IPD3D: Camera-Agnostic Lightweight Monocular 3D Detection via Geometric Decoupling

1st Zihui Gao
State Key Lab of CAD & CG
Zhejiang University
Hangzhou, China
gaozi@zju.edu.cn

2nd Xinhai Zhao
Huawei Technologies
China
zhaoxinhai1@huawei.com

3rd Hao Chen
State Key Lab of CAD & CG
Zhejiang University
Hangzhou, China
haochen.cad@zju.edu.cn

***Abstract*—Monocular 3D object detection is a critical component of autonomous driving systems, offering a low-cost solution for understanding complex 3D environments. While recent Transformer-based approaches have achieved high accuracy, their excessive computational overhead limits their scalability to real-world, resource-constrained edge devices. Conversely, lightweight CNN-based detectors are efficient but suffer from *intrinsic bias*—they tend to overfit the specific camera focal lengths of the training set, leading to inferior object localization when applied to cameras with different intrinsic parameters. In this work, we propose a lightweight and robust framework, IPD3D, to bridge this gap. We first demonstrate that standard CNN detectors implicitly couple depth estimation with camera intrinsics, limiting their generalization. To address this, we propose an Intrinsic Parameters Decoupling (IPD) module that learns feature representations in a unified *virtual camera* space via affine transformations, effectively decoupling the geometric reasoning from specific camera setups. Furthermore, we leverage an Object Reallocation Confidence (ORC) loss to adaptively reallocate regression weights based on the information adequacy of objects. Extensive evaluations on the nuScenes dataset show that our method achieves competitive performance among lightweight models while maintaining a compact size (∼200MB) and real-time speed. Most importantly, IPD3D demonstrates excellent cross-camera generalization: it achieves state-of-the-art results in a challenging domain adaptation setting (training on nuScenes and testing on KITTI) without any fine-tuning, validating its effectiveness for generic 3D perception.**

***Index Terms*—Monocular 3D Object Detection, Real-Time Perception, Domain Adaptation**

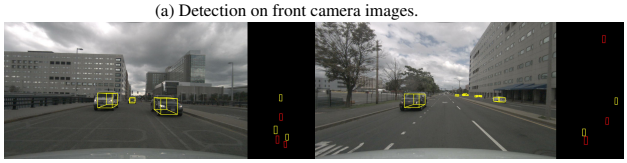
I. Introduction

Monocular 3D object detection is a fundamental and challenging task in computer vision, requiring the reasoning of an object's location, dimension, and orientation in 3D space from a single RGB image. Due to its low hardware cost and ease of deployment compared to LiDAR or stereo systems, it is widely adopted in scenarios such as autonomous driving, visual navigation, and robotics.

Despite the recent progress in 2D object detection [6], [7], recovering 3D information from a monocular image remains an ill-posed problem. Recently, there has been a surge of interest in Vision Transformer (ViT) based approaches, which achieve impressive performance by explicitly modeling long-range dependencies. However, these models often suffer from high latency and massive memory footprints (e.g., several gigabytes), making them impractical for real-time applications on edge devices. On the other hand, lightweight one-stage fully-convolutional approaches [9], [11] offer real-time inference speeds (∼0.03s) and compact model sizes, making them ideal for industrial deployment.

Corresponding author: Hao Chen (haochen.cad@zju.edu.cn)

(a) Detection on front camera images.



(b) Detection on rear camera images.

Fig. 1: **Predicted results for front and rear camera images.** We train a model with images from front camera and inference on both front and rear camera images. Results show that the objects are best predicted with the model trained on same view (*i.e.*, intrinsic parameters). The yellow and red boxes represent predicted bounding boxes and ground-truth.

However, a critical limitation of existing lightweight CNN detectors is their sensitivity to camera intrinsic parameters. Autonomous vehicles (e.g., in the nuScenes dataset [4]) typically mount multiple cameras with different Fields of View (FOV) and focal lengths to cover the surround view. Existing CNN approaches tend to implicitly memorize the relationship between object scale and depth based on the training camera's intrinsics. As a result, simply feeding images from different cameras into the network leads to *geometric overfitting*, causing inferior localization performance on unseen views. In this paper, we strongly advocate that decoupling geometric reasoning from intrinsic parameters is crucial for building robust monocular 3D detectors.

