises 3D attributes, and $\mathcal{L}_{\text{dmap}}$ corresponds to the object-level depth map prediction loss [84].

Inference. During testing, as illustrated in Fig. 3b, MonoIA performs intrinsic-aware prediction without any retraining. For each test image, we retrieve its two nearest seen intrinsics and their corresponding embeddings from the frozen Intrinsic Encoder and Connector. A *Hybrid Interpolation Strategy* is then applied to synthesize the target intrinsic embedding: if the focal difference is within 32 px, the nearest embedding is reused; otherwise, the two nearest embeddings are linearly interpolated. The 32 px threshold corresponds to the smallest perceivable change after the backbone's $32\times$ spatial downsampling, where finer focal variations become indistinguishable in the feature space. The synthesized intrinsic embedding is finally injected into the Intrinsic Adaptation Module to modulate visual features, ensuring consistent and robust 3D detection under unseen camera intrinsics.

4. Experiments

4.1. Experimental Settings

We evaluate MonoIA through three complementary settings designed to assess its generalization, scalability, and benchmark performance. First, we examine *zero-shot generalization* on KITTI [17] using synthetic intrinsic variations generated by our Intrinsic Simulation Module. Second, we investigate *multi-dataset training* on the combination of KITTI, nuScenes [6], and Waymo [68], which measures the ability of MonoIA to unify data with heterogeneous intrinsic configurations. Finally, we report results on standard benchmarks, including KITTI, nuScenes, and Waymo, to verify that intrinsic awareness not only improves cross-intrinsic robustness but also enhances accuracy under conventional evaluation protocols.

Evaluation Metrics. We report AP_{3D} and AP_{BEV} using IoU thresholds of 0.7 (Car) and 0.5 (Pedestrian, Cyclist)[64] for KITTI. On Waymo, we use the APH_{3D} metric[59] and report results for three distance ranges: $[0, 30)$, $[30, 50)$, and $[50, \infty)$ meters. On nuScenes, we follow [89] and adopt KITTI style metrics for simplicity and consistency.

Implementation Details. MonoIA is built on MonoCoP [89]. We employ ChatGPT-4o [51] to generate text prompts per intrinsic and adopt CLIP ViT-H/14 [57] Text Encoder. We design two training settings. For *single-dataset* training, we train for 250 epochs on one NVIDIA A6000 GPU with a batch size of 16 and a learning rate of 2×10^{-4}

Method	Seen Focals (px)				Unseen Focals (px)													
	700	900	1100	1300	600	650	750	800	850	950	1000	1050	1150	1200	1250	1350	1400
MonoDETR [84]	19.15	18.90	16.76	14.22	14.09	16.67	18.55	17.89	16.21	16.54	15.12	15.06	13.66	12.30	11.88	10.08	7.51
MonoDGP [56]	22.51	21.04	19.96	16.74	17.42	19.28	19.78	19.07	18.51	17.33	16.03	15.63	13.18	12.43	12.47	10.27	7.56
MonoCoP [89]	23.88	23.30	22.59	18.50	18.18	21.70	22.49	21.44	20.20	18.61	17.69	16.43	14.57	13.46	13.11	12.73	11.11
MonoIA (Ours)	**24.41**	**24.36**	**23.69**	**21.20**	**22.43**	**23.41**	**24.13**	**22.93**	**23.64**	**22.48**	**22.65**	**22.52**	**19.07**	**20.54**	**20.80**	**19.25**	**16.99**

Table 1. **Results on seen and unseen focal lengths.** Seen focals include 700, 900, 1100, and 1300 px. Unseen focals include interpolated focals that lie within the training interval and extrapolated focals that extend beyond the training range. MonoIA achieves the highest AP_{3D} across all focal lengths and maintains strong robustness even under extrapolated intrinsics.

