Improving Batch Normalization with Test-Time Adaptation for Robust Object Detection in Self-Driving

Dacheng Liao, Mengshi Qi*, Liang Liu*, Huadong Ma

State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications, China
{liaodacheng, qms , liangliu , mhd}@bupt.edu.cn

Abstract

In open real-world autonomous driving scenarios, challenges such as sensor failure and extreme weather hinder the generalization of current autonomous driving perception models to these unseen domain, due to the domain shifts between the test and training data. As the parameter scale of autonomous driving perception models grows, traditional test-time adaptation (TTA) methods become unstable and often degrade model performance in most scenarios. To address these challenges, this paper proposes two new robust methods to improve the Batch Normalization with TTA for object detection in autonomous driving: (1) We introduce a new LearnableBN layer based on Geometric Confidence Maximization and Entropy Minimization. Specifically, we modify the traditional BN layer by incorporating auxiliary learnable parameters, which enables the BN layer to dynamically update the statistics according to the different input data. (2) We propose a novel semantic-consistency based dual-stage adaptation strategy, which encourages the model to iteratively search for the optimal solution and eliminates unstable samples during the adaptation process. Extensive experiments on the NuScenes-C dataset shows that our method achieves a maximum improvement of about 10% using BEVFormer as the baseline across six corruption types and three levels of severity.

Code — https://github.com/Maodou-L/LearnableBN

Introduction

Autonomous driving perception models encounter significant challenges when the distribution of test data diverges from that of the training data, particularly in dynamic and open real-world driving scenarios such as extreme weather conditions or sensor failures, leading to severe degradation in the model's predictive accuracy (Zhu et al. 2023b; Bojarski et al. 2016; Wang et al. 2024), which is unacceptable for autonomous driving tasks. Traditional methods (Yun et al. 2019; DeVries and Taylor 2017; Zhang et al. 2017) for enhancing model robustness typically rely on extensive annotation costs or use data augmentation. However, these methods necessitate prior knowledge of the test data distribution, which is often unknown in real-world driving scenarios. To address these practical issues, a more viable approach is to use Test-Time Adaption (TTA) methods to adjust models promptly when facing unseen domains.

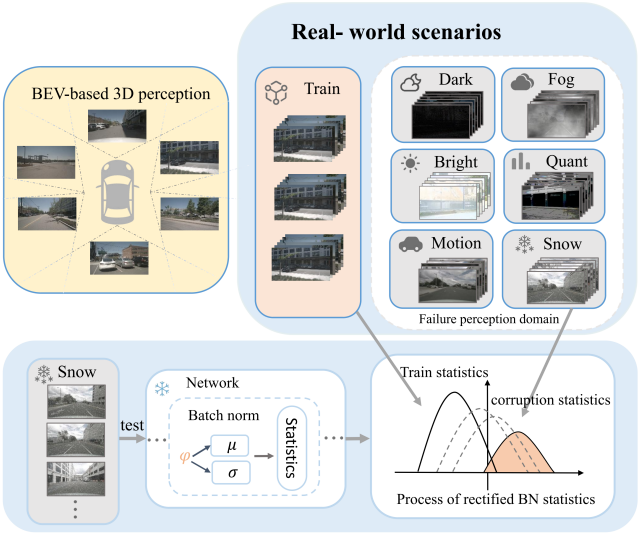


Figure 1: Illustration of the challenges faced by BEV-based 3D object detection models in unseen domains due to extreme weather. To improve robustness, the TTA method estimates BN statistics for unseen domains during testing.