Based on this insight, we analyze the performance of a standard detector when training and testing on different views. As shown in Figure 1, a model trained on front-view images

performs well on the same view but fails catastrophically on rear-view images, especially in depth estimation. This indicates that the model has overfitted to the front camera's focal length, severely limiting its scalability to generic 3D perception across different sensor setups.

To solve this dilemma and achieve a balance between efficiency and robustness, we propose to learn feature representations via decoupling the intrinsic parameters. Specifically, we propose an Intrinsic Parameters Decoupling (IPD) module to transform images or features to an aligned virtual focal length using affine transformation. In this way, the network perceives all inputs through a unified "virtual camera," eliminating the geometric gap caused by varying intrinsics. This enables our lightweight model to generalize to images with various intrinsic parameters without increasing model size. Additionally, to handle the inherent ambiguity of monocular depth, we introduce an Object Reallocation Confidence (ORC) loss in the 3D regression branch. It adaptively reallocates loss weights according to the instance information adequacy, refining both training stability and inference confidence.

Notably, IPD3D exhibits superior cross-camera generalization capability, achieving state-of-the-art performance in zero-shot domain adaptation from nuScenes to KITTI without any fine-tuning, demonstrating its effectiveness as a generic solution for diverse 3D perception scenarios.

We make the following contributions:

(I) We identify the *intrinsic bias* issue in lightweight CNN-based monocular 3D detection and argue that existing approaches are sub-optimal for multi-camera systems due to geometric overfitting.

(II) We propose a plug-and-play Intrinsic Parameters Decoupling (IPD) module to learn unified feature representations, enabling the model to handle varying focal lengths efficiently. We also introduce the ORC loss to improve depth estimation robustness.

(III) We thoroughly evaluate our method on the nuScenes dataset. Our model achieves competitive performance among real-time detectors while maintaining a lightweight parameters (∼200MB). Crucially, it demonstrates superior generalization capability by achieving state-of-the-art results on the real-world domain adaptation setting (*i.e.*, training on nuScenes and testing on KITTI without fine-tuning).

II. Related Work

1) Monocular 3D Object Detection: Significant progress has been made in monocular 3D object detection recently. Existing methods can be broadly categorized into two streams: geometry-based approaches and direct regression approaches.

Geometry-based Approaches. Some methods exploit geometric constraints to infer 3D properties from 2D features. Deep3DBox [13] generates 3D bounding boxes by fitting 3D constraints to detected 2D boxes. To further improve precision, subsequent works utilize keypoints [1], [8] and CAD models [12] as strong geometric priors. Another line of research, such as Pseudo-LiDAR [21], transforms depth maps into point clouds to apply LiDAR-based detectors. AM3D [10] further enhances this by fusing pseudo-LiDAR with RGB information. While these methods utilize geometry explicitly, they often involve multi-stage pipelines that increase computational latency, making them less ideal for real-time edge deployment.

Direct Regression Approaches. To achieve real-time performance, recent trends focus on lightweight, end-to-end frameworks. M3D-RPN [2] introduces depth-aware convolutions with 3D anchors. MonoGRNet [16] predicts depth directly and refines it via feature fusion. Center3D [18] combines classification and regression for better depth estimation. SMOKE [9] significantly simplifies the pipeline by eliminating 2D detection entirely, projecting 3D points directly onto the image plane. Building on SMOKE, MonoDLE [11] re-introduces 2D constraints to filter outliers. However, a critical limitation persists in these direct regression methods (including anchor-free methods like CenterNet [23] and FCOS3D [19]): they typically feed images directly into the network without explicitly accounting for camera intrinsic parameters. This leads to the model implicitly overfitting to the specific focal lengths of the training set (e.g., KITTI [5]). When applied to datasets like nuScenes [4], where front and rear cameras have vastly different intrinsics, these methods suffer from severe domain gaps.

In contrast, our IPD module explicitly decouples intrinsics from feature learning, normalizing all inputs to a virtual camera space for robust generalization across different views.

2) Uncertainty Estimation: In monocular 3D detection, depth ambiguity leads to significant aleatoric uncertainty. Ignoring extreme outliers or hard cases can stabilize training and improve overall performance. Kinematic 3D [3] proposes a self-balanced loss weighted by the rolling mean of recent losses to down-weight hard examples. MonoDLE [11] employs a hard-coding strategy to simply discard samples whose depth exceeds a manually set threshold.