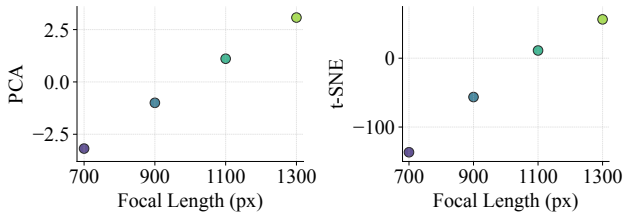


Figure 6. **Visualization of learned intrinsic embeddings.** PCA (Left) and t-SNE (Right). Both views exhibit a smooth, ordered distribution along focal length, indicating that the intrinsic embedding space learned by MonoIA is geometrically consistent and well structured, facilitating interpolation for unseen intrinsics.

using AdamW (weight decay 10^{-4}). For *multi-dataset* training, we train for 120 epochs on four NVIDIA A6000 GPUs with the same batch size and learning rate. Additional implementation details are provided in the Appendix C.

4.2. Generalization on Synthetic Intrinsics

Analysis of Learned Intrinsic Embeddings. MonoIA learns four intrinsic embeddings corresponding to focal lengths $(700, 900, 1100, 1300)$. To examine whether these embeddings capture intrinsic variation, we visualize them using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). As shown in Fig. 6, PCA reveals a clear monotonic trajectory aligned with focal length, while t-SNE forms well-separated clusters, confirming that intrinsic embeddings preserve geometric relationships across cameras. Such structured embedding continuity supports our *Hybrid Interpolation Strategy*, enabling unseen intrinsics to be synthesized via interpolation.

Results on Seen Intrinsics. Existing Mono3D detectors are typically trained under a fixed intrinsic and generalize poorly across cameras. For a fair comparison, we train each baseline (MonoDETR, MonoDGP, and MonoCoP) individually under each focal length, while MonoIA is trained jointly on all four intrinsics, with random sampling to ensure equal total training exposure. Despite this more challenging multi-focal setting, MonoIA achieves the best performance across all seen focal lengths (see Tab. 1), showing that intrinsic-aware modeling enhances both efficiency and accuracy.

Results on Unseen Intrinsics. To comprehensively evaluate intrinsic generalization, we analyze four aspects: (1) interpolation within the training interval, (2) extrapolation beyond the training interval, (3) sensitivity to intrinsic mismatch.

(1) Interpolation. The model is trained on focal lengths $(700, 900, 1100, 1300)$ and tested on intermediate values. As shown in Tab. 1, MonoIA consistently achieves the highest AP_{3D} and remains stable across all interpolated focals.

(2) Extrapolation. We further evaluate focal lengths outside the training range, including values smaller than 700 or larger than 1300. Although extrapolation is naturally more challenging than interpolation, MonoIA still delivers clearly superior performance compared with all baselines, demonstrating strong robustness to unseen intrinsic configurations.

(3) Sensitivity to Intrinsic Mismatch. So far, all experiments assume access to the ground truth intrinsic parameters for each test sample during inference. However, in real world applications this assumption may not be true due to calibration error [95]. To examine the behavior of MonoIA when the provided intrinsics deviate from the ground truth, we perturb the input focal length by ± 5, ± 10, and ± 15 px during inference. For each perturbation magnitude, we report the average accuracy obtained under the perturbed intrinsics. As shown in Tab. 2, MonoIA consistently exhibits the smallest performance drop across all perturbation levels, while existing baselines deteriorate rapidly as the mismatch increases. These results indicate that our intrinsic-aware design offers improved robustness to miscalibrated intrinsics.

Method	GT	± 5 px	± 10 px	± 15 px
MonoDETR [84]	18.55	16.89	14.95	11.21
MonoDGP [56]	19.78	17.76	15.38	12.66
MonoCoP [89]	22.49	20.53	19.22	15.42
MonoIA (Ours)	24.13	23.88	22.34	18.98

Table 2. **Results under intrinsic mismatch with different perturbation levels.** We evaluate performance under the ground truth and perturbed focal lengths (± 5, ± 10, ± 15 px).