The prevalent TTA paradigm (Boudiaf et al. 2022; Sun et al. 2020; Wang et al. 2020; Liu et al. 2021; Qi, Lv, and Ma 2025) typically addresses the issue of the distribution shifts between test and training data by adjusting the statistics of Batch Normalization (BN) layers, As shown in Fig 1. However, this TTA paradigm presenting the following challenges in self-driving (Liang, He, and Tan 2024; Qi et al. 2019):

Firstly, TTA methods that adjust Batch Normalization (BN) parameters exhibit significant instability in autonomous driving perception tasks, due to the BN layers employ an exponential moving average (EMA) approach to estimate the data distribution. The EMA method is highly sensitive to batch size, meanwhile the use of EMA for updating BN statistics is significantly affected by the problem of internal covariate shift in the model. If the prediction of the bottom BN layer's statistic is error, it can lead to the accumulation of errors in subsequent BN layers' prediction (Schnei-

*Corresponding authors: Mengshi Qi and Liang Liu.

der et al. 2020; Cicek and Soatto 2019; Xie et al. 2020). As model parameters and depth increase in autonomous driving perception tasks, the batch size is constrained, making it difficult for TTA methods to accurately predict the real test data distribution and worsening internal covariate shift.

Furthermore, most existing test-time adaptation (TTA) methods assume that the model outputs follow a probability distribution, enabling unsupervised methods such as entropy minimization and KL divergence (Wang et al. 2020; Vu et al. 2019; Sun et al. 2020; Zhang et al. 2023). However, current research lacks an unsupervised loss function that can be directly applied to regression tasks. To adopt similar strategies to regression tasks, the model's output must be modeled as a probability distribution with certain prior assumptions. This approach fundamentally modifies the model's output representation and consequently necessitates model retraining. In the object detection tasks, which involves both classification and regression branches, the regression branch does not naturally produce probabilistic outputs, making it difficult to directly apply existing TTA methods.

TTA methods typically classify test samples first and adjust the model using samples from specific categories (Li et al. 2023; Qi et al. 2021). This requires prior knowledge of the test data distribution. In real-world driving scenarios, the diversity of encountered scenes is often unknown, and the presence of noisy samples is prevalent (Niu et al. 2023; Qi et al. 2020).

To address these challenges, we introduce a learnableBN based on geometric confidence maximization and entropy minimization loss to adjust BN statistics. By adding auxiliary learnable parameters into BN layers, we predict the BN statistics for the test domain, replacing the EMA method. This approach addresses the limitations of BN layers under mini-batch conditions, mitigates model internal covariate shift issues and addresses the instability arising from adjusting BN statics. Additionally, by guiding the optimizing of auxiliary learnable parameters through unsupervised method, we introduce a new geometric confidence maximization loss to regression branch and entropy minimization loss to classification branch. Secondly, to tackle the challenges of TTA in real-world scenarios, we propose a semantic-consistency based dual-stage adaptation method. By adjusting the variation of learning rates and dividing adaptation into two stages, we use the semantic consistency of sample predictions in different stages as guidance to filter out the uncertain samples, thereby making the training process more stable and prevent the model from converging to a local optimum in the solution space.

Our main contributions can be summarized as follows:
(1) We propose a novel TTA paradigm for robust BEV perception in open real-world driving scenarios, by incorporating a LearnableBN for estimating BN statistics with geometric confidence maximization(GCM) and entropy minimization(EM) loss function that effectively addresses the instability issues inherent to traditional BN layers.
(2) We introduce a semantic-consistency based dual-stage adaptation method, which is designed to filters out the noisy samples and prevents the model from converging to a local optimum in the solution space.
(3) We conduct extensive experiments on widely-adopted benchmark, nuScenes-C, and results show that our proposed method achieves a maximum improvement of about 10% using BEVFormer as the baseline model across six corruption types and three levels of severity.

