MonoTTA: Fully Test-Time Adaptation for Monocular 3D Object Detection

Hongbin Lin[1,2*], Yifan Zhang[3,4*], Shuaicheng Niu[5*],
Shuguang Cui[2,1], Zhen Li[2,1†]

[1]FNii-Shenzhen, [2] SSE, CUHK-Shenzhen, [3] NUS, [4] Skywork AI, [5] NTU
{hongbinlin@link.,shuguangcui@,lizhen@}cuhk.edu.cn,
yifan.zhang@u.nus.edu, shuaicheng.niu@ntu.edu.sg

Abstract. Monocular 3D object detection (Mono 3Det) aims to identify 3D objects from a single RGB image. However, existing methods often assume training and test data follow the same distribution, which may not hold in real-world test scenarios. To address the out-of-distribution (OOD) problems, we explore a new adaptation paradigm for Mono 3Det, termed **Fully Test-time Adaptation** which aims to adapt a well-trained model to unlabeled test data by handling potential data distribution shifts at test time. However, applying this paradigm in Mono 3Det poses significant challenges due to OOD test data causing a remarkable decline in object detection scores. This decline conflicts with the pre-defined score thresholds of existing detection methods, leading to severe object omissions (*i.e.*, rare positive detections and many false negatives). Consequently, the limited positive detection and plenty of noisy predictions cause test-time adaptation to fail in Mono 3Det. To handle this problem, we propose a novel **Mono**cular **T**est-**T**ime **A**daptation (**MonoTTA**) method, based on two new strategies. 1) Reliability-driven adaptation: we empirically find that *high-score objects are still reliable* and the optimization of high-score objects can *enhance confidence across all detections*. Thus, we devise a self-adaptive strategy to identify reliable objects for model adaptation, which discovers potential objects and alleviates omissions. 2) Noise-guard adaptation: since high-score objects may be scarce, we develop a negative regularization term to exploit the numerous low-score objects via negative learning, preventing overfitting to noise and trivial solutions. Experimental results show that MonoTTA brings significant performance gains for Mono 3Det models in OOD test scenarios, approximately 190% gains by average on KITTI and 198% gains on nuScenes. The source code is now available at *Hongbin98/MonoTTA*.

Keywords: Test-time Adaptation · Monocular 3D Object Detection

1 Introduction

Three-dimensional (3D) Object Detection is a significant computer vision task, with the objective of identifying objects and determining their spatial and di-

[*] Authors contributed equally.

[†] Corresponding authors.

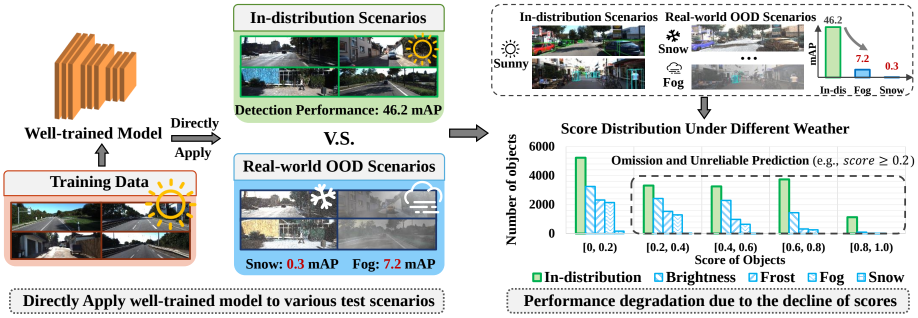
Well-trained Model
Training Data
Directly
Apply
In-distribution Scenarios
Detection Performance: 46.2 mAP
V.S.
Real-world OOD Scenarios
Snow: 0.3 mAP Fog: 7.2 mAP
Directly Apply well-trained model to various test scenarios
In-distribution Scenarios
Real-world OOD Scenarios
Sunny
Snow
Fog
mAP
46.2
7.2
0.3
In-dis
Fog
Snow
Score Distribution Under Different Weather
Number of objects
6000
4000
2000
0
Omission and Unreliable Prediction (e.g., score ≥ 0.2)
[0, 0.2)
[0.2, 0.4)
[0.4, 0.6)
[0.6, 0.8)
[0.8, 1.0)
Score of Objects
In-distribution
Brightness
Frost
Fog
Snow
Performance degradation due to the decline of scores

Fig. 1: An illustration of the generalizability issue of Mono 3Det models. Compared with in-distribution (In-dis) scenarios (*e.g.,* sunny), the detection scores within out-of-distribution (OOD) test data suffer severe degradation when the well-trained model (MonoFlex [48]) is directly applied to test scenarios affected by common natural disruptions, like weather changes (*e.g.,* snow and fog). Since existing Mono 3Det methods mainly adopt a pre-defined score threshold (*e.g.,* 0.2) for object detection, it leads to severe omissions and unreliable detections, thereby suffering serious performance degradation. Note that test images are the same but under different weather conditions.

mensional attributes through diverse sensor inputs [3,5,16,37,38]. To reduce the cost of sensors, there is an increasing trend towards implementing autonomous driving systems via Monocular 3D Object Detection (Mono 3Det) [2,42], where only one single RGB image and the camera calibration information are given. Even if this practical task is challenging, Mono 3Det methods have achieved promising results across various tasks and datasets [4,22,29,40,48]. Behind the success, a common presupposition is assuming that test images have the same distribution as the training images. However, this assumption could be possibly invalidated in many real-world scenarios due to prevalent natural corruptions such as weather changes, diminished sharpness, and other factors that introduce noise and contribute to uncalibrated cameras. In such circumstances, the well-trained model often suffers substantial performance degradation as a consequence of the *data distributional shifts* between the training images and the unlabeled test images. As shown in Fig. 1, the model performance degrades from 46.2 mAP in in-distribution data to 0.3 mAP in Snow and 7.2 mAP in Fog. Considering the widespread application of Mono 3Det in autonomous driving, the severe performance degradation due to out-of-distribution (OOD) test data may lead to unexpected traffic accidents and pose serious safety risks. Therefore, it is crucial to deal with the OOD generalization problem for Mono 3Det.

In addressing the OOD challenges specifically in test scenarios, one paradigm that has emerged as highly promising and gaining traction is *Test-Time Adaptation* (TTA), which seeks to tackle data distribution shifts by adapting a well-trained model to unlabeled test images in real time. Test-time training (TTT) [34] represents an initial approach of TTA in classification tasks, by adjusting the well-trained model to predict rotations through additional model training, while its computation demands at the adaptation stage are prohibitive in Mono 3Det applications, particularly in autonomous driving. To enhance efficiency, Tent [36] and EATA [26] have been developed for *Fully Test-Time Adaptation* (Fully TTA) where only unlabeled test images and a well-trained model are provided. Besides, Ev-TTA [12] and SOD [35] devise TTA methods to handle the event-based object recognition and weakly supervised salient object detection, respectively. Considering the constraints on time of Mono 3Det, we explore the fully TTA paradigm which seeks to deal with OOD test data in real time.

