Adaptive Dual Uncertainty Optimization: Boosting Monocular 3D Object Detection under Test-Time Shifts

Zixuan Hu[1] Dongxiao Li[1] Xinzhu Ma[3] Shixiang Tang[3] Xiaotong Li[1]
Wenhan Yang[2] Ling-Yu Duan[1,2*]
[1]School of Computer Science, Peking University, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China, [3]The Chinese University of Hong Kong, Hongkong, China.
{hzxuan, lingyu}@pku.edu.cn

Abstract

*Accurate monocular 3D object detection (M3OD) is pivotal for safety-critical applications like autonomous driving, yet its reliability deteriorates significantly under real-world domain shifts caused by environmental or sensor variations. To address these shifts, Test-Time Adaptation (TTA) methods have emerged, enabling models to adapt to target distributions during inference. While prior TTA approaches recognize the positive correlation between low uncertainty and high generalization ability, they fail to address the dual uncertainty inherent to M3OD: semantic uncertainty (ambiguous class predictions) and geometric uncertainty (unstable spatial localization). To bridge this gap, we propose **D**ual **U**ncertainty **O**ptimization (**DUO**), the first TTA framework designed to jointly minimize both uncertainties for robust M3OD. Through a convex optimization lens, we introduce an innovative convex structure of the focal loss and further derive a novel unsupervised version, enabling label-agnostic uncertainty weighting and balanced learning for high-uncertainty objects. In parallel, we design a semantic-aware normal field constraint that preserves geometric coherence in regions with clear semantic cues, reducing uncertainty from the unstable 3D representation. This dual-branch mechanism forms a complementary loop: enhanced spatial perception improves semantic classification, and robust semantic predictions further refine spatial understanding. Extensive experiments demonstrate the superiority of DUO over existing methods across various datasets and domain shift types. The source code is available at https://github.com/hzcar/DUO.*

1. Introduction

Monocular 3D object detection (M3OD) [1, 33, 34] serves as a fundamental perception task, enabling agents to understand 3D scenes directly from 2D images, as shown in Fig. 1(a). Due to its low cost and simple hardware configuration, M3OD has attracted widespread attention, leading to the development of numerous detectors [7, 31, 52]. However, when facing adverse weather conditions or sensor failures in real-world deployments, well-trained models often suffer severe performance degradation under data shifts [2, 14, 49]. Therefore, it is crucial to deal with the OOD generalization problem for M3OD.

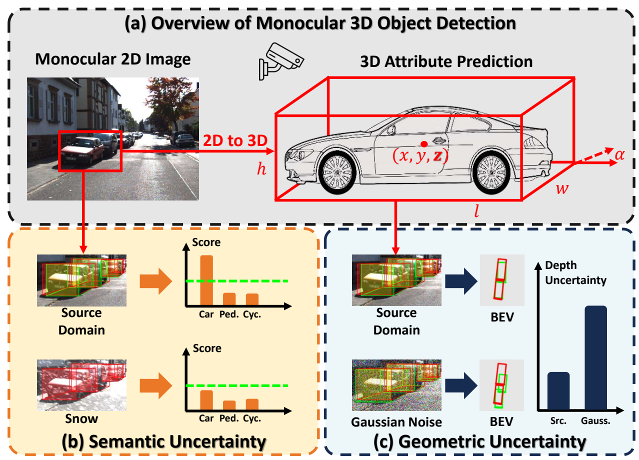

Figure 1. (*a*) Illustration of M3OD. (*b, c*) Under test-time shifts, increased semantic and geometric uncertainties lead to a degradation of semantic perception and spatial localization capabilities.

To address the distribution shifting issue with minimal overhead, Test-Time Adaptation (TTA) has emerged as a critical paradigm, enabling source models to adapt to target distributions via online updates [17, 19, 23]. The dominating strategy in this context involves minimizing prediction entropy, thereby reducing the uncertainty of the model on shifted data [17, 46]. Despite the promising results, existing TTA approaches largely overlook the dual uncertainty inherent in 3D detection—semantic uncertainty (related to class predictions) [42, 50] and geometric uncertainty (related to spatial location) [30, 39], which is a significant dif-

*Corresponding Author.

ference compared to conventional 2D tasks.

To investigate these two types of uncertainty under test-time shifts, we analyze detection outcomes for objects subject to common real-world variations. As shown in Fig. 1 (b)&(c), our empirical study reveals that both uncertainties increase markedly with data shifts, creating compounded error accumulation in 3D detection. Moreover, we find that existing uncertainty optimization techniques exhibit significant limitations: 1) Low-score object neglect. Entropy minimization fails to provide effective supervision for challenging objects with low detection scores, resulting in inevitable omissions. 2) Spatial perception collapse. Direct minimization of depth uncertainty can cause model collapse, compromising the perception capacity of spatial attributes.

In this work, to overcome the above limitations, we propose **D**ual **U**ncertainty **O**ptimization (**DUO**), the first TTA framework for joint semantic-geometric uncertainty minimization. Specifically, through the lens of convex optimization theory [3, 45], we present an Legendre–Fenche structure of the focal loss [25] and reconstruct the semantic uncertainty minimization as a dual optimization problem. Building on this foundation, we apply higher-order approximation analysis to derive a novel Conjugate Focal Loss. This loss breaks the label-dependence barrier in the original objective and introduces a dynamic weighting mechanism for balanced training, effectively reducing the omissions of objects with low scores. In parallel, we introduce a normal field constraint that enforces local consistency of surface normals in regions with high semantic confidence. This spatial coherence clarifies geometric cues, thereby reducing geometric uncertainty. Together, this dual-branch design creates a complementary loop where semantic confident regions bootstrap geometric feature learning, while enhanced spatial perception guides semantic refinement.

We evaluate the effectiveness and generalizability of our method through experiments on the KITTI dataset [11] with 13 corruption shift types, achieving state-of-the-art results with average improvements of **+2.2** $\mathrm{AP}_{3\mathrm{D}|\mathrm{R}_{40}}$ in the Car category. Furthermore, we showcase its superior performance in addressing real-world shift scenarios (daytime $\leftrightarrow$ night, sunny $\leftrightarrow$ rainy) of nuScenes dataset [5], yielding an average gain of **+18%** compared to existing methods.