Related Work

Autonomous driving perception. Monocular 3D object detection (Zou et al. 2021; Ye et al. 2025; Lv et al. 2025) tackles depth estimation from a single image (Ding et al. 2020), often with pre-trained depth modules. SMOKE (Liu et al. 2021) formulates 3D detection as keypoint estimation, while Monoflex (Zhang, Lu, and Zhou 2021) improves performance via flexible object-center definitions for regular and truncated objects.Mainstream BEV-based detectors include object-query methods such as: DETR3D (Wang et al. 2022), which uses Transformer cross-attention to bypass depth estimation; PETR, which introduces 3D position-aware representations; BEVFormer (Li et al. 2022; Yang et al. 2023), which leverages temporal cross-attention and polar coordinates; and Sparse4D (Lin et al. 2022; Zhu et al. 2023a), which uses sparse proposals for feature fusion. We adopt BEVFormer, Sparse4D, and Monoflex as baselines.
Test-Time-Adaptation (TTA). Most TTA methods (Sun et al. 2020; Fleuret et al. 2021; Iwasawa and Matsuo 2021; Qi et al. 2025) adapt models on unlabeled test data during inference. Benz et al. (Benz et al. 2021) adjust BN statistics via forward passes, while Schneider et al. (Schneider et al. 2020) compute mixture coefficients dynamically for BN updates. TENT (Wang et al. 2020) introduces entropy minimization as the sole loss for unsupervised adaptation by optimizing affine BN parameters. Building on TENT, Domain Adaptor (Zhang et al. 2023; Deng, Qi, and Ma 2025) further uses EMA-based mixture coefficients and temperature scaling. However, these unsupervised losses do not apply to the regression branch of object detection.

Method

Problem Definition

In this work, we define the test dataset as $D_c^s = \{x_1, x_2, ..., x_n\}$, where c represents the different conditions in real-world driving scenarios, and s is the severity level of the domain shift between test domain and train domain. We define the model as $f(\cdot|\theta, \phi)$, where the θ is origin model's parameters. We introduce the set of auxiliary learnable parameters in the BN layers, defined as $\phi = \{\phi_1, \phi_2, ..., \phi_m\}$, where m corresponds to the parameters for the m-th BN layer. The learning rate of the model is defined as η.

Overview

Our TTA method applies two stage adaptation to predict the BN statistics (μ, σ) of test domain D_c^s in each BN layer. Specifically, we use Geometric Confidence Maximization and Entropy Minimization as the loss function to optimize the learnable mixture coefficient ϕ that we introduced in the BN layer. After each step of optimizing, we perform secondary correction on the BN statistics (μ, σ) using the optimized ϕ. Additionally, we propose a semantic-consistency