Different from these heuristic or hard-thresholding methods, we propose an Object Reallocation Confidence (ORC) loss. Our approach is fully self-supervised and offers two key benefits: (1) It allows the network to automatically identify and down-weight hard cases based on feature adequacy rather than fixed depth thresholds; (2) It encourages the network to focus on learning samples with sufficient visual cues while outputting reasonable (albeit less precise) predictions for difficult occluded objects.

III. Method

In this section, we present **IPD3D**, a lightweight and robust framework for monocular 3D object detection. Our approach is designed to overcome the intrinsic bias of CNN-based detectors while maintaining real-time performance. As illustrated in Figure 2, the framework consists of two key components: the Intrinsic Parameter Decoupling (IPD) module for geometric alignment, and the Object Reallocation Confidence (ORC) loss for adaptive uncertainty handling.

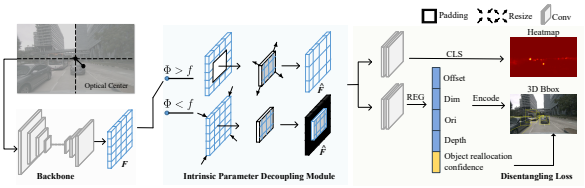


Fig. 2: **Schematic illustration of the proposed IPD3D framework.** Designed for efficient deployment, our method is built upon a lightweight one-stage detector. The core innovation is the **Intrinsic Parameter Decoupling (IPD)** module, which uses affine transformations to align feature maps to a unified virtual focal length. This ensures that the subsequent regression (REG) heads learn *scale-invariant* 3D geometric representations, independent of the specific input camera intrinsics. During training, the **Object Reallocation Confidence (ORC)** branch generates adaptive weights to dynamically handle depth ambiguity and prevent overfitting on ill-posed samples.

A. Problem Formulation

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input RGB image captured by a camera with intrinsic matrix K. The goal is to predict 3D bounding boxes for all objects of interest. Each box is parameterized by its 3D center (x,y,z), dimensions (h,w,l), and yaw orientation R_θ. The projection of a 3D point $P=[x,y,z]^T$ (in the camera coordinate system) onto the 2D image plane $p=[u,v]^T$ is governed by the pinhole camera model:

$$z\cdot[u,v,1]^T = K\cdot[x,y,z]^T, \text{with } K=\begin{bmatrix} f_x & 0 & c_u \\ 0 & f_y & c_v \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

where f is focal length and (c_u,c_v) is the principal point.

Following standard anchor-free detectors like SMOKE [9], our network outputs a heatmap for object classification and a regression map for 3D attributes. The 3D bounding box corners $\hat{B}\in\mathbb{R}^{3\times 8}$ are recovered from the predicted parameters:

$$\hat{B} = R_\theta\cdot[\pm l/2,\pm h/2,\pm w/2]^T + [x,y,z]^T \quad (2)$$

B. Intrinsic Parameter Decoupling (IPD)

1) Geometric Motivation: A fundamental limitation of CNN-based monocular detectors is their implicit reliance on a fixed relationship between object scale in the image (h_{img}) and physical depth (z). From the perspective geometry, $h_{img}\approx f\cdot\frac{h}{z}$. Standard CNNs learn to estimate z by memorizing the scaling factor f from the training set. When test images come from a camera with a different focal length f', the network, unaware of the change, still applies the learned prior for f, leading to a systematic depth error: $z_{pred}\approx z_{gt}\cdot\frac{f}{f'}$.

To address this, we propose to decouple the feature learning from specific camera intrinsics. The core idea is to transform all input images (or feature maps) to a canonical *Virtual Camera* space with a pre-defined focal length Φ. This ensures that the network always observes objects under a consistent geometric projection, regardless of the physical camera used.

2) Affine Transformation Mechanism: We implement the IPD module as a differentiable affine transformation. Given an input feature map $F\in\mathbb{R}^{H'\times W'\times C}$, where $H'=H/R$ and $W'=W/R$ represent the spatial dimensions reduced by the backbone's down-sampling factor R (typically $R=4$ in standard detectors like SMOKE [9]), we aim to transform it to $\hat{F}$ corresponding to the virtual focal length Φ. Mathematically, the scaling of focal length is equivalent to the scaling of the image plane. We define a factor $s=\Phi/f$. The transformation can be formulated as a spatial resampling operation:

$$\hat{F}(u,v) = \mathscr{G}(F,\mathscr{T}_s(u,v)) \quad (3)$$

where $\mathscr{G}$ denotes the bilinear interpolation sampler, and $\mathscr{T}_s$ is the coordinate mapping function. We align the optical centers by setting the coordinate origin to the image center.