Contributions: 1) To the best of our knowledge, we pioneer dual uncertainty optimization in M3OD by establishing the first TTA framework that jointly minimizes semantic and geometric uncertainties, addressing a critical reliability gap in real-world deployments. 2) Through the lens of convex optimization theory, we derive a novel Conjugate Focal Loss that enables label-agnostic uncertainty weighting and balanced learning for low-score objects. This approach is inherently compatible with the source phase, requiring no additional hyperparameter tuning. 3) We introduce a normal field constraint that enforces the stability of geometric representation with semantic guidance, resolving ambiguous spatial predictions. 4) We analyze and verify that our dual-branch design creates a complementary loop of two types of uncertainty optimization, resulting in significantly improved performance over existing TTA methods.

2. Related Work

Monocular 3D Object Detection (M3OD) aims to perceive 3D objects from a single 2D image. Existing methods in M3OD can be broadly categorized based on their use of extra data sources, such as CAD models [28], dense depth maps [8, 47], or LiDAR [18, 41]. In this paper, we focus exclusively on approaches that utilize only monocular images, due to their computational efficiency and lower deployment costs. Previous studies, such as MonoDLE [32] and PGD [48], have identified depth estimation as a critical bottleneck in M3OD. To address this, many works leverage multiple geometric cues to integrate diverse depth predictions. For example, MonoFlex [55] integrates depth prediction by combining direct regression with multi-keypoint estimation; MonoGround [40] incorporates the ground plane as prior information; and MonoCD [53] exploits the complementary properties of multi-head estimation. In this paper, we investigate how to enhance the detection performance of such detectors under test-time shifts.

Test-Time Adaptation (TTA) aims to enhance model performance on out-of-distribution samples during inference. Depending on whether the source training process is modified, existing TTA methods can be mainly divided into two groups: Test-Time Training [13, 27, 44] and Fully Test-Time Adaptation [4, 16, 56]. In this paper, we focus on Fully TTA, which adapts models without source data. Prior studies have demonstrated that reducing prediction uncertainty is an effective strategy for improving model generalization in various tasks [10, 21, 26, 51]. These works have developed various strategies to model and optimize uncertainty. For instance, SAR [37] uses entropy as a measure of classification uncertainty along with sharpness-aware optimization. DeYO [22] incorporates entropy with a disentangled factor for uncertainty modeling. ReCAP [17] models regional uncertainty, enforcing implicit data scaling for uncertainty optimization. MonoTTA [24] optimizes positive and negative class uncertainties separately. However, unlike conventional 2D tasks, M3OD inherently exhibits both semantic and geometric uncertainties, which remain largely unexplored and lack an effective training paradigm. In this work, we explore the joint optimization of this dual uncertainty, enhancing the robustness of M3OD models.

3. Preliminary

Task Definition. M3OD aims to predict the 3D and semantic attributes of objects from a single RGB image. Given an input image $\mathrm{I} \in \mathbb{R}^{H \times W \times 3}$, the goal is to generate accurate

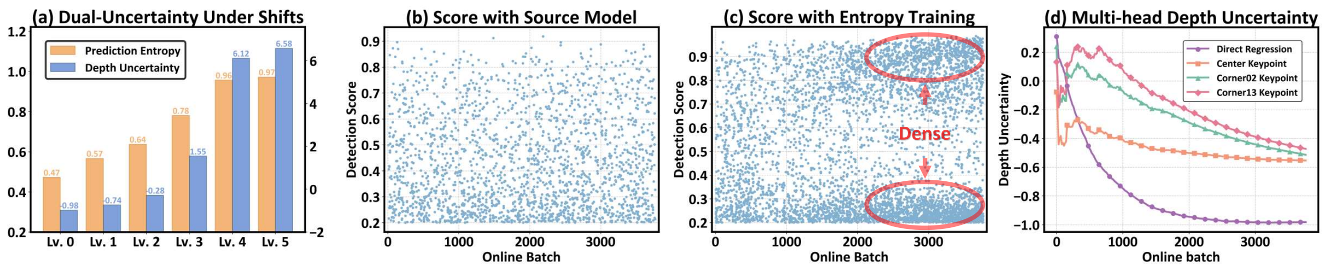

Figure 2. Semantic uncertainty (prediction entropy) and geometric uncertainty (average depth uncertainty) of the M3OD model under domain shifts. (*a*) shows variations in dual-uncertainty as the shift level increases. (*b*) records detection scores of predicted objects during inference with the source model. (*c*) records scores during the entropy minimization process. (*d*) investigates the depth uncertainty of different heads during the optimization with uncertainty regression loss. Experiments are conducted on the Gauss. shift type of KITTI-C.

3D bounding boxes $\{\mathcal{B}_i\}_{i=1}^N$ and semantic labels $\{\mathcal{C}_i\}_{i=1}^N$ for objects present in the scene, where N denotes the number of objects. Each bounding box is typically parameterized as $\mathcal{B}_i = (\mathbf{P}_i, \mathbf{D}_i, \mathbf{O}_i)$, where $\mathbf{P}_i \in \mathbb{R}^3$ denotes the 3D center position, $\mathbf{D}_i \in \mathbb{R}^3$ represents the shape dimensions, and $\mathbf{O}_i \in [-\pi, \pi]$ encodes the orientation. The multi-task nature of M3OD, which involves simultaneous estimation of both geometric and semantic attributes, poses significant challenges for achieving precise and coherent predictions.

Meta-framework. As shown in Fig. 3(a), M3OD models typically employ a backbone network connected with multiple branches to predict various properties, *e.g.*, score heatmaps, depth maps, *etc*. Since depth estimation is widely recognized as a key bottleneck [32, 48], many approaches employ a multi-head depth estimator to improve prediction accuracy. This estimator comprises a direct regression depth head and multiple geometric depth heads, each providing an individual depth prediction along with an associated uncertainty value. These outputs are then fused via an uncertainty-weighted average to produce the final depth prediction. We utilize the average uncertainty across all heads as our depth uncertainty metric. A detailed explanation is provided in the Appendix. D.2.

TTA Setting. TTA addresses the challenge of distribution shifts by enabling a pre-trained model to adapt to the target distribution during inference, without the need for labeled data. Unlike traditional training on fixed datasets, TTA operates in an online manner, where the model h_θ with parameter set θ produces detection outputs while concurrently updating its parameters based on incoming test data:

$$\{\mathcal{B}_t\}, \{\mathcal{C}_t\} \leftarrow h_\theta(\mathrm{I}_t),\ \theta \leftarrow \theta - \nabla \mathcal{L}_{tta}(\mathrm{I}_t), \tag{1}$$

where I_t denotes the incoming test data and $\mathcal{L}_{tta}$ denotes the loss function used for self-training during adaptation.