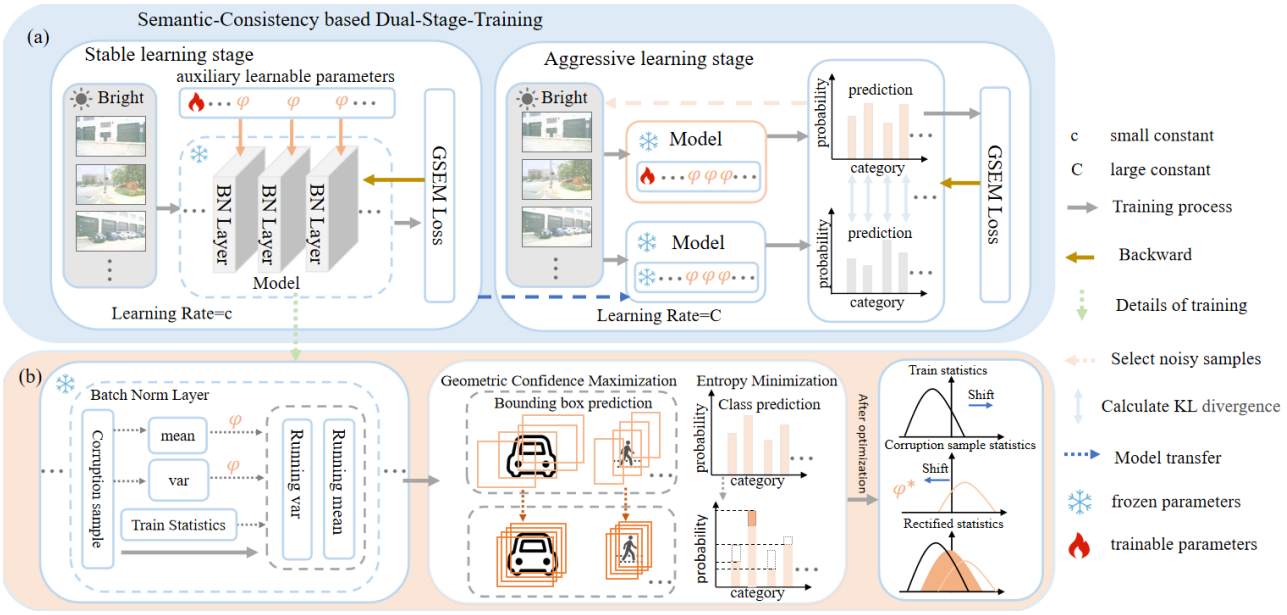


Figure 2: **Method Overview**. Module (a) illustrates the Semantic-Consistency based Dual-Stage Adaptation method. Module (b) is intended to describe the process of the proposed LearnableBN method.

based dual-stage-adaptation method. The first stage is the stable adaptation stage, which uses a smaller learning rate η for conservative estimation of BN statistics. The second stage is the aggressive adaptation stage, using a larger learning rate η to help the ϕ escape local optima and converge to global optima. To ensure the stability of model adaptation, the predictions from the second stage are compared semantic consistency with the predictions from the first stage. This comparison is used to filter noisy samples from the test domain D_c^s. The whole framework is illustrated in Fig. 2.

LearnableBN

Geometric Confidence Maximization and Entropy Minimization. Test-time adaptation (TTA) commonly relies on Entropy Minimization(EM) loss to improve model prediction confidence. However, object detection models consist of not only a classification task but also a regression task that outputs continuous values, such as the position and size of bounding boxes. As a result, the EM loss cannot be applied to the regression branch, leaving it unoptimized. Moreover, during inference, object detectors typically produce a large number of candidate boxes. Simply increasing the classification confidence through EM may lead to increased confidence scores for low-quality candidate boxes, ultimately harming detection performance.

To address this issue, we extend the idea of entropy minimization to the regression task and propose a new Geometric Confidence Maximization loss $\mathcal{L}_{GCM}$. Let the set of candidate boxes be $\mathcal{B} = \{B_1, B_2, \ldots, B_Q\}$, where Q is the query numbers. For any two boxes B_i and B_j, their Intersection over Union (IoU) is defined as:

$$\text{IoU}_{ij} = \frac{|B_i \cap B_j|}{|B_i \cup B_j|}. \tag{1}$$

We assume that the candidate box pairs $\mathcal{P}$ with IoU greater than threshold τ are correspond to the same object:

$$\mathcal{P} = \{(i,j) \mid \text{IoU}_{ij} > \tau,\ i < j\}. \tag{2}$$

Finally, we define the average $1 - \text{IoU}$ of all candidate box pairs $\mathcal{P}$ assumed to correspond to the same object as $\mathcal{L}_{GCM}$:

$$\mathcal{L}_{\text{GCM}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} (1 - \text{IoU}_{ij}). \tag{3}$$

The idea behind $\mathcal{L}_{GCM}$ is analogous to the application of entropy minimization strategies in classification tasks, where minimizing the entropy of the predicted distribution encourages the model to make more confident class predictions. Similarly, the $\mathcal{L}_{GCM}$ aims to improve the stability and consistency of the regression branch by encouraging more concentrated regression results for candidate boxes corresponding to the same object, which can be interpreted as enhancing the "confidence" of bounding box regression.

Formally, $\mathcal{L}_{GCM}$ can be regarded as reducing the variance of the predicted distribution in probability-based regression tasks. In contrast to such methods, $\mathcal{L}_{GCM}$ does not require the introduction of explicit probabilistic modeling components or prior assumptions, nor does it necessitate modifications to the original model architecture, thereby offering greater flexibility and broader applicability.

The LearnableBN integrates $\mathcal{L}_{GCM}$ as the regression loss and additionally incorporates EM loss in the classification branch to improve the confidence of class predictions:

$$\mathcal{L}_{EM} = \lambda \sum_{Q}^{i} \sum_{C}^{j} -p_{i,j} \log p_{i,j}, \tag{4}$$

where Q is the query numbers, C is the numbers of classes, $p_{i,j}$ is the predicted probability of the different classes of query, λ is the weight coefficient.