4.3. Results on Multi-dataset Training

Since MonoIA is intrinsically aware, it can naturally integrate datasets captured with different focal lengths, enabling unified multi-dataset training. As shown in Tab. 4, existing detectors such as MonoCoP fail under heterogeneous intrin-

Method	Extra Data	Test, AP_{3D} (↑)			Test, AP_{BEV} (↑)			Val, AP_{3D} (↑)			Val, AP_{BEV} (↑)		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
OccupancyM3D [54]	LiDAR	25.55	17.02	14.79	35.38	24.18	21.37	26.87	19.96	17.15	35.72	26.60	23.68
OPA-3D [66]	Depth	24.68	17.17	14.14	32.50	23.14	20.30	24.97	19.40	16.59	33.80	25.51	22.13
MonoTAKD [40]	LiDAR	27.91	19.43	16.51	38.75	27.76	24.14	34.36	22.61	19.88	42.86	29.41	26.47
MonoUNI [25]	None	24.75	16.73	13.49	–	–	–	24.51	17.18	14.01	–	–	–
MonoDETR [84]	None	25.00	16.47	13.58	33.60	22.11	18.60	28.84	20.61	16.38	37.86	26.95	22.80
MonoCD [79]	None	25.53	16.59	14.53	33.41	22.81	19.57	26.45	19.37	16.38	34.60	24.96	21.51
MonoMAE [26]	None	25.60	18.84	16.78	34.15	24.93	21.76	30.29	20.90	17.61	40.26	27.08	23.14
MonoDGP [56]	None	26.35	18.72	15.97	35.24	25.23	22.02	30.76	22.34	19.02	39.40	28.20	24.42
MonoCoP [89]	None	27.54	19.11	16.33	36.77	25.57	22.62	32.06	23.98	20.64	42.20	31.29	27.58
MonoIA (Ours)	None	**29.52**	**20.29**	**17.93**	**37.55**	**26.59**	**23.26**	**33.61**	**24.40**	**20.80**	**44.69**	**32.17**	**27.93**

Table 3. **KITTI Leaderboard (Test) and Val results at $IoU_{3D} \geq 0.7$.** MonoIA achieves SoTA performance across all metrics, demonstrating that our intrinsic aware design also improves standard 3D benchmarks.

Method	Trained on	AP_{3D}^{KIT}	AP_{3D}^{NU}
MonoCoP [89]	KIT	23.98	–
MonoCoP [89]	NU	–	7.39
MonoCoP [89]	KIT + NU	17.26	6.21
MonoCoP [89] + VD [5]	KIT + NU	23.15	7.42
MonoIA	KIT	24.40	–
MonoIA	NU	–	8.12
MonoIA	KIT + NU	26.54	9.81
MonoIA	KIT + NU + Way	**28.91**	**11.48**

Table 4. **Multi dataset training results.** Our intrinsic aware design helps bridge inter dataset discrepancies and improves overall detection performance across KITTI and nuScenes. [Key: KIT = KITTI, NU = nuScenes, Way = Waymo, VD = Virtual Depth]

sics, dropping from $23.64\% \rightarrow 17.26\%$ on KITTI and from $7.39\% \rightarrow 6.21\%$ on nuScenes. Applying virtual-depth (VD) normalization [5] alleviates but does not eliminate this degradation. In contrast, MonoIA improves from $24.40\%/8.12\%$ (single-dataset) to $26.54\%/9.81\%$ (joint KITTI +nuScenes) and further to $28.91\%/11.48\%$ when scaled to three datasets (KITTI, nuScenes, Waymo). These results confirm that our intrinsic-aware design bridges inter-dataset discrepancies and generalizes across diverse camera intrinsics. We provide more detailed results in Appendix D

4.4. Results on Standard 3D Benchmarks

KITTI Leaderboard (Test) Results. Tab. 3 presents the official KITTI test results for the Car at IoU ≥ 0.7, with all numbers sourced from the KITTI leaderboard. MonoIA achieves SoTA performance in both AP_{3D} and AP_{BEV}, surpassing all previous image-only methods. Notably, under the Moderate level which is considered the primary criterion on KITTI, MonoIA outperforms MonoCoP by $+1.18\%$ in 3D detection and $+1.02\%$ in BEV detection. Remarkably, even when compared to models that utilize additional LiDAR or depth inputs (*e.g.*, MonoTAKD and OPA-3D), MonoIA still

Method	AP_{3D}		AP_{BEV}	
	Easy	Mod.	Easy	Mod.
DEVIANT [31]	9.69	8.33	16.28	14.36
MonoDETR [84]	9.53	8.19	16.39	14.41
MonoDGP [56]	10.04	8.78	16.55	14.53
MonoCoP [89]	<u>10.85</u>	<u>9.71</u>	<u>17.83</u>	<u>15.86</u>
MonoIA (Ours)	**12.33**	**10.74**	**19.56**	**17.33**

Table 5. **nuScenes Val Results.** MonoIA achieves SoTA performance on 3D detection and BEV detection under IoU ≥ 0.7. [Key: **First**, <u>Second</u>]

delivers superior results, highlighting the effectiveness of our intrinsic-aware design. We also provide more detailed results on KITTI in Appendix E.