4. Uncovering Dual-Uncertainty under Shifts

Since TENT [46] revealed the positive correlation between prediction uncertainty and generalization error under distribution shifts, numerous TTA methods have emerged to optimize uncertainty metrics. While previous works have developed diverse measures for conventional 2D tasks [22, 37], the compound semantic and geometric uncertainties in 3D perception remains critically unexplored in the TTA field. In this work, we investigate both types of uncertainty within the context of M3OD to understand their distinct roles.

Specifically, to quantify the variation of uncertainty under distribution shifts, we track two metrics: semantic prediction entropy and geometric depth uncertainty (average uncertainty of multi-head depth estimator). As shown in Fig. 2(a), both metrics demonstrate a consistent upward trend as distribution shifts intensify, indicating that more severe shifts lead to higher model uncertainty. Furthermore, we analyze the independent optimization effects of two uncertainties and make the following key observations:

Observation 1: *Conventional entropy minimization exacerbates imbalance distribution of detection scores.* Different from classification scenarios, object detection suffers from extreme foreground-background imbalance, which hinders the effective optimization of hard positive objects [6, 25]. As shown in Fig. 2(b)&(c), entropy minimization yields a marginal gain for low-score objects while significantly boosting high-score predictions, further exacerbating the imbalance and leading to omissions of low-score objects.

Observation 2: *Minimizing depth uncertainty directly causes the model collapse of the multi-head depth estimator.* To minimize depth uncertainty, we apply the uncertainty regression loss across multiple depth heads (detailed in the Appendix. D.2). As shown in Fig. 2(d), the regression head which lacks any geometric constraints, exhibits a significantly faster decline of uncertainty compared to keypoint heads. This rapid convergence reduces the multi-head system to a single deterministic head, undermining the model's ability to perform robust spatial understanding.

5. Methodology

Based on the above observations, we propose a TTA method **D**ual **U**ncertainty **O**ptimization (**DUO**) for M3OD, which mainly leverages two novel designs, *i.e.*, conjugate optimization framework and normal consistency constraint, to compatibly optimize semantic and geometric uncertainties.

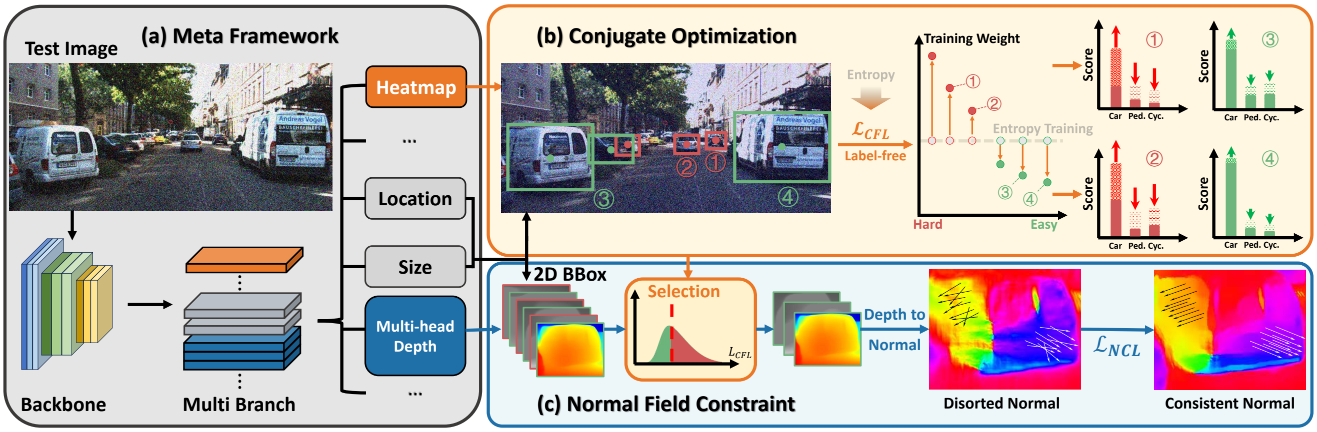

Figure 3. Overview of our DUO method. (*a*) Meta-framework of the base models, where a backbone network connects to multi-branch predictors for estimating various object properties. (*b*) Through our theoretical conjugate optimization framework, we derive a conjugate focal loss $\mathcal{L}_{\text{CFL}}$ that can adaptively adjust training weights across all classes without the ground-truth labels, while also identifying low-semantic-uncertainty regions for geometric constraints (Sec. 5.1). (*c*) DUO employs the efficient Sobel operator to convert the depth map to a normal field and design a normal consistency constraint $\mathcal{L}_{\text{NCL}}$ to enhance geometric consistency in selected regions (Sec. 5.2).

5.1. Semantic Uncertainty Conjugate Optimization

To address the imbalance problem, it is crucial to devote more attention to challenging objects with high uncertainty. A straightforward approach is to employ the focal loss [25] which increases the weight of predictions with low probability. However, focal loss cannot assign an appropriate weight without the ground-truth label and fails to adapt effectively in unsupervised settings. To provide a more flexible weighting solution, we leverage convex optimization theory [3, 45] (a classical analytical tool) to explore the focal loss via a Legendre–Fenchel structure. Building on this foundation, we perform a higher-order approximation analysis to derive a novel loss that can dynamically adjust the weighting without relying on labels, as shown in Fig. 3(b).

Legendre–Fenchel Structure. Given the source model h_θ pre-trained with the focal loss, we denote $h \triangleq h_\theta(x)$ and its corresponding loss function can be formulated as:

$$\mathcal{L}_{\text{FL}}(x, y) = -\alpha(1-p)^\gamma \cdot y \log p, \tag{2}$$

where $p \triangleq \texttt{softmax}(h)$ is the normalized probability over the classes, x is a detected object, and y is the one-hot coding of the ground-truth label. α and γ are hyperparameters controlling the weighting of uncertain predictions. Motivated by [12], we reformulate it as the following structure:

$$\begin{gathered}\mathcal{L}_{\text{FL}} = f(h) - y^\top g(h), \\ f(h) = \alpha \log s,\ g(h) = \alpha h + \alpha((1-p)^\gamma - 1)\log p,\end{gathered} \tag{3}$$

where $s \triangleq e^{h_1} + ... + e^{h_c}$ is the sum over the exponential outputs of the model and c is the number of classes. Refer to the Appendix. A for missing proofs of this subsection.