Case 1: Zoom-in ($\Phi>f$). The virtual camera has a narrower Field of View (FOV). We perform a center-crop followed by resizing. The effective region of interest in the original feature map is defined by a window of size $(\frac{W'}{s},\frac{H'}{s})$.

Case 2: Zoom-out ($\Phi<f$). The virtual camera has a wider FOV. The original content occupies a smaller portion of the virtual view. We resize the feature map by factor s and pad the borders with zeros to match the target spatial resolution.

This process is computationally negligible and can be inserted either at the image level or the feature level. In our experiments, applying IPD on feature maps yields the best trade-off between efficiency and performance. By normalizing inputs to Φ, the network effectively learns scale-invariant representations, solving the geometric domain gap.

C. Object Reallocation Confidence (ORC) Loss

Monocular depth estimation is inherently ambiguous, leading to varying levels of uncertainty across objects. Standard losses treat all samples equally, which can be detrimental when "hard" samples (e.g., severe occlusion or truncation) dominate the gradients. Unlike previous works that use hard thresholds to discard difficult samples [11], we propose a soft, self-supervised re-weighting strategy.

We add a lightweight branch to predict a confidence score o_i for each object i. This score is mapped to an instance-wise weight $m_i\in(0,1)$ via a sigmoid function: $m_i=\sigma(o_i)$. To prevent the network from trivially minimizing the loss by predicting all weights as zero, we normalize the weights within each batch and introduce a regularization term.

Let N be the number of objects in an image. The normalized weight w_i is computed as:

$$w_i = N\cdot\frac{m_i}{\sum_{k=1}^{N} m_k+\varepsilon} \quad (4)$$

The total regression loss combines the weighted L1 loss and a regularization term:

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N}\left(w_i\cdot\|\mathscr{D}_i-\hat{\mathscr{D}}_i\|_1 + w_i^2\right) \quad (5)$$

where $\mathscr{D}_i$ represents the set of 3D attributes for object i. The regularization term $\sum w_i^2$ penalizes the variance of weights,

encouraging the network to maintain a balanced distribution unless the sample is truly uninformative.

During training, the ORC loss allows the model to dynamically down-weight ambiguous samples. During inference, the predicted weight m_i serves as a quality indicator, which we fuse with the classification score to produce a more reliable final confidence score: $Score_{final} = Score_{cls} \cdot m_i$.

IV. Experiments

A. Dataset & Evaluation metrics

The nuScenes is a large-scale multi-modal dataset that provides a full set of sensor data, *i.e.*, LiDAR, Camera, and Radar. For monocular 3D detection task, it provides RGB images of 6 cameras from 1000 scenes with 3D bounding box annotations for 23 classes and 8 attributes. In evaluation, the center-distance $\mathbb{D}$ is matching criterion. The mean Average Precision (mAP) and NDS are calculated with:

$$\text{mAP} = \frac{1}{|\mathbb{C}||\mathbb{D}|} \sum_{c \in \mathbb{C}} \sum_{d \in \mathbb{D}} \text{AP}_{c,d} \tag{6}$$

$$\text{NDS} = \frac{1}{10}[5\,\text{mAP} + \sum_{\text{mTP} \in \mathbb{TP}} (1 - \min(1, \text{mTP}))] \tag{7}$$

where the mTP is the set of the five mean True Positive metrics, which is average translation, velocity, scale, orientation and attribute errors and they are represented by ATE, AVE, ASE, AOE and AAE.

B. Implementation details

a) Network Architecture.: We adopt the Hourglass-104 [15] network as our backbone for feature extraction. This architecture is chosen for its superior ability to capture multi-scale features while maintaining computational efficiency. Besides, our full framework achieves an inference speed of **33 FPS** ($\sim$0.03s) on a single NVIDIA Tesla V100 GPU with a model size of $\sim$200MB, fully satisfying the requirements for real-time deployment on autonomous vehicles.

b) Training Settings.: We train our model for 24 epochs with a batch size of 32 on 8 Tesla V100 GPUs. We use the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 1×10^{-4}. The initial learning rate is set to 2.5×10^{-4} and is decayed by a factor of 10 at the 12-th and 20-th epochs, respectively.

c) Geometric Configuration.: The nuScenes dataset presents a significant challenge with varying intrinsic parameters: the front camera has a focal length of $f_{front} \approx 1266$, while the rear camera is $f_{rear} \approx 800$. To unify the geometric reasoning space, we set the target virtual focal length in our IPD module to $\Phi = 1000$. This value serves as an intermediate canonical scale, balancing the zoom-in and zoom-out ratios for different views to preserve feature fidelity.