To investigate this paradigm for Mono 3Det, we dig into the detection outcomes for objects within test scenarios with variations or corruptions that are commonly caused by weather or cameras. Specifically, we directly apply the well-trained model to the validation set of KITTI which has been artificially injected with four distinct types of weather-related corruptions, namely Brightness, Frost, Fog, and Snow. Subsequently, we plot their distributions of detection scores (c.f. Fig. 1). It is observed that the detection scores of test objects tend to *markedly decline* as well as the *high-score objects are scarce* in the extreme scenario (Snow) when the well-trained model is directly applied to the scenarios with corruptions. This phenomenon indicates that: 1) The pre-trained Mono 3Det model struggles to discriminate between objects and the background within OOD test data, presenting as quantities of *omissions* and *unconfident detections*. 2) Directly applying existing fully TTA methods to Mono 3Det could only get suboptimal performance since they struggle to optimize the model *without enough high-score (positive) detections*, especially in certain extreme scenarios.

To handle it in Mono 3Det, we propose a **Mono**cular **T**est-**T**ime **A**daptation (**MonoTTA**) method, consisting of the reliability-driven adaptation and noise-guard adaptation strategies: 1) Reliability-driven adaptation. Specifically, data distribution shifts lead to omissions and noisy detections while our empirical analysis suggests that *high detection score objects are still reliable* (c.f. Fig. 3 **(a)**). Moreover, even if we only optimize the model via high-score objects (*e.g.,* $\geq$0.5), both the numbers of low-score and high-score objects increase (c.f. Fig. 3 **(b)**). These investigations motivate us that exploiting high-score objects rather than all objects for model adaptation would be a more reliable way to alleviate data distribution shifts and discover potential objects. Hence, we develop a self-adaptive strategy for the identification of reliable high-score objects in test images and devise the adaptive optimization loss $\mathcal{L}_{AO}$ to exploit the reliable subset for model adaptation, alleviating the detection score decline issue of OOD test data and digging out more potential objects. 2) Noise-guard adaptation. In addition, data distribution shifts may also result in a scarcity of high-score objects, *i.e.*, the majority of objects presenting low scores as the 'Snow' scenario in Fig. 3 **(a)**. To this end, we develop a negative regularization term to make rational use of the numerous low-scoring objects in the Negative Learning manner [13]. On the one hand, the negative regularization term $\mathcal{L}_{Nreg}$ allows the model to conduct adaptation via numerous noisy low-scoring objects. Thus, the model can achieve more high-score objects after alleviating distribution shifts. On the other hand, this term also prevents the model from overfitting to noise and trivial solutions, *i.e.*, assigning all classes of one object with high scores.

We summarize the main contributions as follows:

- To the best of our knowledge, we are the first to explore Fully Test-Time Adaptation to address OOD generalization problems for Mono 3Det. We show that the explored novel paradigm can bring significant improvements to Mono 3Det models in OOD test scenarios, *e.g.,* **137%** and **244%** average performance gains across 13 types of OOD shifts on KITTI.
- Our empirical investigation reveals an important insight that high-score objects maintain their reliability amidst various corruptions, while optimizing these high-score objects significantly boosts model confidence across all detections. This motivates the first effective test-time adaptation method (*i.e.*, our MonoTTA) in Mono 3Det.
- Extensive experiments on 13 types of corruptions of KITTI and 2 real scenarios (daytime ↔ night) of nuScenes demonstrate the effectiveness of our MonoTTA in boosting existing Mono 3Det methods [29,48] to handle test-time OOD problems. Even for instance-level methods [40], MonoTTA also maintains sufficient improvement, which further confirms its applicability.

2 Related Work

We first review the literature on Monocular 3D Object Detection, and then discuss Source-free Domain Adaptation and Test-Time Adaptation methods. More discussions on Unsupervised Domain Adaptation [18,41,47] are in Appendix ***A***.
Monocular 3D Object Detection aims to perceive 3D objects from a single 2D image. Existing Mono 3Det methods could be divided into two groups according to the use of extra information. On the one hand, some existing methods leverage extra pre-trained depth estimation modules [6,39,49] to solve one of the most difficult problems in Mono 3Det, *i.e.*, depth estimation from a single image. Other methods utilize LiDAR information, *e.g.,* generating pseudo-LiDAR [23,31,37]. It is worth noting that Monoground [29] proposes to introduce the ground plane as prior information, and MonoNeRD [40] proposes to utilize scene geometric clues to enhance the detector's performance in the implicit reconstruction manner. On the other hand, some Mono 3Det methods try to detect 3D objects without extra data. For example, SMOKE [20] proposes to detect 3D objects as the key points estimation task. Then, Monoflex [48] improves this idea by providing a flexible definition of object centers, which unifies the centers of regular and truncated objects. GrooMeD-NMS [15] proposes a grouped mathematically differentiable Non-Maximal Suppression for Mono 3Det.
Source-free Domain Adaptation (SFDA) aims to adapt the pre-trained source model to an unlabeled target domain without using the source data due to privacy issues [17,30]. SF-UDA^3D [32] first explores the SFDA framework to adapt the PointRCNN 3D detector to target domains, which consists of pseudo-labeling, reversible scale-transformations and motion coherency. Recently, the authors [10] seek to exploit the source model more reliably and propose an uncertainty-aware teacher-student framework to filter incorrect pseudo labels during model adaptation, alleviating the negative impact of label noise.

Nonetheless, SFDA assumes all target data to be known in advance and makes predictions after multiple epochs of optimization, which may not be viable for real-time applications due to computational or time constraints.

Test-Time Adaptation (TTA) seeks to improve model performance on test data via model adaptation through test samples even if data shifts exist. Early TTA methods [19, 34] endeavor to conduct additional model optimization on training data by self-supervised objectives, and then adapt the well-trained model to the test data via self-supervised objectives. However, in Mono 3Det applications like autonomous driving, the computation demands of such methods are prohibitive. To solve this, *Fully Test-Time Adaptation* methods are developed to adapt the well-trained model, where only unlabeled test images are available. Specifically, certain methods [25, 27, 33] tackle data distribution shifts by adapting the batch normalization layer statistics, while others alleviate this issue either by the entropy minimization of test data [8, 36] or maximizing the prediction consistency of different augmentations [43–46]. As for object detection tasks, Ev-TTA [12] and SOD [35] try to handle the event-based object recognition and weakly supervised salient object detection offline, respectively.

However, existing fully TTA methods struggle to optimize the model and solve distribution shifts in Mono 3Det due to numerous false negative detections. To the best of our knowledge, our MonoTTA stands as the first fully TTA method that handles distribution shifts for Mono 3Det models in real time.

3 Monocular Test-Time Adaptation

3.1 Problem Statement

Without loss of generality, we denote the pre-trained (or well-trained) model as $f_{\Theta_0}(\mathbf{x})$, which is achieved via training on labeled training images $\{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^N$. The training images follow the training distribution $P(\mathbf{x})$ (*i.e.*, $\mathbf{x}^s \sim P(\mathbf{x})$). Here, Θ_0 represents the parameters of the pre-trained model and N is the number of training data. During the training stage, the model is optimized to fit (or overfit) the training data. Then, at the test stage, the model will be able to perform well if the unlabeled test images $\mathcal{D}_t{=}\{\mathbf{x}_i\}_{i=1}^{N_t}$ follows the identical data distribution, *i.e.*, $\mathbf{x} \sim P(\mathbf{x})$ where N_t is the total number of test images. However, in real applications, it is possible for the pre-trained model to encounter Out-Of-Distribution (OOD) test samples due to prevalent natural corruptions, namely distribution shifts, *i.e.*, $\mathbf{x} \sim Q(\mathbf{x})$ and $P(\mathbf{x}) \neq Q(\mathbf{x})$.