Problem Reconstruction. Under this structure, the optimization problem can be regarded as finding an optimal representation h that minimizes the empirical loss. According to the Legendre–Fenchel condition [45], the existence of the conjugate function f^* is equivalent to the invertibility of the function g. This critical condition is formally established in our work (see Appendix. A), serving as a fundamental theoretical guarantee for our method. Therefore, the minimum value of the objective can be expressed as follows:

$$\min_h\{f(h) - y^\top g(h)\} = \min_{z=g(h)}\{f \circ g^{-1}(z) - y^\top z\} = f^*(y). \tag{4}$$

Building on the common assumption that the representation h pre-trained from the large source dataset is already close to a locally optimal solution h_0 [9, 35], we can convert the problem into the following relationships:

$$f \circ g^{-1}(z) - y^\top z = f^*(y), \quad \nabla_z(f \circ g^{-1}) = y. \tag{5}$$

Conjugate Focal Loss. By applying the chain rule of gradient computation and higher-order approximation, we can obtain the following estimation:

$$\begin{aligned} y_0 &\triangleq \frac{\nabla_h(f \circ g^{-1})}{\nabla_h z}|_{z=g(h)} = \nabla_h g(h)^{-1} \cdot \nabla_h f(h) \\ &\approx (I + \gamma(1-\log p)\cdot pp^\top - \gamma \log p \cdot \text{diag}(p)))^{-1} p,\end{aligned} \tag{6}$$

where $\text{diag}(\cdot)$ denotes the diagonal matrix and I denotes the identity matrix. This estimation y_0 ensures that the loss aligns closely with the conjugate function. Ultimately, we substitute it into Equ. 5, yielding an unsupervised approximation of the conjugate function:

$$\begin{aligned}\mathcal{L}_{\text{CFL}}(x) = &-\alpha(1-p)^\gamma (I + \gamma(1-\log p)\cdot p^\top p \\ &-\gamma \log p \cdot \text{diag}(p))^{-1} p \log p.\end{aligned} \tag{7}$$

In contrast to the focal loss in Equ. 2, our derived **C**onjugate **F**ocal **L**oss (CFL) offers a dynamic adjustment of uncertainty weighting without relying on ground-truth labels and it confers the following advantages:

Static vs. Dynamic Adjustment. The vanilla focal loss uses a fixed weighting term for the ground-truth class to focus on high-uncertainty predictions. In contrast, our CFL not only incorporates $(1-p)^\gamma$ to address the imbalance training issue but also dynamically adjusts the weighting across all classes based on the term $(I+\gamma(1-\log p)pp^\top - \gamma \log p \cdot \mathrm{diag}(p))^{-1}$, which encodes inter-class prediction relationships.
Ground-Truth Independence. While focal loss requires labels to compute the weight for the loss term, CFL operates solely based on the prediction probability, eliminating the need for labeled data during TTA.
Compatible Hyperparameters with Source Phase. Our theoretical analysis suggests that hyperparameters α, γ should remain consistent with their values used in focal loss during the source training. This compatibility provides a practical advantage for TTA scenarios, eliminating the need for extensive hyperparameter tuning. The validity of this consistent setting is empirically verified in Appendix C.

5.2. Semantic-Guided Normal Field Constraint

Despite progress in uncertainty modeling, current methods struggle to handle geometric uncertainty in TTA scenarios due to critical limitations: 1) *Model Collapse*: Direct optimization of model-predicted uncertainty leads to the degenerated predictor, as discussed in Sec. 4. 2) *Substantial Overhead:* Geometry-aware methods typically require additional data or offline training for uncertainty estimation, limiting their real-time applicability [15, 39]. To overcome these challenges, we propose an efficient normal field constraint derived from a single image, which enhances the geometric coherence of 3D predictions, thereby reducing the uncertainty stemming from the unstable geometric representation, as shown in Fig. 3(c).

Normal Field. Given a depth map D, we restore it to the original image resolution using bilinear interpolation, allowing for the alignment with the pixel grid. Then, we compute the spatial gradients of the depth map using efficient Sobel operators [20], which approximate the rate of change in the depth values along horizontal and vertical directions:

$$\nabla D_x = \mathbf{S}_x * D, \quad \nabla D_y = \mathbf{S}_y * D, \tag{8}$$

where $\mathbf{S}_x$ and $\mathbf{S}_y$ denote horizontal and vertical Sobel kernels, respectively. These gradient maps capture the variation in depth across neighboring pixels. The surface normal field, which encodes the orientation of the surface at each pixel, is then derived from the gradients as:

$$\mathbf{N}(u,v) = \frac{1}{\sqrt{\nabla D_x^2 + \nabla D_y^2 + 1}} \begin{bmatrix} -\nabla D_x \\ -\nabla D_y \\ 1 \end{bmatrix}, \tag{9}$$

where $\mathbf{N}(u,v)$ denotes the normal orientation at pixel (u,v).

Normal Consistency Loss. To quantify geometric uncertainty, we employ an edge-aware **N**ormal **C**onsistency **L**oss (NCL), that encourages smoothness in the local surface by penalizing inconsistencies between neighboring pixels:

$$\begin{aligned} \psi_x(u,v) &= \|2\mathbf{N}(u,v) - \mathbf{N}(u+1,v) - \mathbf{N}(u-1,v)\|_2^2, \\ \psi_y(u,v) &= \|2\mathbf{N}(u,v) - \mathbf{N}(u,v+1) - \mathbf{N}(u,v-1)\|_2^2, \end{aligned} \tag{10}$$

where the smoothness terms $\psi_x(u,v), \psi_y(u,v)$ enforce horizontal and vertical consistency in the surface normal field. The total normal consistency loss is then given by:

$$\mathcal{L}_{\text{NCL}}(u,v) = (\psi_x(u,v) + \psi_y(u,v)) \cdot \exp(-\|\nabla \mathbf{I}(u,v)\|_2), \tag{11}$$

where the edge-aware weighting term $\exp(-\|\nabla \mathbf{I}(u,v)\|_2) = \exp(-\sqrt{|\mathbf{S}_x * \mathbf{I}(u,v)|^2 + |\mathbf{S}_y * \mathbf{I}(u,v)|^2})$ preserves discontinuities at boundaries while enforcing smoothness in homogeneous regions. NCL encourages the model to learn a spatially consistent normal field, thereby reducing uncertainty stemming from unstable 3D representations.