KITTI Val Results. Tab. 3 shows MonoIA achieves consistent SoTA performance on the KITTI Val split. It surpasses the previous best method MonoCoP by $+0.42\%$ AP_{3D} on the Moderate level and $+1.55\%$ on the Easy level, aligning with the trends on the official KITTI leaderboard. These results confirm the effectiveness of intrinsic-aware modeling.

nuScenes Val Results. Tab. 5 shows MonoIA achieves SoTA performance on the nuScenes Val dataset. For instance, MonoIA outperforms MonoCoP by $+1.48\%$ on the AP_{3D} Easy level.

Waymo Val Results. MonoIA also achieves SoTA performance on the Waymo Val and nuScenes Val. Due to space limitations, detailed results of Waymo and nuScenes are provided in the Appendix F and G respectively.

4.5. Efficiency Analysis

Beyond accuracy, as shown in Tab. 7, MonoIA remains highly efficient. It introduces only a marginal increase of $+0.13$M parameters and identical GFLOPs compared to MonoCoP, yet yields consistent performance gains. This indicates that the improvement comes from a more effective

Changed	Row Index	From → To	AP_{3D}, IoU ≥ 0.7 Easy	Mod.	Hard	AP_{3D}, IoU ≥ 0.5 Easy	Mod.	Hard
Baseline	1	Single Focal	32.40	23.64	20.31	71.30	54.70	48.66
	2	Synthetic Images	29.77	21.71	17.46	69.53	51.20	46.87
Intrinsic Encoder	3	Yes → No	29.80	22.16	17.76	69.61	52.63	46.57
	4	Frozen → Trainable	29.76	21.85	18.77	68.81	52.30	47.18
Connector	5	Yes → No	31.97	22.85	19.40	69.54	52.04	48.25
Feature-Level Adaptation	6	Yes → No	32.48	23.43	20.03	71.04	54.02	47.94
Query-Level Adaptation	7	Yes → No	**34.02**	23.99	20.32	71.50	54.25	47.85
MonoIA (Ours)	8	–	33.61	**24.40**	**20.80**	**71.96**	**55.29**	**50.63**

Table 6. **Ablation studies** on KITTI validate the effectiveness of each module in enabling intrinsic-aware detection.

Method	AP_{3D}	Efficiency #Param (M)	GFLOPs
MonoDETR [84]	20.61	35.93	59.72
MonoDGP [56]	22.34	38.90	68.99
MonoCoP [89]	23.98	42.50	71.77
MonoIA (Ours)	**24.40**	42.63	71.77

Table 7. **Efficiency comparison on the KITTI Val set.** MonoIA achieves the SoTA performance while maintaining comparable model size and computational cost to prior works.

design rather than increased model capacity.

4.6. Ablation Study

In this section, we conduct ablation studies to understand the effects of each component of MonoIA on the KITTI Val set. Unless otherwise specified, we adopt AP_{3D} at IoU ≥ 0.7 (Moderate) as the primary evaluation metric. The results summarized in Tab. 6, progressively reveal how (1) intrinsic simulation alone is insufficient, (2) semantic intrinsic encoding enables generalization, and (3) hierarchical adaptation and connector design further enhance the alignment between intrinsic knowledge and visual representations. We also provide additional ablations in Appendix H.

Intrinsic Simulation Module. We first examine whether the performance gain primarily comes from using the synthetic multi-intrinsic dataset. As shown in Tab. 6 (Row 1 vs. 2), directly applying the synthetic data to the baseline MonoCoP leads to a 1.93% performance drop, demonstrating that simply increasing data diversity without intrinsic-aware modeling is ineffective. This highlights the necessity of explicitly encoding and adapting intrinsic information rather than relying on raw data augmentation alone.