To address this issue, fully test-time adaptation [36] seeks to tackle distribution shifts by adapting the pre-trained model $f_{\Theta_0}(\mathbf{x})$ to unlabeled test images $\{\mathbf{x}_i\}_{i=1}^{N_t}$ in real time. To achieve this goal, existing methods typically endeavor to update the model through the minimization of unsupervised objectives defined on test samples by $\min_{\hat{\Theta}} \mathcal{L}(\mathbf{x}; \Theta)$, where $\mathbf{x} \sim Q(\mathbf{x})$ and $\hat{\Theta} \subseteq \Theta$. Here, $\hat{\Theta}$ denotes the subset of model parameters that should be updated (*i.e.*, *batch normalization layers* following existing methods [26, 36]). Most existing fully TTA methods focus on classification tasks, heavily relying on sufficient positive predictions for

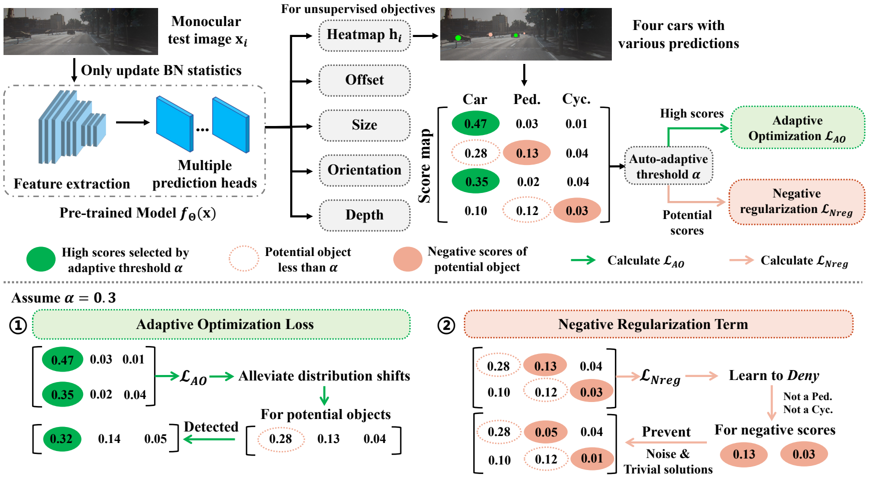
Monocular test image x_i
Only update BN statistics
Feature extraction
Multiple prediction heads
Pre-trained Model $f_\Theta(x)$
For unsupervised objectives
Heatmap h_i
Offset
Size
Orientation
Depth
Four cars with various predictions
Score map
Car
Ped.
Cyc.
0.47 0.03 0.01
0.28 0.13 0.04
0.35 0.02 0.04
0.10 0.12 0.03
Auto-adaptive threshold α
High scores
Potential scores
Adaptive Optimization $\mathcal{L}_{AO}$
Negative regularization $\mathcal{L}_{Nreg}$
High scores selected by adaptive threshold α
Potential object less than α
Negative scores of potential object
Calculate $\mathcal{L}_{AO}$
Calculate $\mathcal{L}_{Nreg}$
Assume $\alpha = 0.3$
① Adaptive Optimization Loss
0.47 0.03 0.01
0.35 0.02 0.04
$\mathcal{L}_{AO}$
Alleviate distribution shifts
For potential objects
0.32 0.14 0.05
Detected
0.28 0.13 0.04
② Negative Regularization Term
0.28 0.13 0.04
0.10 0.12 0.03
$\mathcal{L}_{Nreg}$
Learn to *Deny*
Not a Ped.
Not a Cyc.
0.28 0.05 0.04
0.10 0.12 0.01
Prevent
Noise & Trivial solutions
For negative scores
0.13 0.03

Fig. 2: An illustration of our MonoTTA. During the test phase, only the pre-trained model $f_{\Theta_0}(\mathbf{x})$ and unlabeled test images $\{\mathbf{x}_i\}_{i=1}^{N_t}$ are given. To conduct model adaptation, we initialize the model $f_\Theta(\mathbf{x})$ by Θ_0 and *only update the parameters of batch normalization layers*. When a batch of test images arrives, we first compute test object scores and refine the adaptive threshold α to select the reliable high-score objects, thereby optimizing Θ via the adaptive optimization loss $\mathcal{L}_{AO}$. Meanwhile, we devise a negative regularization term $\mathcal{L}_{Nreg}$ to facilitate the model to avoid overfitting to noise and trivial solutions. Here, Ped. and Cyc. represent Pedestrian and Cyclist in KITTI.

model adaptation. Nevertheless, there is a significant difference between conventional classification tasks and Mono 3Det. As previously indicated, the detection scores of test images $\mathbf{x}$ derived from $f_{\Theta_0}(\mathbf{x})$ are prone to markedly decrease in the presence of corruptions as shown in Fig. 1, leading to severe omissions (numerous false negatives) in Mono 3Det. In such circumstances, the scarcity of positive detections presents a significant challenge for model adaptation to test distributions while adapting the model with unreliable low-score detections may significantly introduce the noise. Therefore, existing fully TTA methods tend to fail in the OOD generalization problems of Mono 3Det.

3.2 Overall Scheme

After thoroughly examining the characteristics and challenges of Mono 3Det, we introduce a Monocular Test-Time Adaptation (MonoTTA) method to address the OOD problems for Mono 3Det models, which seeks to solve the object score declining issue within unlabeled OOD test data. As shown in Fig. 2, MonoTTA consists of two strategies: 1) Reliability-driven adaptation and 2) Noise-guard adaptation. We first briefly introduce the two strategies below.

First, we develop a reliability-driven adaptation strategy (c.f. Section 3.3) to conduct reliable model adaptation for OOD test data based on dependable test objects. Our empirical investigations inspire us to exploit those relatively reliable

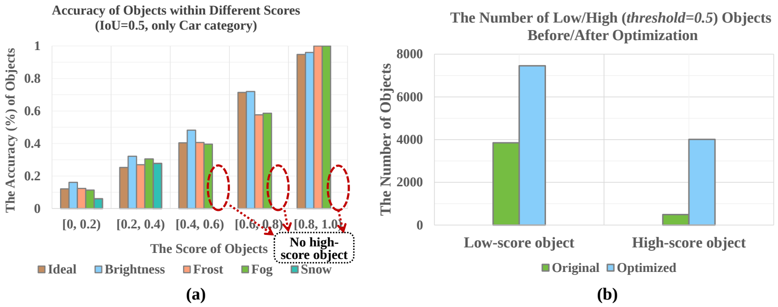

Fig. 3: Based on MonoGround [29], we conduct two empirical studies (Car, KITTI), with the 3D IoU threshold of 0.5. (a) We visualize the accuracy of the objects across varied scoring ranges, which shows that the accuracy of objects with high scores remains relatively stable even in the presence of diverse corruptions (Ideal means in-distribution scenarios). (b) We visualize the number of low & high-score objects before and after optimization. Although only high-score objects are optimized, the model treats low-score objects with more confidence.

test objects for alleviating distribution shifts, thereby discovering more potential objects. To this end, MonoTTA excludes unreliable test objects involving an adaptive threshold α for any unlabeled test data. Subsequently, the model is optimized by the adaptive optimization loss $\mathcal{L}_{AO}$ via the selected reliable objects.