Semantic Guidance. To enforce geometric–semantic coherence, we generate masks by integrating 2D bounding boxes with semantic predictions, ensuring synchronized focus regions for dual-branch uncertainty optimization. Let $\{\mathcal{B}_i\}_{i=1}^n$ denote detected bounding boxes with scores $\{s_i\}_{i=1}^n$ and semantic uncertainties $\{U_i\}_{i=1}^n$ derived from the Conjugate Focal Loss ($\mathcal{L}_{\text{CFL}}$ in Equ. 7). To ensure reliable supervision, we select boxes with low semantic uncertainty using an exponentially moving average threshold:

$$R = \{i | U_i \le \overline{U}\}, \ \overline{U} \leftarrow \beta \cdot \frac{\Sigma_{i=1}^n U_i}{n} + (1-\beta) \cdot \overline{U}, \tag{12}$$

where $\beta \in [0,1]$ is a moving average factor (set to 0.1 in default). We then construct the region mask as follows:

$$\mathcal{M}(u,v) = \max_{i \in R} s_i \cdot \mathbb{I}_{\text{inside}} \left(u, v \mid \mathcal{B}_i\right), \tag{13}$$

where $\mathbb{I}_{\text{inside}}(u, v \mid \mathcal{B}_i)$ is an indicator function that returns 1 if pixel (u,v) is inside $\mathcal{B}_i$, and 0 otherwise. This semantic-guided mask ensures that only low semantic-uncertainty regions contribute to the normal field constraint, enhancing the reliability of the geometric constraint.

5.3. Overall Objective

The overall objective integrates the conjugate focal loss with the normal consistency constraint into a unified framework, simultaneously addressing semantic and geometric uncertainties. This dual optimization enables a complementary feedback loop: low-uncertainty spatial location enhances the model's ability to perform precise semantic classification, while confident semantic predictions in turn improve spatial understanding. The procedure is as follows:

$$\min_\theta \ \Sigma_{x \in \mathrm{I}} \mathcal{L}_{\text{CFL}}(x) + \lambda \Sigma_{(u,v) \in \mathrm{I}} \mathcal{M}(u,v) \cdot \mathcal{L}_{\text{NCL}}(u,v), \tag{14}$$

where x iterates over all detected objects and (u,v) spans all pixels in the image I. $\mathcal{L}_{\text{CFL}}$, $\mathcal{L}_{\text{NCL}}$, and $\mathcal{M}$ are defined in Equ. 7, Equ. 11, and Equ. 13. $\lambda = 0.7$ by default.

Table 1. Comparisons with state-of-the-art methods on the KITTI-C *validation* set (severity level 5) with MonoFlex as the base model. We highlight the best and second results with **bold** and underline respectively.

Method	Reference	Noise			Blur			Weather				Digital			Avg
		Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Snow	Frost	Fog	Brit.	Contr.	Pixel	Sat.	
Car Category															
MonoFlex	CVPR'21	0.00	0.00	0.00	0.00	11.02	0.00	5.07	12.94	7.70	14.46	0.00	2.13	5.65	4.54
• TENT	ICLR'21	5.31	11.09	6.40	2.22	25.49	2.14	23.88	28.96	35.40	37.07	24.67	22.59	30.63	19.68
• EATA	ICML'22	5.44	12.12	4.67	2.75	25.66	2.45	24.77	28.99	35.67	36.95	24.33	22.47	34.10	20.03
• DeYO	ICLR'24	5.78	12.52	4.52	3.01	26.05	2.98	24.91	29.40	35.19	37.59	23.75	23.82	34.33	20.30
• MonoTTA	ECCV'24	5.93	13.34	4.05	3.35	28.10	3.21	25.86	29.10	36.43	37.18	25.90	25.01	33.89	20.87
• Ours	This paper	**7.30**	**15.40**	**9.36**	**4.34**	**30.23**	**6.89**	**29.09**	**29.76**	**38.38**	**37.72**	**29.35**	**25.88**	**34.97**	**22.97**
Pedestrian Category															
MonoFlex	CVPR'21	0.00	0.00	0.00	0.00	4.09	0.00	0.51	1.67	1.77	2.09	0.00	0.32	0.95	0.88
• TENT	ICLR'21	0.71	2.07	0.86	1.36	12.70	0.81	6.72	8.33	11.89	12.79	7.42	6.45	9.78	6.30
• EATA	ICML'22	0.98	2.15	0.76	1.44	12.23	0.80	6.77	8.67	11.60	13.62	7.53	6.47	10.33	6.41
• DeYO	ICLR'24	1.03	2.25	0.91	1.64	11.85	0.80	6.63	**9.09**	11.77	**13.99**	7.59	6.39	10.57	6.50
• MonoTTA	ECCV'24	1.77	2.88	0.34	1.78	12.38	0.82	7.03	9.02	12.31	13.11	7.75	7.10	11.08	6.72
• Ours	This paper	**1.89**	**3.08**	**1.54**	**1.86**	**13.53**	**1.75**	**7.68**	**9.09**	**12.66**	**13.99**	**7.81**	**7.27**	**11.27**	**7.19**
Cyclist Category															
MonoFlex	CVPR'21	0.00	0.00	0.00	0.00	0.24	0.00	2.14	2.33	1.72	4.41	0.00	0.00	0.00	0.83
• TENT	ICLR'21	0.06	0.14	0.04	0.04	4.55	0.93	6.63	8.23	11.94	15.16	7.72	1.85	2.81	4.62
• EATA	ICML'22	0.05	0.15	0.03	0.02	4.66	1.10	6.73	7.58	13.77	14.99	7.32	2.03	2.82	4.71
• DeYO	ICLR'24	0.06	0.19	0.03	0.03	4.91	1.08	6.48	6.91	**13.94**	14.67	7.57	1.79	2.82	4.65
• MonoTTA	ECCV'24	0.05	0.12	0.01	0.02	4.80	1.25	6.75	8.24	13.31	14.95	7.55	2.11	2.88	4.77
• Ours	This paper	**0.11**	**0.22**	**0.07**	**0.10**	**6.00**	**2.00**	**6.89**	**8.41**	13.58	**15.94**	**7.94**	**2.13**	**2.92**	**5.10**

Table 2. Comparisons with state-of-the-art methods on the KITTI-C *validation* set (severity level 5) with MonoGround as the base model. Due to the space limit, the complete results of three categories are provided in Tab. 7 of Appendix. B.