C. Results on nuScenes

Table I compares our method with state-of-the-art CNN-based models. Focusing on the efficiency-accuracy trade-off, we observe: **(I) Superiority among CNN Baselines.** IPD3D achieves the best NDS (45.2%) and mAP among all compared real-time detectors. It outperforms the strong competitor PGD [20] while maintaining a high inference speed of $\sim$33 FPS, proving the effectiveness of handling intrinsics over complex architectures. **(II) Exceptional Geometric Accuracy.** Notably, we achieve the lowest orientation error (mAOE) of **0.34**, significantly outperforming PGD (0.45) and even surpassing LiDAR-based CenterPoint v2 (0.35). This confirms that decoupling intrinsic parameters yields highly consistent geometric representations for precise pose estimation. **(III) Robustness on Large Objects.** We observe substantial gains on large objects sensitive to perspective distortion. Compared to PGD, our method improves AP by $+2.3\%$ for Truck and $+2.1\%$ for Trailer. It shows that IPD effectively normalizes scale variations, enhancing robustness for large instances.

D. Ablation studies

Based on SMOKE, we add individual components gradually to verify their efficacy on nuScenes val set in Table II: (I) We first extend the baseline by increasing positive samples on heatmap, which improves 2.8 detection mAP. Specifically, we do so by enlarging the size of the Gaussian kernel on the ground truth heatmap. (II) Then our proposed IPD transform module allows images with different camera views can be feed into the head without considering intrinsic parameters. It brings a clear 2.1 mAP improvement and substantial gains on all categories. (III) The multi-group refinement strategy [24] brings 2.3 detection AP improvement for Trailer, but impairs some large-amount categories slightly, such as Car, Truck and Bus. (IV) Adding object reallocation confidence loss results in an increase of 1.7 mAP on detection.

E. Further analysis

We also analyze in detail how the IPD transform module and ORC loss boost the module performance.

1) Intrinsic parameter decoupling module: We verify the efficacy of the IPD module by analyzing the performance gap between cameras with dominant focal lengths (Front & Side, $f \approx 1266$) and outlier focal lengths (Rear, $f \approx 800$) in Table III. The baseline model exhibits *geometric overfitting*, yielding a higher Average Translation Error (ATE) on the Rear camera (0.752m) compared to Front & Side (0.726m). By unifying the feature space to a virtual focal length, IPD effectively mitigates this bias. It reduces the ATE on the Rear camera to 0.747m and further improves the Front & Side cameras to 0.716m, demonstrating consistent gains in depth estimation robustness across varying intrinsics.

2) Zero-Shot Cross-Sensor Generalization: To strictly verify the robustness of our approach against intrinsic variations, we conduct a **zero-shot cross-sensor evaluation**. Specifically, we train our model solely on nuScenes and directly inference on the KITTI [5] validation set without any fine-tuning or domain adaptation techniques. This setting is extremely challenging due to the significant differences in camera sensors and mounting heights between the two datasets.

As shown in Table IV, we compare our zero-shot model against the baseline SMOKE [9] model that is fully

Methods	Modality	Time (s)↓	NDS↑	mAP↑	mAOE↓	Car	Truck	Bus	Trailer	C.V.	Ped.	Motor.	Bicy	T.C.	Barr
CenterFusion [14]	C & R	-	44.9	32.6	0.52	50.9	25.8	23.4	23.5	7.7	37.0	31.4	20.1	57.5	48.4
CenterPoint v2 [22]	C & L & R	-	71.4	67.1	0.35	87.0	57.3	69.3	60.4	28.8	90.4	71.3	49.0	86.8	71.0
MonoDIS [17]	Camera	0.10	38.4	30.4	0.55	47.8	22.0	18.8	17.6	7.4	37.0	29.0	24.5	48.7	51.1
CenterNet [23]	Camera	0.04	40.0	33.8	0.63	53.6	27.0	24.8	25.1	8.6	37.5	29.1	20.7	58.3	53.3
FCOS3D [19]	Camera	0.08	42.8	35.8	0.45	52.4	27.0	27.7	25.5	11.7	39.7	34.5	29.8	55.7	53.8
Mono3D	Camera	4.20	42.9	36.6	0.52	54.4	28.9	30.5	24.7	9.3	40.0	36.2	28.8	59.1	54.1
PGD [20]	Camera	**0.03**	44.8	38.6	0.45	56.1	29.9	28.5	26.6	13.4	**44.1**	39.7	**31.4**	**60.5**	**56.1**
Ours	Camera	**0.03**	**45.2**	**38.7**	**0.34**	**58.3**	**32.2**	28.4	**28.7**	13.1	**44.1**	38.4	30.2	59.6	53.5