Intrinsic Encoder. We then evaluate the contribution of the Intrinsic Encoder. As shown in Tab. 6 (Row 3), replacing the Intrinsic Encoder with a simple linear layer that directly encodes raw intrinsic values leads to a notable performance drop, confirming the importance of semantically meaningful intrinsic embeddings. We further find that freezing the intrinsic embeddings during training is crucial for maintaining stable performance: without freezing, the accuracy drops from 24.40% to 21.85% (Row 4 in Tab. 6). This indicates that updating these embeddings distorts the semantic structure inherited from the pretrained LLM and CLIP encoders, while freezing them preserves the intrinsic knowledge necessary for effective generalization.

Intrinsic Adaptation Module. Finally, we assess the effectiveness of the Intrinsic Adaptation Module. Removing the Connector leads to a noticeable performance drop from 24.40% to 22.85% in AP_{3D}, confirming its necessity for projecting intrinsic embeddings into a learnable feature space and aligning them with visual representations. We further evaluate the two components of the hierarchical fusion design: removing the Feature-Level Adaptation (Row 6) results in the most significant degradation, while removing the Query-Level Adaptation (Row 7) causes a smaller yet consistent decline (-0.41%). These results demonstrate that both adaptation stages contribute to intrinsic-aware learning, with feature-level integration playing a more critical role in preserving geometric consistency.

Visualizations. Appendix I includes further visualizations of MonoIA under KITTI, nuScenes and Waymo.

5. Conclusion

We presented MonoIA, a unified intrinsic-aware framework that converts numeric intrinsics into language-grounded representations capturing their perceptual and geometric implications, and integrates them hierarchically into the detection pipeline. This design enables detectors to interpret how intrinsic changes affect perception and adapt their features accordingly. Extensive experiments across multiple benchmarks show that MonoIA generalizes well to unseen focal lengths, supports multi-dataset training, and achieves new SoTA results. We believe modeling camera intrinsics as semantic representations offers a promising path toward geometry- and perception-aware 3D vision systems that remain reliable across diverse real-world cameras.

References

[1] Jinze Bai, Shuai Bai, and et al. Yunfei Chu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3

[2] Garrick Brazil and Xiaoming Liu. M3D-RPN: Monocular 3D region proposal network for object detection. In *ICCV*, 2019. 2

[3] Garrick Brazil and Xiaoming Liu. Pedestrian detection with autoregressive network phases. In *CVPR*, 2019. 2

[4] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *ECCV*, 2020. 2

[5] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3D object detection in the wild. In *CVPR*, 2023. 1, 3, 7

[6] 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 *CVPR*, 2020. 5

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 5

[9] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teuliere, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image. In *CVPR*, 2017. 2

[10] Dian Chen, Jie Li, Vitor Guizilini, Rares Andrei Ambrus, and Adrien Gaidon. Viewpoint equivariance for multi-view 3D object detection. In *CVPR*, 2023. 2

[11] Junwen Chen, Jie Zhu, and Yu Kong. Atm: Action temporality modeling for video question answering. In *ACM MM*, 2023. 18

[12] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. MonoPair: Monocular 3D object detection using pairwise spatial relationships. In *CVPR*, 2020. 2

[13] Yiwei Chen, Soumyadeep Pal, Yimeng Zhang, Qing Qu, and Sijia Liu. Unlearning isn't invisible: Detecting unlearning traces in llms from model outputs. *arXiv preprint arXiv:2506.14003*, 2025. 18

[14] Yiwei Chen, Yuguang Yao, Yihua Zhang, Bingquan Shen, Gaowen Liu, and Sijia Liu. Safety mirage: How spurious correlations undermine vlm safety fine-tuning and can be mitigated by machine unlearning. *arXiv preprint arXiv:2503.11832*, 2025. 18

[15] Zhili Chen, Shuangjie Xu, Maosheng Ye, Zian Qian, Xiaoyi Zou, Dit-Yan Yeung, and Qifeng Chen. Learning high-resolution vector representation from multi-camera images for 3D object detection. In *ECCV*, 2024. 2

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICLR*, 2024. 3

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3, 5

[18] Xiao Guo, Jie Zhu, Anil Jain, and Xiaoming Liu. On the holistic approach for detecting human image forgery. *arXiv preprint arXiv:2601.04715*, 2026. 18

[19] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[20] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. 2023. 3