Method	Reference	Noise			Blur			Weather				Digital			Avg
		Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Snow	Frost	Fog	Brit.	Contr.	Pixel	Sat.	
Car Category															
MonoGround	CVPR'22	0.00	0.00	0.00	0.00	11.63	0.29	1.95	6.59	3.14	19.25	0.00	4.66	3.74	3.94
• TENT	ICLR'21	6.82	14.81	8.21	4.88	28.38	2.65	23.92	28.08	33.06	36.70	20.22	30.63	33.27	20.90
• EATA	ICML'22	7.12	15.26	8.81	5.09	29.08	2.52	24.18	28.03	33.43	36.78	21.61	30.50	33.42	21.22
• DeYO	ICLR'24	7.35	15.72	9.38	5.74	30.01	2.99	25.03	28.55	34.32	37.31	23.41	30.99	34.16	21.92
• MonoTTA	ECCV'24	7.88	16.73	10.35	5.97	31.19	3.06	25.24	28.99	34.85	37.82	25.00	31.61	34.79	22.57
• Ours	This paper	**9.72**	**18.88**	**12.74**	**7.24**	**33.02**	**5.24**	**28.50**	**30.73**	**37.27**	**39.40**	**28.34**	**33.22**	**37.24**	**24.73**

6. Experiments

6.1. Experimental Setup

For a fair comparison, we follow the identical evaluation pipeline with prior work [24], including baseline models, training recipes, and evaluation protocols.

Datasets. We conduct experiments on the KITTI [11] and nuScenes [5] datasets. For KITTI, we use the KITTI-C version, which includes 13 distinct data corruption types [14] and five severity levels per type. Results represent the average performance across three difficulty levels, *i.e*. Easy, Moderate, and Hard. Note that we also provide more results of different severity levels of KITTI in Appendix. B.1. For nuScenes, we adopt the front-view images and construct the Daytime, Night, Sunny, and Rainy scenarios via their scene descriptions following [29]. Based on these scenes, we define 4 real-world adaptation tasks, *i.e*., Daytime ↔ Night and Sunny ↔ Rainy. Following the MonoTTA, we transfer the nuScenes dataset into the KITTI format. More details are provided in Appendix. D.3.

Compared Methods. Based on two representative base models MonoFlex [55] and MonoGround [40], we compare our DUO with several state-of-the-art methods: EATA [36] identifies reliable samples during entropy training. DeYO [22] leverages probability variations under augmentations as an additional cue to enhance entropy optimization. MonoTTA [24] boosts probabilities for high-score classes while applying negative learning to low-score classes.

Implementation Details. We implement our method and other baselines in PyTorch [38]. We employ the Stochastic

Table 3. Comparison with baselines on **D**aytime $\leftrightarrow$ **N**ight and **S**unny $\leftrightarrow$ **R**ainy of nuScenes dataset, regarding $\mathrm{AP}_{3\mathrm{D}|\mathrm{R}_{40}}$.

Method	MonoFlex			MonoGround		
	D $\to$ N	N $\to$ D	Avg.	D $\to$ N	N $\to$ D	Avg.
Source model	1.53	2.75	2.14	6.97	1.09	4.03
• TENT	3.33	3.45	3.39	8.36	1.66	5.01
• DeYO	4.72	4.87	4.79	11.01	1.40	6.21
• MonoTTA	6.92	3.68	5.30	13.61	1.29	7.45
• Ours	**9.05**	**5.41**	**7.23**	**15.70**	**1.91**	**8.81**
Method	S $\to$ R	R $\to$ S	Avg.	S $\to$ R	R $\to$ S	Avg.
Source model	6.86	10.91	8.89	8.03	7.44	7.73
• TENT	8.53	11.61	10.07	9.94	8.71	9.33
• DeYO	9.33	12.04	10.68	10.54	9.10	9.82
• MonoTTA	9.47	12.55	11.01	10.89	8.90	9.90
• Ours	**11.54**	**13.21**	**12.38**	**12.88**	**9.54**	**11.21**

Gradient Descent (SGD) optimizer with the same learning rate as MonoTTA, a momentum of 0.9, and a batch size of 16 for KITTI, 4 for nuScenes. Parameters λ, α, γ are assigned default values of 0.7, 4, and 2, respectively.

Evaluation Protocols. We report experimental results using the Average Precision for 3D bounding boxes, denoted as $\mathrm{AP}_{3D|R_{40}}$ [43]. Results represent the mean values across three difficulty levels, with Intersection over Union thresholds set to 0.5 for Cars and 0.25 for both Ped. and Cyc.

6.2. Main Results

Evaluation on Corruption Shifts. We first compare our DUO with previous methods on the KITTI-C dataset at severity level 5. Due to space constraints, detailed comparisons with other severity levels are provided in the Appendix. B. The results, reported in Tab. 1 & 2, reveal several observations: 1) Under test-time distribution shifts, pretrained detectors experience significant performance degradation across all categories. 2) Existing TTA methods partially mitigate the adverse effects of distribution shifts in M3OD, but their performance remains suboptimal as they cannot address the dual uncertainty. 3) Our method consistently outperforms compared approaches across all categories and base models. Notably, the proposed DUO achieves the best or comparable performance under 13 types of corruptions, with performance gains of **+2.1** and **+2.2** $\mathrm{AP}_{3\mathrm{D}|\mathrm{R}_{40}}$ in the Car category for two different models. These observations validate the crucial role of our adaptation framework in addressing dual uncertainty in M3OD models and enhancing robustness against distribution shifts.

Evaluation on Real-World Scenario. We further evaluate different methods on four real-world scenarios, as shown in Tab. 3. The experimental results yield the following observations: 1) Under real-world shifts, the pre-trained model still suffers severe performance degradation. 2) For Sunny $\leftrightarrow$ Rainy adaptation tasks on MonoFlex, existing methods achieve only marginal improvements, whereas our method boosts performance by **+4.7** $\mathrm{AP}_{3\mathrm{D}|\mathrm{R}_{40}}$. 3) Our DUO brings significant performance improvement on two base models, maintaining the best performance across all four scenarios, further demonstrating its effectiveness and superiority.

Table 4. Effects of components in our method. We conduct ablation studies of the conjugate focal loss $\mathcal{L}_{\mathrm{CFL}}$, normal consistency loss $\mathcal{L}_{\mathrm{NCL}}$, and semantic guidance $\mathcal{M}$ on the KITTI-C validation set. "Src." denotes the source model without adaptation.