Second, we tend to utilize plenty of low-score objects to adapt the model in an indirect manner instead of directly optimizing the model since low-score objects are noisy. Hence, we devise a noise-guard adaptation strategy (c.f. Section 3.4) to prevent the model from overfitting to noisy predictions and falling into trivial solutions. Specifically, we randomly choose one of the negative classes of low-score objects and minimize the scores (*e.g.,* score 0.03 of $[0.10, 0.12, 0.03]$ c.f. Fig. 2) after simply filtering out extremely low-score detections. Even though the positive class is noisy (*i.e.*, score 0.12), this term is capable of optimizing the model indirectly, *i.e.*, learn to deny the negative category of the object.

Overall, the training scheme of MonoTTA is as follows:

$$\min_{\hat{\Theta}} \mathcal{L}_{AO}(\hat{\Theta}) + \lambda \mathcal{L}_{Nreg}(\hat{\Theta}), \tag{1}$$

where λ is the balance hyper-parameter. The pseudo-code of MonoTTA is summarized in Algorithm 1.

3.3 Reliability-Driven Adaptation

To identify dependable test objects and conduct test-time model adaptation, we propose a reliability-driven adaptation strategy that consists of two components: 1) Reliable object identification and 2) Adaptive model optimization.

Reliable object identification. When unlabeled test images arrive, it is difficult for the pre-trained model to get accurate detections on OOD test data due to the decline in detection scores. To resolve this, we dig into the detection accuracy of the pre-trained model in test scenarios with various corruptions. Specifically, based on MonoGround [29], we visualize the detection precision for the Car category within the KITTI dataset across varied scoring ranges, with the 3D Intersection over Union (IoU) threshold of 0.5. As shown in Fig. 3 **(a)**, we find that high-score objects are more reliable and relatively stable even in the presence of diverse corruptions. Following this, we propose to select reliable high-score objects to conduct model adaptation via an adaptive threshold α. Specifically, we exploit the model $f_{\Theta}(\cdot)$ (initialized by Θ_0) to infer a batch of test images $\{\mathbf{x}_b\}_{b=1}^{B}$ and obtain the heatmap $\mathbf{h}_i$ of each image $\mathbf{x}_i$ by $\mathbf{h}_i = f_{\Theta}(\mathbf{x}_i)$, where i ranges from 0 to $B-1$ and B denotes the batch size. Then, we could achieve object score maps $\mathbf{s}_i \in \mathbb{R}^{B \times N_m}$ like the peaks in $\mathbf{h}_i$ [7] after normalization, where K and N_m denotes the number of classes and the maximum of detected objects, respectively. With the score map $\mathbf{s}_i$, we update the adaptive threshold α_t at the iteration t in the exponential moving average manner by:

$$\alpha_t = \begin{cases} \gamma, & \text{if } t = 1 \\ \beta \bar{m}_t + (1-\beta)\alpha_{t-1}, & \text{if } t > 1 \end{cases}, \tag{2}$$

$$\bar{m}_t = \frac{1}{B} \sum_{i=1}^{B} \frac{\sum_{j=1}^{N_m} s_{ij} \cdot \mathbb{I}(s_{ij} \geq \gamma)}{\sum_{j=1}^{N_m} \mathbb{I}(s_{ij} \geq \gamma)}. \tag{3}$$

In Eqn. (2), $\bar{m}_t$ denotes the average score of all detected objects in a single batch of B test samples at iteration t, while $\beta \in [0,1]$ is a decay coefficient. As for Eqn. (3), $s_{ij} \in (0,1)$ denotes the score of the j-th object in the i-th test image while $\mathbb{I}(\cdot)$ is the indicator function. Note that $\gamma \in \mathbb{R}^1$ is a pre-defined object detection threshold adopted from existing methods at their original inference stage as well as N_m.

Adaptive model optimization. As shown in Fig. 3 **(b)**, the optimization of high-score objects can also enhance the confidence of the model for relatively low-score objects. It motivates us that exploiting high-score objects rather than all objects for model adaptation would be a more reliable way to learn from OOD test data. Therefore, with the adaptive threshold α_t, we select the reliable subset of high scores from $\mathbf{s}_i$ and calculate the adaptive optimization loss $\mathcal{L}_{AO}$ to adapt the model by:

$$\mathcal{L}_{AO} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{N_m} \log(s_{ij} \cdot \mathbb{I}(s_{ij} \geq \alpha_t)), \tag{4}$$

The adaptive optimization loss $\mathcal{L}_{AO}$ alleviates the potential data distribution shifts in OOD test data by allowing the model to confidently identify high-score test objects. As we mentioned, it solves the score decline issue of OOD test data in a more reliable way rather than directly optimizing all test objects, thereby avoiding overfitting to noise and discovering more potential test objects.

Algorithm 1 The pipeline of the proposed MonoTTA

```
Require: Unlabeled test data $\mathcal{D}_t$=$\{\mathbf{x}_i\}_{i=1}^{N_t}$; Pre-trained model $f_{\Theta_0}(\mathbf{x})$; Batch size $B$;
    Parameters $\lambda$, $\beta$, $\eta$.
for a batch images $\{\mathbf{x}_b\}_{b=1}^{B}$ in $\mathcal{D}_t$ do
    Update the adaptive threshold $\alpha$ based on Eqn. (2);
    Calculate adaptive optimization loss $\mathcal{L}_{AO}$ based on Eqn. (4);
    Calculate negative regularization term $\mathcal{L}_{Nreg}$ based on Eqn. (6);
    Update $\hat{\Theta}$ by optimizing Eqn. (1)
end for
return Detection Results for all $\mathbf{x} \in \mathcal{D}_t$.
```