[21] Jinghua Hou, Tong Wang, Xiaoqing Ye, Zhe Liu, Xiao Tan, Errui Ding, Jingdong Wang, and Xiang Bai. OPEN: Object-wise position embedding for multi-view 3D object detection. In *ECCV*, 2024. 2

[22] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3D object detection with depth-aware transformer. In *CVPR*, 2022. 2

[23] Rui Huang, Henry Zheng, Yan Wang, Zhuofan Xia, Marco Pavone, and Gao Huang. Training an open-vocabulary monocular 3d detection model without 3d data. *NeurIPS*, 2024. 2

[24] Haoxuanye Ji, Pengpeng Liang, and Erkang Cheng. Enhancing 3D object detection with 2D detection-guided query anchors. In *CVPR*, 2024. 2

[25] Jinrang Jia, Zhenjia Li, and Yifeng Shi. Monouni: A unified vehicle and infrastructure-side monocular 3D object detection network with sufficient depth clues. In *NeurIPS*, 2023. 7, 15

[26] Xueying Jiang, Sheng Jin, Xiaoqin Zhang, Ling Shao, and Shijian Lu. MonoMAE: Enhancing monocular 3D detection through depth-aware masked autoencoders. In *NeurIPS*, 2024. 7, 15

[27] Zheng Jiang, Jinqing Zhang, Yanan Zhang, Qingjie Liu, Zhenghui Hu, Baohui Wang, and Yunhong Wang. FSD-BEV: Foreground self-distillation for multi-view 3D object detection. In *ECCV*, 2024. 2

[28] Sanmin Kim, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, and Dongsuk Kum. LabelDistill: Label-guided cross-modal knowledge distillation for camera-based 3D object detection. In *ECCV*, 2024. 2

[29] Abhinav Kumar, Tim Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. LUVLi face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, 2020. 2

[30] Abhinav Kumar, Garrick Brazil, and Xiaoming Liu. GrooMeD-NMS: Grouped mathematically differentiable nms for monocular 3D object detection. In *CVPR*, 2021. 2

[31] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3D object detection. In *ECCV*, 2022. 1, 2, 3, 7, 15, 16, 17

[32] Abhinav Kumar, Yuliang Guo, Xinyu Huang, Liu Ren, and Xiaoming Liu. SeaBird: Segmentation in bird's view with dice loss improves monocular 3D detection of large objects. In *CVPR*, 2024. 2

[33] Abhinav Kumar, Yuliang Guo, Zhihao Zhang, Xinyu Huang, Liu Ren, and Xiaoming Liu. Charm3r: Towards unseen camera height robust monocular 3d detector. In *ICCV*, 2025. 2

[34] Hyo-Jun Lee, Hanul Kim, Su-Min Choi, Seong-Gyun Jeong, and Yeong Jun Koh. Baam: Monocular 3D pose and shape reconstruction with bi-contextual attention module and attention-guided modeling. In *CVPR*, 2023. 2

[35] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. BEVStereo: Enhancing depth estimation in multi-view 3D object detection with dynamic temporal stereo. In *AAAI*, 2023. 2

[36] Yangguang Li, Bin Huang, Zeren Chen, Yufeng Cui, Feng Liang, Mingzhu Shen, Fenggang Liu, Enze Xie, Lu Sheng, Wanli Ouyang, and Jing Shao. Fast-BEV: A fast and strong bird's-eye view perception baseline. In *NeurIPS Workshops*, 2023. 2

[37] Zhenxin Li, Shiyi Lan, Jose Alvarez, and Zuxuan Wu. BEVNeXt: Reviving dense BEV frameworks for 3D object detection. In *CVPR*, 2024. 2

[38] Hongbin Lin, Zilu Guo, Yifan Zhang, Shuaicheng Niu, Yafeng Li, Ruimao Zhang, Shuguang Cui, and Zhen Li. Drivegen: Generalized and robust 3d detection in driving via controllable text-to-image diffusion generation. In *CVPR*, 2025. 2

[39] Feng Liu, Tengteng Huang, Qianjing Zhang, Haotian Yao, Chi Zhang, Fang Wan, Qixiang Ye, and Yanzhao Zhou. Ray Denoising: Depth-aware hard negative sampling for multi-view 3D object detection. In *ECCV*, 2024. 2