Src.	$\mathcal{L}_{\mathrm{CFL}}$	$\mathcal{L}_{\mathrm{NCL}}$	$\mathcal{M}$	MonoFlex				MonoGround			
				Car	Pedes.	Cyc.	Avg.	Car	Pedes.	Cyc.	Avg.
✔				4.54	0.88	0.83	2.08	3.94	1.79	0.52	2.08
	✔			20.98	6.60	4.32	10.63	22.89	8.86	2.51	11.42
		✔		12.38	4.63	2.82	6.61	15.69	6.40	1.91	8.00
	✔	✔		<u>22.17</u>	<u>6.99</u>	4.72	<u>11.29</u>	<u>24.29</u>	<u>9.27</u>	2.78	<u>12.11</u>
		✔	✔	16.49	6.23	<u>4.87</u>	9.20	19.68	7.98	<u>2.94</u>	10.20
	✔	✔	✔	**22.97**	**7.19**	**5.10**	**11.75**	**24.73**	**9.62**	**3.02**	**12.46**

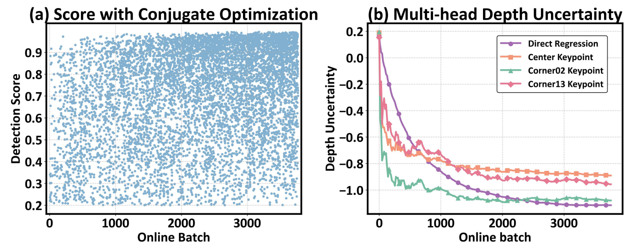

Figure 4. Effects of components in our method. (*a*) shows the distribution of detection scores during adaptation with conjugate focal loss. (*b*) records the variation of depth uncertainty in different heads during adaptation with normal field constraints.

6.3. Ablation Study

In this section, without loss of generality, we conduct ablation studies on Monoflex under Gaussian shift in KITTI-C for the sake of brevity. Focusing on two pivotal components of DUO, *Conjugate Focal Loss* and *Normal Field Constraint*, we perform extensive experiments to analyze their independent roles and complementary effects, gaining insights into key factors contributing to their effectiveness.

Effectiveness of Components. We investigate the impact of individual components by comparing the full method with variations that omit key parts. As shown in Tab. 4, incorporating either the conjugate focal loss or the normal field constraint significantly enhances detection performance, yielding average gains of $+8.9$ and $+7.6$ $\mathrm{AP}_{3\mathrm{D}|\mathrm{R}_{40}}$, respectively. Notably, using the normal consistency loss ($\mathcal{L}_{\mathrm{NCL}}$) alone results in unstable, marginal improvements; its effectiveness depends on being combined with $\mathcal{M}$, highlighting the necessity of semantic guidance for effective geometric uncertainty reduction. These findings underscore the effectiveness of each component.

Furthermore, a comparison of score distributions in Fig. 4(a) and Fig. 2(c) shows that the conjugate focal loss consistently boosts detection scores, ensuring effective training for challenging low-score objects. Similarly, Fig. 4(b) versus Fig. 2(d) demonstrates that our normal field constraint consistently reduces geometric uncertainty across all heads, preventing the model collapse observed in baseline meth-

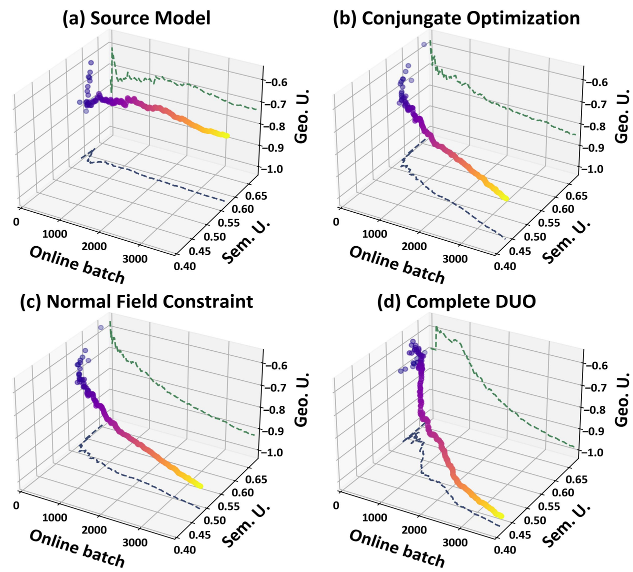

Figure 5. Impacts of different components on dual uncertainty. Colored lines in three-dimensional space show the cumulative average of dual uncertainty during the adaptation process. Green and blue lines in two-dimensional planes record geometric uncertainty and semantic uncertainty, respectively.

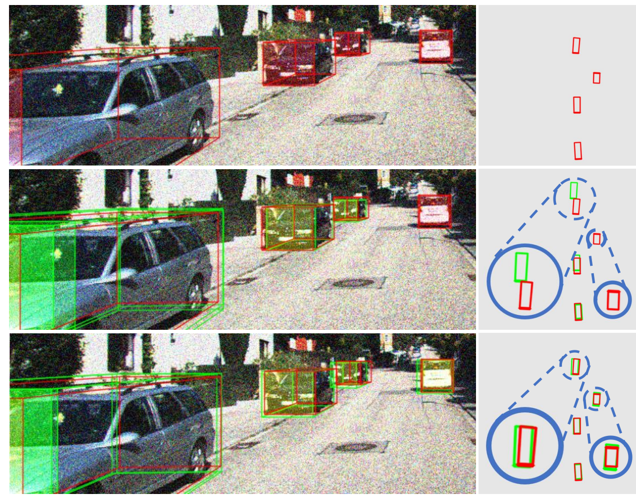

Figure 6. Qualitative examples on KITTI-C. In each row, we provide the front view (*left*) and the bird's-eye view (*right*) visualizations. Red represents the ground truth of boxes, while Green represents the predictions. We circle some objects to highlight differences in predictions. The first row corresponds to the source model, the second to MonoTTA, and the third to our method.

ods. Therefore, these innovations effectively overcome the limitations of existing methods identified in Sec. 4, leading to more robust and reliable detection.

Complementary Effects. As shown in Tab. 4, our complete method, which integrates all components, consistently achieves the best detection performance, demonstrating the compatibility of different components. To further investigate the complementary effects of our dual-branch design, we visualize the dual-uncertainty optimization process. Specifically, comparing Fig. 5(a)&(b), we observe that applying the conjugate focal loss significantly reduces semantic uncertainty, while interestingly, geometric uncertainty also shows a modest decline. A similar phenomenon is evident when employing the normal field constraint in Fig. 5(c). This synchronous behavior suggests an intrinsic interdependence between the two types of uncertainty, validating the potential of joint optimization.