3.4 Noise-Guard Adaptation

Through the optimization of adaptive optimization loss $\mathcal{L}_{AO}$, the model is refined to yield more confident detection outcomes. Nevertheless, high-scoring objects may be scarce due to the distribution shifts, making the adaptation procedure difficult. Meanwhile, exclusive reliance on $\mathcal{L}_{AO}$ for adaptation may result in trivial solutions, whereby the model indiscriminately assigns high scores to all categories. Previous studies [13, 14] indicate that deep neural networks could learn from noisy pseudo labels in classification tasks through negative learning. Thus, MonoTTA proposes to learn from noisy low-score objects in a negative learning manner for Mono 3Det. Specifically, we denote $\hat{\mathbf{s}}_i \in \mathbb{R}^{B \times N_m \times K}$ as the multi-class score map for the test image $\mathbf{x}_i$, *i.e.* the multi-class score map $\hat{\mathbf{s}}_i$ contains not only the highest score of objects but also the scores for other classes. Here $\mathbf{s}_i = \arg\max_k \hat{\mathbf{s}}_{ik}$ where k is the class index. Employing a simple constant threshold η, we filter out extremely low-score objects and then randomly select a negative class $\bar{k}$ for each object. Subsequently, we compute the regularization loss e_k for each class k with relatively low object scores $s_{ij} \in [\eta, \alpha_t)$ by:

$$e_k = -\sum_{i=1}^{B} \sum_{j=1}^{N_m} \bar{y}_{ij} \log(1 - s_{ij\bar{k}} \cdot \mathbb{I}(\bar{k} = k)), \tag{5}$$

where $\bar{y}_{ij}$=$1 - s_{ij\bar{k}}$ is a constant weight. We further define n_k as the frequency of negative scores corresponding to class k in the test batch and balance e_k by:

$$\mathcal{L}_{Nreg} = \sum_{k=1}^{K} \frac{e_k}{n_k}. \tag{6}$$

As we mentioned before, detections with low and intermediate scores tend to be more noisy (c.f. Fig. 3 **(a)**), which is unreliable for direct model adaptation. To this end, we introduce the regularization term $\mathcal{L}_{Nreg}$ to leverage noisy low-score detections and improve the model for assimilating potentially accurate information (c.f. Term 2 in Fig. 2). Moreover, this term also prevents the model from trivial solutions, *i.e.*, indiscriminately assigning high scores to all classes of a single object (*e.g.,* TENT [36]). In addition, high-score objects may be absent

Table 1: Comparison with baselines on the KITTI-C validation set, severity **level 1** regarding Mean $AP_{3D|R_{40}}$. The **bold** number indicates the best result.

Car, IoU @ 0.7, 0.5, 0.5													
Method	Noise			Blur			Weather				Digital			Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Snow	Frost	Fog	Brit.	Contr.	Pixel	Sat.	
Monoflex [48]	3.84	7.48	5.31	2.59	3.73	11.05	0.23	7.77	7.57	24.87	6.92	28.16	31.46	10.84
• BN adaptation [33]	13.58	21.93	18.78	15.87	8.59	24.32	5.42	21.45	24.63	31.80	30.58	41.04	30.71	22.21
• TENT [36]	17.80	27.09	23.18	21.66	11.90	28.75	6.84	26.58	30.78	35.65	34.72	41.71	35.91	26.35
• EATA [26]	16.67	26.42	25.07	22.54	13.23	27.73	7.87	26.58	31.10	35.39	35.28	41.40	36.72	26.62
• ActMAD [24]	12.26	21.49	20.47	11.10	7.53	23.79	5.57	19.11	26.54	26.24	23.92	38.51	33.75	20.79
• MonoTTA (Ours)	**21.15**	**28.65**	**26.64**	**25.91**	**19.26**	**31.48**	**12.43**	**30.24**	**33.75**	**36.84**	**36.83**	**41.97**	**38.13**	**29.48**
MonoGround [29]	2.40	4.10	3.31	3.71	2.67	8.13	0.22	5.54	4.59	25.37	4.00	33.57	28.08	9.67
• BN adaptation [33]	13.49	23.52	19.69	16.33	7.61	23.99	7.98	20.71	24.00	31.34	29.03	43.06	32.99	22.60
• TENT [36]	17.94	29.60	19.90	23.45	13.90	29.39	10.32	26.65	33.35	35.96	36.39	43.35	37.79	27.54
• EATA [26]	16.03	26.08	18.08	20.28	12.40	27.37	9.22	23.79	29.49	33.65	32.58	43.61	36.00	25.28
• ActMAD [24]	13.65	22.39	20.56	14.99	8.17	21.09	9.13	18.77	19.52	28.34	25.32	42.63	31.58	21.24
• MonoTTA (Ours)	**26.13**	**33.11**	**28.60**	**30.38**	**25.48**	**32.44**	**18.72**	**32.60**	**37.75**	**37.87**	**39.57**	**43.67**	**37.98**	**32.64**
Pedestrian, IoU @ 0.5, 0.25, 0.25														
Method	Noise			Blur			Weather				Digital			Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Snow	Frost	Fog	Brit.	Contr.	Pixel	Sat.	
Monoflex	0.19	1.62	0.32	3.72	**8.47**	6.22	0.00	4.27	2.25	9.19	2.08	1.83	9.11	3.79
• BN adaptation [33]	6.21	8.20	9.20	7.83	5.35	7.52	2.89	6.47	9.24	9.12	9.93	12.73	9.76	8.03
• TENT [36]	6.02	7.96	9.57	7.75	6.06	8.63	2.63	6.71	9.91	10.26	**10.55**	12.33	10.27	8.36
• EATA [26]	6.05	7.96	**9.74**	**7.93**	6.06	8.01	2.48	6.24	9.94	9.70	10.02	12.41	10.12	8.21
• ActMAD [24]	2.41	4.27	4.68	1.46	1.62	6.48	1.17	4.25	4.84	6.42	6.19	8.29	7.07	4.55
• MonoTTA (Ours)	**6.54**	**8.41**	9.39	7.63	7.12	**8.99**	**3.25**	**7.64**	**10.26**	**10.55**	10.06	**13.28**	**10.66**	**8.75**
MonoGround	0.61	0.93	0.73	8.41	7.57	8.03	0.00	3.19	1.39	**13.82**	1.83	3.70	7.26	4.42
• BN adaptation [33]	5.09	7.08	7.77	7.63	6.63	9.45	2.24	6.72	8.40	9.72	11.06	16.70	12.43	8.53
• TENT [36]	7.27	10.10	10.02	8.53	8.30	11.03	3.54	8.73	9.11	11.74	12.12	**17.70**	**15.03**	10.25
• EATA [26]	5.92	8.05	8.58	8.12	7.59	10.95	3.31	7.91	10.10	11.02	10.95	17.30	14.34	9.55
• ActMAD [24]	3.96	5.75	6.48	5.61	5.73	7.14	1.58	5.37	5.74	7.27	8.10	13.12	9.27	6.55
• MonoTTA (Ours)	**8.58**	**11.18**	**11.79**	**9.22**	**9.40**	**13.20**	**4.83**	**9.95**	**14.46**	12.85	**13.25**	17.13	14.85	**11.59**
Cyclist, IoU @ 0.5, 0.25, 0.25														
Method	Noise			Blur			Weather				Digital			Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Snow	Frost	Fog	Brit.	Contr.	Pixel	Sat.	
Monoflex	0.28	1.64	0.47	0.59	4.97	3.60	0.00	7.42	3.81	**13.07**	3.79	3.80	8.39	3.99
• BN adaptation [33]	2.39	6.26	4.36	5.78	6.76	9.09	**1.70**	8.53	9.16	12.91	11.26	10.55	11.02	7.67
• TENT [36]	2.72	**7.94**	5.63	6.27	7.20	9.49	1.07	8.94	10.96	12.75	**12.72**	9.64	**11.28**	8.20
• EATA [26]	2.33	7.46	5.46	**7.19**	**7.23**	7.51	1.24	8.60	10.18	12.86	11.02	10.89	10.30	7.87
• ActMAD [24]	2.22	5.32	4.36	2.29	1.57	5.64	0.48	5.68	6.46	10.56	9.31	9.97	7.80	5.51
• MonoTTA (Ours)	**3.01**	7.24	**5.98**	7.00	6.09	**9.51**	1.45	**10.63**	**11.22**	12.85	11.85	**11.59**	10.81	**8.40**
MonoGround	0.12	0.46	0.48	0.33	0.72	2.03	0.00	0.56	0.35	5.55	0.52	2.08	5.24	1.42
• BN adaptation [33]	1.76	3.58	2.08	3.61	3.05	5.41	0.57	3.82	4.47	6.30	5.60	11.02	8.87	4.63
• TENT [36]	1.79	4.85	3.00	3.36	3.49	6.05	0.49	4.43	6.20	7.19	6.50	10.43	9.23	5.15
• EATA [26]	1.89	4.20	2.41	4.13	3.17	5.73	0.38	4.06	6.27	6.41	6.14	10.93	7.72	4.88
• ActMAD [24]	1.14	2.84	1.53	3.3	3.08	3.15	0.32	2.74	3.47	6.36	5.75	9.69	6.95	3.87
• MonoTTA (Ours)	**3.93**	**5.78**	**4.55**	**5.43**	**4.70**	**6.09**	**0.69**	**4.66**	**7.53**	**7.69**	**7.74**	**11.71**	**9.43**	**6.15**