[40] Hou-I Liu, Christine Wu, Jen-Hao Cheng, Wenhao Chai, Shian-Yun Wang, Gaowen Liu, Hugo Latapie, Jhih-Ciang Wu, Jenq-Neng Hwang, Hong-Han Shuai, et al. Monotakd: Teaching assistant knowledge distillation for monocular 3d object detection. In *CVPR*, 2025. 2, 7

[41] Xianpeng Liu, Ce Zheng, Kelvin B Cheng, Nan Xue, Guo-Jun Qi, and Tianfu Wu. Monocular 3D object detection with bounding box denoising in 3D by perceiver. In *ICCV*, 2023. 2

[42] Xianpeng Liu, Ce Zheng, Ming Qian, Nan Xue, Chen Chen, Zhebin Zhang, Chen Li, and Tianfu Wu. Multi-view attentive contextualization for multi-view 3D object detection. In *CVPR*, 2024. 2

[43] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3D object detection. In *ICCV*, 2021. 2

[44] Yunfei Long, Abhinav Kumar, Daniel Morris, Xiaoming Liu, Marcos Castro, and Punarjay Chakravarty. RADIANT: RADar Image Association Network for 3D object detection. In *AAAI*, 2023. 2

[45] Yunfei Long, Abhinav Kumar, Xiaoming Liu, and Daniel Morris. Riccardo: Radar hit prediction and convolution for camera-radar 3d object detection. In *CVPR*, 2025. 2

[46] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3D object detection. In *ICCV*, 2021. 2, 15, 16

[47] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving. In *ICCV*, 2019. 2

[48] Xinzhu Ma, Wanli Ouyang, Andrea Simonelli, and Elisa Ricci. 3D object detection from images for autonomous driving: a survey. *TPAMI*, 2023. 1, 2

[49] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric BEV perception: A survey. *arXiv preprint arXiv:2208.02797*, 2022. 2

[50] Zhixiang Min, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Enrique Dunn, and Manmohan Chandraker. NeurOCS: Neural NOCS supervision for monocular 3D object localization. In *CVPR*, 2023. 2

[51] OpenAI. Chatgpt. 2024. May 14 version. 2, 5, 13

[52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[53] Nadia Payet and Sinisa Todorovic. From contours to 3D object detection and pose estimation. In *ICCV*, 2011. 2

[54] Liang Peng, Junkai Xu, Haoran Cheng, Zheng Yang, Xiaopei Wu, Wei Qian, Wenxiao Wang, Boxi Wu, and Deng Cai. Learning occupancy for monocular 3D object detection. In *CVPR*, 2024. 1, 7, 15

[55] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 3

[56] Fanqi Pu, Yifan Wang, Jiru Deng, and Wenming Yang. Monodgp: Monocular 3D object detection with decoupled-query and geometry-error priors. *arXiv preprint arXiv:2410.19590*, 2024. 1, 2, 6, 7, 8, 13, 15, 16, 17

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 4, 5

[58] Yasiru Ranasinghe, Deepti Hegde, and Vishal M Patel. Monodiff: Monocular 3D object detection and pose estimation with diffusion models. In *CVPR*, 2024. 2

[59] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3D object detection. In *CVPR*, 2021. 2, 5

[60] Christoph Schuhmann, Romain Beaumont, Richard Vencu, and Cade W Gordon et al. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[61] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[62] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *CVPR*, 2019. 1

[63] Changyong Shu, Fisher Yu, and Yifan Liu. 3DPPE: 3D point positional encoding for multi-camera 3D object detection transformers. In *ICCV*, 2023. 2

[64] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3D object detection. In *ICCV*, 2019. 5

[65] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel Lopez Antequera, and Peter Kontschieder. Disentangling monocular 3D object detection: From single to multi-class recognition. *TPAMI*, 2020. 1

[66] Yongzhi Su, Yan Di, Guangyao Zhai, Fabian Manhardt, Jason Rambach, Benjamin Busam, Didier Stricker, and Federico Tombari. Opa-3d: Occlusion-aware pixel-wise aggregation for monocular 3D object detection. *RAL*, 2023. 7, 15