Moreover, by combining two innovations in our DUO framework, it achieves the fastest and most pronounced decrease in two uncertainties compared to individual components, as shown in Fig. 5(d). These observations further validate that our dual-branch architecture creates a complementary loop for dual-uncertainty optimization, effectively harnessing their complementary effects.

Robustness and Efficiency. To validate the robustness and efficiency of our method, we extend our analysis by examining the sensitivity of key hyper-parameters and comparing running times in Appendix. C.

6.4. Qualitative Results

Based on the qualitative results shown in Fig. 6, our DUO framework produces predictions that are more precisely aligned with the ground-truth annotations. Compared with the latest SOTA method, DUO not only significantly improves the accuracy of 3D location estimation but also reduces missed detections for challenging distant or small-scale objects, which are typically prone to large geometric errors. These visualizations further confirm that our dual uncertainty optimization effectively adapts the source model, achieving precise and reliable 3D perception that closely matches the ground truth.

7. Conclusion and Future Work

In this paper, we propose a synergistic dual-branch framework designed to address semantic-geometric uncertainty inherent in 3D vision systems under test-time domain shifts. Our approach derives a novel conjugate loss to offer an adaptive, label-free weighting mechanism for balanced training on semantic uncertainty. Meanwhile, it incorporates a normal consistency constraint to reduce uncertainty from inconsistent geometric representation. Our extensive experiments demonstrate that the proposed dual-branch optimization creates a complementary loop, consistently improving performance across diverse domain shifts. In future work, we tend to expand our framework beyond M3OD to cover a broader range of 3D vision tasks. We hope that our work will deepen the understanding of uncertainty-aware model adaptation while providing transferable insights for related fields, such as unsupervised learning.

Acknowledgements

This work was supported by the Program of Beijing Municipal Science and Technology Commission Foundation (No.Z241100003524010), in part by the National Natural Science Foundation of China under Grant 62088102, in part by AI Joint Lab of Future Urban Infrastructure sponsored by Fuzhou Chengtou New Infrastructure Group and Boyun Vision Co. Ltd, and in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation.

References

[1] Eduardo Arnold, Omar Y Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 1

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 1

[3] Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*. Athena Scientific, 2003. 2, 4

[4] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022. 2

[5] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 6, 18

[6] Joya Chen, Qi Wu, Dong Liu, and Tong Xu. Foreground-background imbalance problem in deep object detectors: A review. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 285–290. IEEE, 2020. 3

[7] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 1

[8] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*, pages 1000–1001, 2020. 2

[9] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010. 4

[10] Junyu Gao, Xuan Yao, and Changsheng Xu. Fast-slow test-time adaptation for online vision-and-language navigation. In *Forty-first International Conference on Machine Learning*, 2024. 2

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2, 6

[12] Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35:6204–6218, 2022. 4

[13] Gustavo A Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Ismail Ben Ayed, and Christian Desrosiers. Clust3: Information invariant test-time training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6136–6145, 2023. 2

[14] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 1, 6, 18

[15] Julia Hornauer and Vasileios Belagiannis. Gradient-based uncertainty for monocular depth estimation. In *Proceedings of the European Conference on Computer Vision*, pages 613–630. Springer, 2022. 5

[16] Zixuan Hu, Yichun Hu, and Ling-Yu Duan. Seva: Leveraging single-step ensemble of vicinal augmentations for test-time adaptation. *arXiv preprint arXiv:2505.04087*, 2025. 2

[17] Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Ling-Yu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation. *arXiv preprint arXiv:2505.20704*, 2025. 1, 2

[18] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4012–4021, 2022. 2

[19] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021. 1

[20] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2): 358–367, 1988. 5

[21] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024. 2

[22] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 6, 17

[23] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, pages 1–34, 2024. 1

[24] Hongbin Lin, Yifan Zhang, Shuaicheng Niu, Shuguang Cui, and Zhen Li. Monotta: Fully test-time adaptation for monocular 3d object detection. In *Proceedings of the European Conference on Computer Vision*, pages 96–114. Springer, 2025. 2, 6, 17

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 3, 4

[26] Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22952–22961, 2023. 2

[27] Weihuang Liu, Xi Shen, Haolun Li, Xiuli Bi, Bo Liu, Chi-Man Pun, and Xiaodong Cun. Depth-aware test-time training for zero-shot video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19218–19227, 2024. 2

[28] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15641–15650, 2021. 2

[29] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 6, 18

[30] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 1

[31] Yueru Luo, Chaoda Zheng, Xu Yan, Tang Kun, Chao Zheng, Shuguang Cui, and Zhen Li. Latr: 3d lane detection from monocular images with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7941–7952, 2023. 1

[32] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4721–4730, 2021. 2, 3, 17

[33] Xinzhu Ma, Wanli Ouyang, Andrea Simonelli, and Elisa Ricci. 3d object detection from images for autonomous driving: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[34] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, and Xinge Zhu. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[35] Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of sgd. In *International Conference on Machine Learning*, pages 15959–15975. PMLR, 2022. 4

[36] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 6, 17

[37] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023. 2, 3

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 2019. 6

[39] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3227–3237, 2020. 1, 5

[40] Zequn Qin and Xi Li. Monoground: Detecting monocular 3d objects from the ground. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, 2022. 2, 6, 17

[41] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 2

[42] Prabhu Teja S and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9613–9623, 2021. 1

[43] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1991–1999, 2019. 7

[44] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 2

[45] Hugo Touchette. Legendre-fenchel transforms in a nutshell. *URL http://www. maths. qmul. ac. uk/ht/archive/lfth2. pdf*, page 25, 2005. 2, 4

[46] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2020. 1, 3, 17

[47] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 454–463, 2021. 2

[48] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 2, 3

[49] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020. 1

[50] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9092–9101, 2021. 1

[51] Yanshuo Wang, Ali Cheraghian, Zeeshan Hayder, Jie Hong, Sameera Ramasinghe, Shafin Rahman, David Ahmedt-Aristizabal, Xuesong Li, Lars Petersson, and Mehrtash Harandi. Backpropagation-free network for 3d test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23231–23241, 2024. 2

[52] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6814–6824, 2023. 1

[53] Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, and Yihua Tan. Monocd: Monocular 3d object detection with complementary depths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10257, 2024. 2, 17

[54] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 17

[55] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 2, 6, 17

[56] Bowen Zhao, Chen Chen, and Shu-Tao Xia. DELTA: DEGRADATION-FREE FULLY TEST-TIME ADAPTATION. In *The Eleventh International Conference on Learning Representations*, 2023. 2