TABLE I: Comparison of efficiency (Time) and accuracy (NDS, mAP, mAOE) with representative **CNN-based** methods on the nuScenes test set. The top section lists multi-modal methods for reference. Note that our method achieves the **best NDS and mAOE among all CNN-based monocular methods** while maintaining the fastest inference speed (0.03s).

Methods	mAP[%] ↑	Car	Truck	Bus	Trailer	C.V.	Ped.	Motor.	Bicycle	T.C.	Barrier
SMOKE (baseline)	29.5	46.5	23.9	32.0	6.2	2.7	36.2	23.7	30.3	48.8	44.2
+ Positive sample selection	32.3	48.4	26.0	34.7	8.2	4.8	40.1	27.3	32.9	52.2	48.0
+ IPD module	34.4	53.2	28.2	38.0	10.4	5.9	42.0	30.0	34.0	53.5	49.0
+ Multi-group refinement [24]	34.9	52.9	27.4	36.6	12.7	6.1	42.4	31.2	34.0	55.4	50.3
+ ORC loss	36.6	56.2	29.4	37.3	14.0	6.1	45.0	33.2	35.5	58.7	50.9

TABLE II: Ablation study results on nuScenes validation set.

Camera	Ablation	ATE[m]↓	ASE[1-IOU]↓	AOE[rad]↓
Front & Side	w/o IPD	0.726	0.260	0.356
	w/ IPD	**0.716**	0.264	0.366
Rear	w/o IPD	0.752	0.393	0.465
	w/ IPD	**0.747**	0.395	**0.444**

TABLE III: **Errors comparison on Front&Side and Rear camera results.** In 6 cameras, the front and side cameras have similar focal length of $f_{front} = 1266$ and the rear camera has a different one of $f_{rear} = 800$. We choose to report results from the front & rear cameras with two typical f. The IPD transforms focal length to $f_0 = 1000$.

Method	AP_{BEV} @ $IoU = 0.7$ Easy	Moderate	Hard
SMOKE [9] (*Supervised*: KITTI→KITTI)	**19.99**	15.61	15.28
Ours (***Zero-Shot***: nuScenes→KITTI)	19.53	**16.11**	**15.54**

TABLE IV: Zero-shot generalization results on KITTI val set. "nuScenes→KITTI" indicates the model is trained on nuScenes and tested on KITTI. Note that our zero-shot model outperforms the fully supervised baseline on Moderate and Hard settings, demonstrating superior geometric robustness.

trained on KITTI. **Remarkably, our cross-domain model (nuScenes→KITTI) outperforms the fully supervised baseline (KITTI→KITTI) on both Moderate and Hard BEV metrics.** Specifically, we achieve a gain of **+0.50%** and **+0.26%** AP on Moderate and Hard levels, respectively.

This result strongly suggests that standard CNN detectors (like SMOKE) suffer heavily from geometric overfitting to the specific training camera. In contrast, our IPD successfully decouples the geometric reasoning from sensor intrinsics, enabling the network to learn universal depth cues that generalize across different sensor setups. This marks a significant step towards generic, plug-and-play monocular 3D perception.

3) Effectiveness of ORC Mechanism: We further analyze how the Object Reallocation Confidence (ORC) loss contributes to robustness. **Adaptive Training.** As visualized in Figure 3, the ORC mechanism automatically assigns lower weights (low m_i) to occluded or distant instances with inadequate depth information, allowing the network to focus on high-quality samples. This prevents the model from overfitting to ambiguous visual cues. **Inference Refinement.** During inference, these learned weights serve as a proxy for depth uncertainty. By fusing them with 2D classification scores, we effectively downgrade the confidence of unreliable predictions.