Moreover, directly using unsupervised losses to update model parameters may amplify the effect of incorrect predictions during adaptation, potentially leading to error accumulation. In contrast, the LearnableBN method leverages the two proposed losses to optimize the auxiliary learnable parameters ϕ, which do not directly modify the model parameters but instead indirectly influence the model by adjusting the statistical information in the BN layers.

Optimizing BN layers. The inherent instability of the BN layer is mainly attributed to the following factors:

(1) The exponential moving average (EMA) method used to predict statistics in the BN layer is highly dependent on the batch size. If the batch size is too small, it might not accurately reflect the full distribution of the test data domain, potentially leading to erroneous shifts in BN statistics.

(2) Within the neural network, deeper layer information is found to exhibit greater transferability, while shallow layers information often requires more frequent updates. Therefore, the update strategy for the BN statistics should be different for each layer.

(3) Predictions of BN statistics are highly sensitive to internal covariate shift, where the accuracy of statistical predictions in deep BN layers significantly influences those in shallow BN layers.

Therefore, we propose a novel BN layer method for predicting BN statistics to replace the EMA method:

In m-th BN layer the equation in forward propagation:

$$\mu = (1 - \phi_m)\mu_h + \phi\mu_p, \tag{5}$$

$$\sigma^2 = (1 - \phi_m)(\sigma_h)^2 + \phi(\sigma_p)^2, \tag{6}$$

$$\hat{z} = \frac{z - \mu}{\sqrt{\sigma^2 + \epsilon}}\gamma + \beta, \tag{7}$$

where z and $\hat{z}$ are the input and output of the BN layer, (μ_h, σ_h) are the historical BN statistics, and (μ_p, σ_p) are the BN statistics from the current sample. (γ, β) are the affine parameters, and ϵ is a small constant for numerical stability. We introduce a new learnable parameter ϕ_m to each BN layer and apply the leakyrelu function to prevent negative values. This enables each BN layers to have independent mixture coefficient. At this stage, the BN statistic (μ_p, σ_p) calculated from Eq. 5 and Eq. 6 are utilized as a temporary variables to influence the model's predictions.

After optimizing with $\mathcal{L}_{GCM}$ and $\mathcal{L}_{EM}$, we introduce a quadratic correction:

$$\phi_m^{(t+1)} = \phi_m^{(t)} - \eta \cdot \nabla_\phi(\mathcal{L}_{GCM} + \mathcal{L}_{EM}), \tag{8}$$

$$\mu = (1 - \phi_m^{(t+1)})\mu_h + \phi_m^{(t+1)}\mu_p, \tag{9}$$

$$\sigma^2 = (1 - \phi_m^{(t+1)})(\sigma_h)^2 + \phi_m^{(t+1)}(\sigma_p)^2, \tag{10}$$

where $\{..., t-1, t, t+1, ...\}$ represent each optimization step in training iterations. The first correction is necessary because the BN layer dynamically mixes the current sample's statistics with history statistics during prediction process, helping to reduce domain shift and enabling the model to predict the mixture coefficient for the current sample more accurately.

The second revision is due to the delay in the impact of the ϕ_m on the BN statistics. The ϕ_m after optimization should be the mixing coefficients of the current samples. If we use Eq. 5 and Eq. 6 as the BN statistic, it will result in the ϕ_m optimized by current $\mathcal{L}_{GCM}$ and $\mathcal{L}_{EM}$ to be used in the next sample's mixing coefficients. Therefore, we made specific adjustments to the model training process. After optimizing ϕ_m using Eq. 8, we applied Eq. 9 and Eq. 10 to correct the statistics of the BN layers.