[67] Yiyang Su, Minchul Kim, Jie Zhu, Christopher Perry, Feng Liu, Anil Jain, and Xiaoming Liu. Localscore: Local density-aware similarity scoring for biometrics. *arXiv preprint arXiv:2602.01012*, 2026. 18

[68] Pei Sun, Henrik Kretzschmar, and et al. Dotiwalla. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 5

[69] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2, 3

[70] Yingqi Tang, Zhaotie Meng, Guoliang Chen, and Erkang Cheng. SimPB: A single model for 2D and 3D object detection from multiple cameras. In *ECCV*, 2024. 2

[71] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. In *CVPR*, 2019. 2

[72] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. DistillBEV: Boosting multi-camera 3D object detection with cross-modal knowledge distillation. In *ICCV*, 2023. 2

[73] Zengran Wang, Chen Min, Zheng Ge, Yinhao Li, Zeming Li, Hongyu Yang, and Di Huang. STS: Surround-view temporal stereo for multi-view 3D detection. In *AAAI*, 2023. 2

[74] Zizhang Wu, Yuanzhu Gan, Yunzhe Wu, Ruihao Wang, Xiaoquan Wang, and Jian Pu. FD3D: Exploiting foreground depth map for feature-supervised monocular 3D object detection. In *AAAI*, 2024. 2, 15

[75] Tianfu Wu Xianpeng Liu, Nan Xue. Learning auxiliary monocular contexts helps monocular 3D object detection. In *AAAI*, 2022. 15

[76] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. MonoNeRD: NeRF-like representations for monocular 3D object detection. In *ICCV*, 2023. 2

[77] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024. 3

[78] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *CVPR*, 2023. 3, 13

[79] Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, and Yihua Tan. MonoCD: Monocular 3D object detection with complementary depths. In *CVPR*, 2024. 1, 2, 7, 15

[80] Sunghun Yang, Minhyeok Lee, Jungho Lee, and Sangyoun Lee. Monoclue: Object-aware clustering enhances monocular 3d object detection. *arXiv preprint arXiv:2511.07862*, 2025. 2

[81] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3D object detection and tracking. In *CVPR*, 2021. 1

[82] Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *CVPR*, 2025. 3

[83] Hao Zhang, Hongyang Li, Xingyu Liao, Feng Li, Shilong Liu, Lionel Ni, and Lei Zhang. DA-BEV: Depth aware BEV transformer for 3D object detection. *arXiv preprint arXiv:2302.13002*, 2023. 2

[84] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3D object detection. In *ICCV*, 2023. 1, 5, 6, 7, 8, 15, 16, 17

[85] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3D object detection. In *CVPR*, 2021. 2, 15

[86] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. BEVerse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 2

[87] Zhiahao Zhang, Yiwei Chen, Weizhan Zhang, Caixia Yan, Qinghua Zheng, Qi Wang, and Wangdu Chen. Tile classification based viewport prediction with multi-modal fusion transformer. In *ACM MM*, 2023. 2

[88] Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. Tamm: Triadapter multi-modal learning for 3d shape understanding. In *CVPR*, 2024. 3

[89] Zhihao Zhang, Abhinav Kumar, Girish Chandar Ganesan, and Xiaoming Liu. Unleashing the power of chain-of-prediction for monocular 3d object detection. In *CVPR*, 2026. 1, 4, 5, 6, 7, 8, 13, 15, 16, 17

[90] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. MonoEF: Extrinsic parameter free monocular 3D object detection. *TPAMI*, 2021. 2

[91] Jie Zhu, Yiyang Su, Minchul Kim, Anil Jain, and Xiaoming Liu. A quality-guided mixture of score-fusion experts framework for human recognition. In *ICCV*, 2025. 18

[92] Jie Zhu, Xiao Guo, Yiyang Su, Anil Jain, and Xiaoming Liu. Fusionagent: A multimodal agent with dynamic model selection for human recognition. In *CVPR*, 2026.

[93] Jie Zhu, Yiyang Su, and Xiaoming Liu. Can textual reasoning improve the performance of mllms on fine-grained visual classification? *arXiv preprint arXiv:2601.06993*, 2026. 18

[94] Shengjie Zhu, Garrick Brazil, and Xiaoming Liu. The edge of depth: Explicit constraints between segmentation and depth. In *CVPR*, 2020. 2

[95] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. In *NeurIPS*, 2023. 6