in certain extreme scenarios. For instance, the Snow scenario (c.f. Fig. 3 **(a)**) lacks objects with scores exceeding 0.4 when the pre-trained model is directly applied. Under such a circumstance, $\mathcal{L}_{Nreg}$ plays a more important role in model adaptation since it can alleviate distribution shifts even only low-score objects, *i.e.*, deny the negative category. In other words, $\mathcal{L}_{Nreg}$ enables the model to alleviate distribution shifts and achieve more relatively high-score objects, thereby laying a crucial foundation for $\mathcal{L}_{AO}$ in extremely challenging scenarios.

4 Experiments

We conduct experiments based on KITTI [9] and nuScenes [1]. The results presented in this manuscript represent the *average value across three difficulty levels*, *i.e.*, Easy, Moderate, Hard. Note that we provide more results of higher severity levels of KITTI and more detailed results in Appendix ***D***.

Datasets. For KITTI, we adopt the protocol established by Monoflex [48] to split the training images into the training set (3712) and validation set (3769) to perform model training and adaptation, respectively. Following the official KITTI evaluation criteria, we evaluate detection results on three levels of difficulty. Then, we construct the KITTI-C dataset through the incorporation of 13 distinct types of data corruptions [11] to the validation set and each corruption

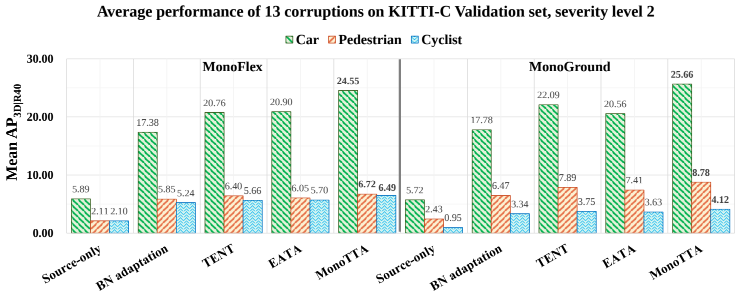

Fig. 4: We visualize the comparison with baselines on the KITTI-C validation set, severity **level 2** regarding Mean $AP_{3D|R_{40}}$. The **bold** number indicates the best result.

has 5 severity levels. The model is trained on the original training set and then tested on the KITTI-C validation set within one of the corruptions.

As for nuScenes, we adopt the front-view images and construct the Daytime and Night scenarios via their scene descriptions following [21]. Specifically, there exist 24.7k/5.4k train/test images in Daytime while 3.3k/0.6k train/test images in Night. Based on these splits, we construct two real-world adaptation tasks, *i.e.*, Daytime → Night and Night → Daytime. For simplification, we transfer the nuScenes dataset into the KITTI format and only consider the Car category. More details of data construction are provided in Appendix ***B***.

Implementation details. We implement our method and other baselines in PyTorch [28]. In MonoTTA, we conduct model adaptation based on the public pre-trained weights and the parameter settings provided by their authors [29, 40, 48]. Besides, we employ the Stochastic Gradient Descent (SGD) optimizer with a half learning rate of the initial rate used in base training over different methods, a momentum of 0.9 and a batch size of 16 for KITTI, 4 for nuScenes. Parameters λ, β, η are assigned default values of 1, 0.1, and 0.05, respectively. More training details of MonoTTA are provided in Appendix ***C***.

Compared methods. Based on three typical or state-of-the-art (SOTA) Mono 3Det methods [29, 40, 48], we fully compare MonoTTA with following methods: 1) source-only, *i.e.*, directly apply the pre-trained model to the test data within corruptions; 2) BN adaptation [33] updates batch normalization statistics via target data; 3) TENT [36] minimizes the entropy loss of test data; 4) EATA [26] identifies reliable samples to update the model by entropy loss minimization. Here we compare MonoTTA with its variant Efficient Test-time Adaptation. 5) ActMAD [24] matches activation statistics for the distribution alignment.

Evaluation protocols. In order to fully evaluate the proposed method, we report our experimental results in the Average Precision (AP) for 3D bounding boxes, denoted as $AP_{3D|R_{40}}$. The results present the mean values across three levels of difficulty and the Intersection over Union (IoU) thresholds are set to 0.7, 0.5, 0.5 for Cars and 0.5, 0.25, 0.25 for Pedestrians and Cyclists.

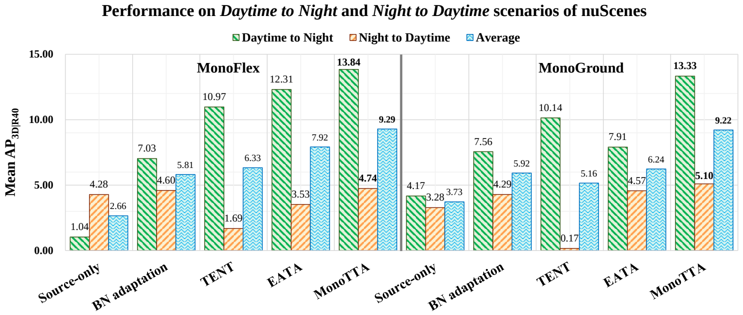

Fig. 5: Comparison with baselines on **D**aytime → **N**ight and **N**ight → **D**aytime of nuScenes, regarding Mean $AP_{3D|R_{40}}$. The **bold** number indicates the best result.