We propose such a method to optimize the BN layer by introducing a auxiliary learnable parameters ϕ_m to replace the EMA method. It mitigates the limitation where the accuracy of BN statistics predicted using the EMA method is highly dependent on batch size, resulting in a more stable process for predicting BN statistics. Applying different BN statistics shift strategies for each BN layers, effectively utilized the transferability of the deep BN layers. Unlike the traditional model parameter, ϕ_m is an auxiliary parameter that will initialized at the start of each domain adaptation, enabling specific adaptation strategies for different domains.

Semantic-Consistency based Dual-Stage Adaptation

In our LearnableBN method, only the auxiliary learnable parameters ϕ are optimized. Adapting with a very small learning rate often results in the model converging to a local optimum due to the limited number of trainable parameters. Conversely, the peculiarities of $\mathcal{L}_{GCM}$ and $\mathcal{L}_{EM}$ can cause the model to converge to a trivial solution if an excessively large learning rate is used. To further enhance the generalization of our method and effectively handle noisy samples encountered in real-world scenarios, we propose a new semantic-consistency based dual-stage adaptation method.

As shown in figure 2, during the first stage, a small learning rate is used to allow the model to find the local optimum. In the second stage, we use a large learning rate to allow the model to escape from the local optimum. In order to guarantee the reliability of the adapting process, we compare the semantic consistency between the first-stage model and the second-stage model by evaluating the Kullback-Leibler (KL) divergence of their predictions. We consider a sample to be stable for model adapting if the KL value is in the lowest 10% of historical KL values.

The rationale behind the first adapting stage is that the model often exhibits instability when confronted with unseen domains. The original model cannot be used directly as a semantic comparison model. The local optimums obtained by the model in the first adapting stage are more transferable. Consequently, we use the predictive power of the local optimums to filter the unstable samples. During the second stage of adapting, the learning rate is increased in order to encourage the model to converge to the global optimum.

Concurrently, this method is used for sample selection, which considers the hidden layer features of samples. This approach guarantees the adapting stability while minimizing the risk of learning noisy samples during TTA.

	Low Light				Fog				Motion blur			
Severity	**Easy**	**Mid**	**Hard**	**Avg**	**Easy**	**Mid**	**Hard**	**Avg**	**Easy**	**Mid**	**Hard**	**Avg**
Baseline	0.4011	0.3352	0.2274	0.3212	**0.4908**	0.4825	0.4655	0.4796	0.4661	0.3002	0.2328	0.3330
ReviseBN	0.3382	0.2798	0.1715	0.2631	0.4296	0.4188	0.4048	0.4177	0.4316	0.3444	0.2905	0.3555
TENT	0.2636	0.2085	0.149	0.2070	0.3416	0.333	0.3161	0.3323	0.3185	0.1842	0.1442	0.2823
AdaBn	0.104	0.075	0.0528	0.0772	0.1388	0.1345	0.1266	0.1333	0.1325	0.1296	0.1218	0.1279
ARM(BN)	0.1319	0.0978	0.0587	0.0961	0.1473	0.1449	0.1372	0.1431	0.1621	0.1535	0.1297	0.1484
LearnableBN	**0.4203**	**0.3712**	**0.3097**	**0.3671**	0.4899	**0.4829**	**0.4720**	**0.4816**	**0.4707**	**0.3783**	**0.3325**	**0.3938**

Table 1: Comparison of different TTA methods on **Nuscenes-C** across three levels of severity. The baseline model is **BEVFormer** with **ResNet-101** as the backbone. **Bold**: Best in the category. Underline: Second best in the category.

Experiments

Experimental Setup

Dataset. To simulate a dynamic and real-world autonomous driving scenarios, the experiments are conducted on the Nuscenes-C (Kong et al. 2023) dataset and Kitti-C (Lin et al. 2024) dataset. NuScenes-C adds natural corruption, including exterior environments, interior sensors factors, and temporal factors, based on NuScenes (Caesar et al. 2020). It includes six types of corruption and three levels of severity: EASY, MID, and HARD. The KITTI-C dataset introducing 12 distinct types of data corruptions to the validation set based on KITTI dataset (Geiger et al. 2013). Our method compares with TTA methods, without introducing additional source data and without relying on annotations.