V. Conclusion

In this paper, we addressed the critical issue of *geometric overfitting* in lightweight monocular 3D object detection. We identified that standard CNN detectors implicitly couple depth estimation with training camera intrinsics, severely limiting their scalability across multi-camera systems. To bridge this gap, we proposed **IPD3D**, a real-time framework that learns intrinsic-invariant representations via a novel Intrinsic Parameter Decoupling (IPD) module and handles depth uncertainty with the Object Reallocation Confidence (ORC) loss. Extensive experiments on nuScenes demonstrate that our method achieves a superior trade-off between speed and accuracy (∼33 FPS on a V100). Most importantly, our approach exhibits remarkable **zero-shot generalization** ability when transferring from nuScenes to KITTI, proving its potential as a robust and practical solution for real-world autonomous driving and robotics applications.

	mAP[%] ↑	Car	Truck	Bus	Trailer	C.V.	Ped.	Motor.	Bicycle	T.C.	Barrier
w/o ORC	34.4	53.2	28.2	38.0	10.4	5.9	42.0	30.0	34.0	53.5	49.0
w/ ORC (train)	36.1	55.4	29.4	37.2	14.0	6.1	43.9	32.3	35.2	57.4	50.5
w/ ORC (infer)	36.6	56.2	29.4	37.3	14.0	6.1	45.0	33.2	35.5	58.7	50.9

TABLE V: Ablation study on ORC loss at training and inference.

Fig. 3: **Qualitative analysis of adaptive training via ORC loss.** Comparison between the baseline (left) and our method with ORC (right). The ORC mechanism enables the network to distinctively prioritize targets: it improves depth estimation for objects with rich geometric cues (fully visible), while adaptively down-weighting ambiguous instances (severely occluded) to prevent overfitting. Different colored 3D boxes represent successful matches under tightening distance error thresholds (from loose to strict).

Fig. 4: **Zero-shot qualitative results on KITTI.** Predictions (**yellow**) vs. gt (**red**) from a model trained *solely* on nuScenes. The accurate localization demonstrates our model's strong cross-sensor generalization capability without any fine-tuning.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62576315)

References

[1] Barabanau, I., Artemov, A., Burnaev, E., Murashkin, V.: Monocular 3d object detection via geometric reasoning on keypoints. arXiv preprint (2019) 2

[2] Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: ICCV (2019) 2

[3] Brazil, G., Pons-Moll, G., Liu, X., Schiele, B.: Kinematic 3d object detection in monocular video. In: ECCV (2020) 2

[4] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020) 1, 2

[5] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012) 2, 4

[6] Girshick, R.: Fast r-cnn. In: ICCV (2015) 1

[7] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 1

[8] Li, P., Zhao, H., Liu, P., Cao, F.: Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In: ECCV (2020) 2

[9] Liu, Z., Wu, Z., Tóth, R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. In: CVPR (2020) 1, 2, 3, 4, 5

[10] Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., Fan, X.: Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In: ICCV (2019) 2

[11] Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W.: Delving into localization errors for monocular 3d object detection. In: CVPR (2021) 1, 2, 3

[12] Manhardt, F., Kehl, W., Gaidon, A.: Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In: CVPR (2019) 2

[13] Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J.: 3d bounding box estimation using deep learning and geometry. In: CVPR (2017) 2

[14] Nabati, R., Qi, H.: Centerfusion: Center-based radar and camera fusion for 3d object detection. In: CVPR (2021) 5

[15] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016) 4

[16] Qin, Z., Wang, J., Lu, Y.: Monogrnet: A geometric reasoning network for monocular 3d object localization. In: AAAI (2019) 2

[17] Simonelli, A., Bulo, S.R., Porzi, L., López-Antequera, M., Kontschieder, P.: Disentangling monocular 3d object detection. In: ICCV (2019) 5

[18] Tang, Y., Dorn, S., Savani, C.: Center3d: Center-based monocular 3d object detection with joint depth understanding. In: GCPR (2020) 2

[19] Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. arXiv preprint (2021) 2, 5

[20] Wang, T., Zhu, X., Pang, J., Lin, D.: Probabilistic and geometric depth: Detecting objects in perspective. arXiv preprint (2021) 4, 5

[21] Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In: CVPR (2019) 2

[22] Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: CVPR (2021) 5

[23] Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint (2019) 2, 5

[24] Zhu, B., Jiang, Z., Zhou, X., Li, Z., Yu, G.: Class-balanced grouping and sampling for point cloud 3d object detection. arXiv preprint (2019) 4, 5