4.1 Comparisons with Previous Methods

We first compare our MonoTTA with previous methods in severity level 1 of KITTI-C. The results are reported in Table 1, which gives the following observations: 1) Due to distribution shifts, directly applying the pre-trained model to the test data (*i.e.*, source-only) suffers severe performance degradation in all categories. 2) Existing TTA methods are able to mitigate the negative effect of distribution shifts for Mono 3Det to some degree. However, they only achieve suboptimal performance since they tend to increase the scores of all positive detections, containing severe noise. 3) MonoTTA consistently outperforms all compared methods over all categories within various base models in terms of mean $AP_{3D|R_{40}}$. Specifically, MonoTTA achieves the best or comparable performance in all categories under all corruptions, attaining a large performance gain over TENT and EATA (*e.g.,* improving an average $AP_{3D|R_{40}}$ about 5.1 and 7.4 of the Car category based on MonoGround).

4.2 More Severe Corruption and Real Scenario

On the one hand, to fully validate the effectiveness of our MonoTTA, we visualize experimental results under more severe corruption conditions at severity level 2 as shown in Fig. 4, which clearly gives additional observations: 1) With the escalation of severity level, the pre-trained models suffer a larger performance decline within various corruptions, enlarging the difficulty of TTA. 2) The performance improvements of existing TTA methods become relatively limited, particularly in Pedestrian and Cyclist classes. 3) Even if the tasks are more challenging, MonoTTA still stably obtains the best average performance within all corruptions since $\mathcal{L}_{Nreg}$ plays an important role in alleviating distribution shifts for certain extreme scenarios, *i.e.*, only low-score objects exist.

On the other hand, we further validate different methods within real scenarios as shown in Fig. 5. The experimental results also give the following observations: 1) Under real corruptions, the pre-trained model still suffers severe performance degradation due to the data distribution shifts. 2) TENT tends to increase the

Table 2: Comparison with baselines based on the instance-level method (*i.e.*, batch size is 1) on the KITTI-C validation set, regarding Mean $AP_{3D|R_{40}}$ and the severity levels 1 and 2. The **bold** number indicates the best result.

Method	Level 1				Level 2			
	Car	Pedes.	Cyclist	Avg.	Car	Pedes.	Cyclist	Avg.
MonoNeRD	19.84	5.96	2.27	9.36	13.02	3.83	1.61	6.15
• BN adaptation [33]	30.73	8.85	3.81	14.46	26.47	6.94	2.91	12.11
• TENT [36]	35.72	9.99	**4.75**	16.82	31.85	7.81	**3.40**	14.35
• EATA [26]	34.60	10.04	4.20	16.28	30.66	7.86	3.28	13.93
• MonoTTA (Ours)	**37.40**	**10.39**	4.35	**17.38**	**33.99**	**8.25**	3.33	**15.19**

Table 3: Based on MonoGround [29], we conduct ablation studies of $\mathcal{L}_{AO}$ and $\mathcal{L}_{Nreg}$ on the KITTI-C validation set, regarding $AP_{3D|R_{40}}$.

Backbone	$\mathcal{L}_{AO}$	$\mathcal{L}_{Nreg}$	Level 1				Level 2			
			Car	Pedes.	Cyclist	Avg.	Car	Pedes.	Cyclist	Avg.
✓			9.67	4.42	1.42	5.17	5.72	2.43	0.95	3.03
✓	✓		27.94	10.17	5.10	14.40	21.27	7.79	3.82	10.96
✓		✓	23.26	8.60	4.60	12.15	18.93	6.81	3.55	9.77
✓	✓	✓	**32.64**	**11.59**	**6.15**	**16.79**	**25.66**	**8.78**	**4.12**	**12.85**

confidence of all positive detections and thus overfits to noise, *i.e.*, failing to handle the extremely challenging task N→D. 3) EATA still achieves sub-optimal performance while our MonoTTA brings sufficient average performance improvement on both MonoFlex (6.23 mAP) and MonoGround (8.26 mAP), maintaining the best performance in real scenarios . Detailed results are in Appendix ***D***.

4.3 Application to Instance-Level Inference Method

In this section, we seek to investigate whether the proposed MonoTTA can be used to effectively enhance Mono 3Det methods which only process a single image sequentially, *i.e.*, $B = 1$. To be specific, it may be crucial for Mono 3Det methods to make immediate decisions based on the most recent scene (image) in real-world scenarios like autonomous driving. To this end, it is essential for TTA methods devised for Mono 3Det to allow a single image as input and then conduct model adaptation. To validate it, we integrate MonoTTA into the SOTA Mono 3Det method namely MonoNeRD [40] which accepts one image each time at the test phase. As shown in Table 2, MonoTTA achieves the best or comparable performance across all categories at both severity levels 1 and 2, illustrating the applicability of our method to boost these approaches for handling OOD test data. Detailed results can be also found in Appendix ***D***.

4.4 Ablation Studies and Quantitative Results

Ablation studies. To examine the effectiveness of the losses in MonoTTA, we show the results of the models optimized by different losses. As shown in Table 3, introducing $\mathcal{L}_{AO}$ or $\mathcal{L}_{Nreg}$ enhances the model performance compared to directly applying the pre-trained model (*i.e.*, source only). On the one hand, such a result verifies that our strategy is able to alleviate the score decline issue in test OOD data with high-score objects. On the other hand, introducing $\mathcal{L}_{Nreg}$ could obtain a higher average $AP_{3D|R_{40}}$ value, which also verifies that the negative regularization term is able to enhance the pre-trained model even only

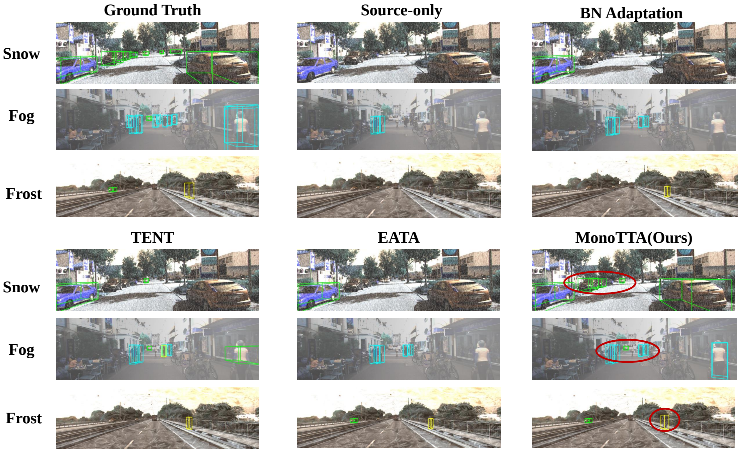

Fig. 6: Qualitative results of baselines and the proposed MonoTTA based on MonoGround [29]. We visualize the results on KITTI-C validation set, where predicted cars, pedestrians and cyclists are in lime green, sky blue and yellow, respectively.

with low-score objects. When combining the losses ($i.e.\mathcal{L}_{AO}$, $\mathcal{L}_{Nreg}$) together, we obtain the best performance.

Quantitative Results. We provide visualizations within Snow, Fog and Frost based on Monoground as shown in Fig. 6. It is evident that the source-only setting suffers severe omissions, while BN adaptation, TENT and EATA alleviate distribution shifts to some degree and give more detections. As for our MonoTTA, it can produce superior detections even in severe conditions, including fewer omissions and accurate detections as highlighted by red circles.