Metrics. In the BEV based 3D object detection, We evaluate the performance of our method with the official nuScenes mertric, *nuScenes Detection Score (NDS)*, which calculates a weighted sum of mAP, mATE, mASE, mAOE, mAVE, and mAAE. For the monocular 3D object detection task, we present our experimental results in terms of Average Precision (AP) for 3D bounding boxes, denoted as $AP_{3D|R_{40}}$.

Implementation Details. We implement our model based on Pytorch on a single NVIDIA L20 GPU. The baseline models used are BEVFormer (Li et al. 2022), Sparse4D (Lin et al. 2022) and MonoFlex. In Nuscenes-C, to evaluate the stability of TTA methods, the batch size was set to 1. In the semantic-consistency-based dual-stage adaptation, we set the learning rates η to 2e-8 and 2e-7, and learning ratio α set at 0.1. The initial value of auxiliary learnable parameters ϕ in BN layers is set to 1e-5.

Quantitative Results

We compare our method with several TTA methods as shown in Table 1, which can be classified into two main categories, (1) adjusting model parameters based on unsupervised training (ie,TENT (Wang et al. 2020)), (2) modifying the BN statistics (ie, ReviseBN (Benz et al. 2021), AdaBn (Li et al. 2018), ARM (Zhang et al. 2021))

The experimental results demonstrate that the BEV based model is highly sensitive to Batch Normalization (BN) statistics due to the number of parameters and model depth. ARM and AdaBn have caused the model to collapse. These methods have failed to predict the true distribution of the test domain, particularly when the batch size is minimal.

In response to this situation, the ReviseBN adjusts the mixture coefficient of the EMA method in accordance with the specific test domains. ReviseBN show significant improvements compared to the baseline in cases where the test domain greatly shifted from the training domain. For example, in the Motion blur corruption type, the average results improved from 0.3330 to 0.3555 compared to the baseline, However, when the test domain was similar to the training domain, ReviseBN led to a degradation of the model's predictive ability. For example, the average results of the low light corruption type decreased from 0.3212 to 0.2631. While the TENT method fine-tunes BN layer affine parameters using the EM loss. However, experiments show that TENT decrease the accuracy across all scenarios. For example, in the Low Light scenario, accuracy decreased from 0.4011 to 0.2636. This is because the TENT method, which relies solely on the EM loss and directly modifies the model parameters, causes the model to fully trust its initial predictions, resulting in overconfidence and error accumulation.

Compared to these methods, our approach adaptively learns the mixture coefficients, by adaptively learning the mixture coefficients based on the different corruption scenarios and the varying depths of BN layers, which has overcome the instability issues commonly encountered in adjust BN statistics methods, and has effectively prevented model collapse. Moreover, the introduction of the $\mathcal{L}_{GCM}$ significantly improves the model performance. The results of the experiments demonstrated that our method significantly enhanced the model's capacity for generalization in scenarios with severe domain shifts, including those involving fog, motion blur, and low light. Furthermore, as the degree of corruption increased, the efficacy of our method became increasingly evident. To illustrate, compared to the baseline in the low light corruption scenario, our method showed an improve the average performance from 0.3212 to 0.3671. Notably, in the hard severity, the performance improved significantly from 0.2274 to 0.3097. At the same time, our method also demonstrated high stability in minimal domain shift scenarios. In the Fog corruption scenarios, our method avoids the performance degradation in model predictions that is commonly observed with common TTA methods. achieved the best average performance in the fog corruption scenarios.