5 Conclusion

In this paper, we propose a monocular test-time adaptation method to improve the pre-trained model on the shifted test data for monocular 3D object detection. Specifically, our method consists of two strategies: 1) Reliability-driven adaptation. To discover more potential objects, we devise a self-adaptive strategy to identify reliable objects for adaptive model adaptation. 2) Noise-guard adaptation. To avoid overfitting to noise and trivial solutions, we devise the negative regularization term to mitigate the negative effects of noisy detections and alleviate distribution shifts. Experiments on KITTI-C and nuScenes datasets demonstrate the effectiveness of MonoTTA in handling fully test-time adaptation for monocular 3D object detection.

Future directions. 1) Our work focuses on 2D images, while future studies could explore 3D information in handling distribution shifts. 2) This work explores TTA by assuming one OOD distribution at a time, where the forgetting issue is not severe as the source model weights are recoverable. Exploring scenarios with dynamically OOD distributions offers a compelling future direction, where the forgetting issue would become more severe.

Acknowledgements

This work was supported by NSFC with Grant No. 62293482, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by Shenzhen General Program No. JCYJ20220530143600001, by Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, by the Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010 001), by the Guangdong Provincial Key Laboratory of Big Data Computing, CHUK-Shenzhen, by the NSFC 61931024&12326610, by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS2017072 51409055), and by Tencent & Huawei Open Fund.

References

1. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
2. Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., Urtasun, R.: Monocular 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2147–2156 (2016)
3. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A.G., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. Advances in neural information processing systems **28** (2015)
4. Chen, Y., Tai, L., Sun, K., Li, M.: Monopair: Monocular 3d object detection using pairwise spatial relationships. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12093–12102 (2020)
5. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21674–21683 (2023)
6. Ding, M., Huo, Y., Yi, H., Wang, Z., Shi, J., Lu, Z., Luo, P.: Learning depth-guided convolutions for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops. pp. 1000–1001 (2020)
7. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6569–6578 (2019)
8. Fleuret, F., et al.: Test time adaptation through perturbation robustness. In: NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications (2021)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3361. IEEE (2012)
10. Hegde, D., Kilic, V., Sindagi, V., Cooper, A.B., Foster, M., Patel, V.M.: Source-free unsupervised domain adaptation for 3d object detection in adverse weather.

In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 6973–6980. IEEE (2023)

11. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2018)
12. Kim, J., Hwang, I., Kim, Y.M.: Ev-tta: Test-time adaptation for event-based object recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17745–17754 (2022)
13. Kim, Y., Yim, J., Yun, J., Kim, J.: Nlnl: Negative learning for noisy labels. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 101–110 (2019)
14. Kim, Y., Yun, J., Shon, H., Kim, J.: Joint negative and positive learning for noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9442–9451 (2021)
15. Kumar, A., Brazil, G., Liu, X.: Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8973–8983 (2021)
16. Li, P., Chen, X., Shen, S.: Stereo r-cnn based 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7644–7652 (2019)
17. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International conference on machine learning. pp. 6028–6039. PMLR (2020)
18. Lin, H., Zhang, Y., Qiu, Z., Niu, S., Gan, C., Liu, Y., Tan, M.: Prototype-guided continual adaptation for class-incremental unsupervised domain adaptation. In: European Conference on Computer Vision. pp. 351–368 (2022)
19. Liu, Y., Kothari, P., Van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: Ttt++: When does self-supervised test-time training fail or thrive? In: Advances in Neural Information Processing Systems. vol. 34, pp. 21808–21820 (2021)
20. Liu, Z., Wu, Z., Tóth, R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 996–997 (2020)
21. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE international conference on robotics and automation (ICRA). pp. 2774–2781. IEEE (2023)
22. Luo, Y., Zheng, C., Yan, X., Kun, T., Zheng, C., Cui, S., Li, Z.: Latr: 3d lane detection from monocular images with transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7941–7952 (2023)
23. Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., Ouyang, W.: Rethinking pseudo-lidar representation. In: European Conference on Computer Vision (2020)
24. Mirza, M.J., Soneira, P.J., Lin, W., Kozinski, M., Possegger, H., Bischof, H.: Actmad: Activation matching to align distributions for test-time-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24152–24161 (2023)
25. Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., Snoek, J.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint arXiv:2006.10963 (2020)
26. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: The Internetional Conference on Machine Learning (2022)

27. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: Internetional Conference on Learning Representations (2023)
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. vol. 32 (2019)
29. Qin, Z., Li, X.: Monoground: Detecting monocular 3d objects from the ground. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3793–3802 (2022)
30. Qiu, Z., Zhang, Y., Lin, H., et al.: Source-free domain adaptation via avatar prototype generation and adaptation. In: International Joint Conference on Artificial Intelligence (2021)
31. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8555–8564 (2021)
32. Saltori, C., Lathuiliére, S., Sebe, N., Ricci, E., Galasso, F.: Sf-uda 3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection. In: 2020 International Conference on 3D Vision (3DV). pp. 771–780. IEEE (2020)
33. Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., Bethge, M.: Improving robustness against common corruptions by covariate shift adaptation. In: Advances in neural information processing systems. pp. 11539–11551 (2020)
34. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International conference on machine learning. pp. 9229–9248. PMLR (2020)
35. Veksler, O.: Test time adaptation with regularized loss for weakly supervised salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7360–7369 (2023)
36. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: The Internetional Conference on Machine Learning (2021)
37. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8445–8453 (2019)
38. Wu, H., Wen, C., Shi, S., Li, X., Wang, C.: Virtual sparse convolution for multimodal 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21653–21662 (2023)
39. Xu, B., Chen, Z.: Multi-level fusion based 3d object detection from monocular images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2345–2353 (2018)
40. Xu, J., Peng, L., Cheng, H., Li, H., Qian, W., Li, K., Wang, W., Cai, D.: Mononerd: Nerf-like representations for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6814–6824 (2023)
41. Yang, J., Shi, S., Wang, Z., Li, H., Qi, X.: St3d: Self-training for unsupervised domain adaptation on 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10368–10378 (2021)
42. Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., Ding, E.: Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21341–21350 (2022)

43. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. In: Advances in Neural Information Processing Systems. vol. 35, pp. 38629–38642 (2022)
44. Zhang, Y., Hooi, B., Hong, L., Feng, J.: Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In: Advances in Neural Information Processing Systems. vol. 35, pp. 34077–34090 (2022)
45. Zhang, Y., Hooi, B., Hu, D., Liang, J., Feng, J.: Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. In: Advances in Neural Information Processing Systems. vol. 34, pp. 29848–29860 (2021)
46. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 10795–10816 (2023)
47. Zhang, Y., Wei, Y., Wu, Q., Zhao, P., Niu, S., Huang, J., Tan, M.: Collaborative unsupervised domain adaptation for medical image diagnosis. IEEE Transactions on Image Processing **29**, 7834–7844 (2020)
48. Zhang, Y., Lu, J., Zhou, J.: Objects are different: Flexible monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3289–3298 (2021)
49. Zou, Z., Ye, X., Du, L., Cheng, X., Tan, X., Zhang, L., Feng, J., Xue, X., Ding, E.: The devil is in the task: Exploiting reciprocal appearance-localization features for monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2713–2722 (2021)