To learn more about how LearnableBN helps models per-

	Noise			Blur			Weather			Digital		
Method	**Gauss.**	**Shot**	**Impul.**	**Defoc.**	**Glass**	**Motion**	**Frost**	**Fog**	**Brit.**	**Contr.**	**Pixel.**	**Sat.**
Baseline	0.19	1.62	0.32	3.72	8.47	6.22	4.27	2.25	9.19	2.08	1.83	9.11
BN adaptation	6.21	8.20	9.20	7.83	5.35	7.52	6.47	9.24	9.12	9.93	12.73	9.76
TENT	6.02	7.96	9.57	7.75	6.06	8.63	6.71	9.91	10.26	10.55	12.33	10.27
EATA	6.05	7.96	9.74	7.93	6.06	9.01	6.24	9.94	9.07	10.02	12.41	10.12
MonoTTA	6.54	8.41	9.39	7.63	7.12	8.99	7.64	**10.26**	10.55	10.06	13.28	10.66
LearnableBN	**9.73**	**10.00**	**9.65**	**10.16**	**9.05**	**10.85**	**8.09**	9.62	**13.13**	**14.74**	**18.27**	**12.60**

Table 2: Comparison of different TTA methods on the **KITTI-C** validation set regarding Mean $AP_{3D|R_{40}}$ with IoU threshold set to 0.25 for the **Pedestrian** category. The baseline model is **Monoflex** . **Bold**: Best in the category.

	Motion Blur			
Severity	**Easy**	**Mid**	**Hard**	**Avg**
Sparse4D	0.4809	0.3189	0.269	0.3563
TENT	0.4868	0.368	0.3188	0.3912
ReviseBN	0.4533	0.374	0.3359	0.3877
AdaBN	0.1713	0.1737	0.1442	0.1630
LearnableBN	**0.4962**	**0.3795**	**0.3360**	**0.4035**

Table 3: Comparison of different TTA methods across three levels of severity Motion Blur in **Sparse4D** with **ResNet-101** as the backbone. **Bold**: Best in the category. Underline: Second best in the category.

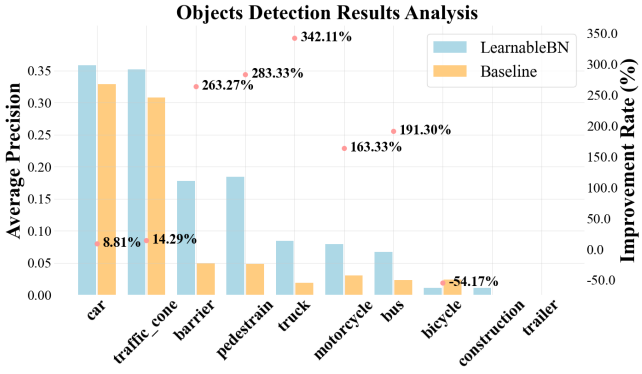


Figure 3: Comparison of detection results for Categories in Snow scenarios using BEVFormer as the baseline.

form, we conducted experiments across ten different categories under snow corruption scenarios. As shown in Fig 3, the experimental results indicate that the introduction of the LearnableBN method did not result in a significant performance improvement when the baseline was already performing well. For instance, in the detection task for traffic cones, LearnableBN achieved only about a 14% improvement. However, in the categories where the baseline model performs poorly, the LearnableBN method delivers significant improvements. Specifically, in the truck category, LearnableBN achieved a remarkable improvement of up to 342% over the baseline. These improvements are crucial for enhancing the reliability and safety of autonomous driving.

Generalization Evaluation

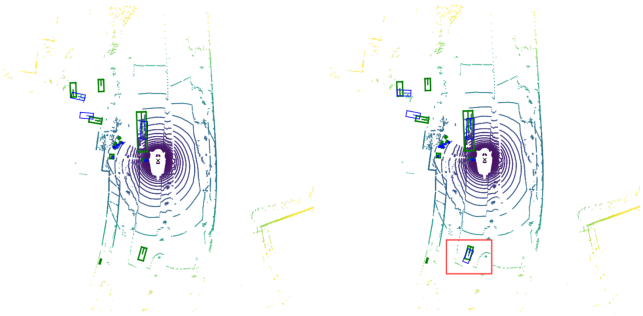
Additionally, we substituted the baseline model with the Sparse4D model and compared it with three representative TTA methods in the Motion Blur corruption scenario. We selected Motion Blur for comparison because it presents significant domain shifts across the easy, mid and hard severity levels, which helps us assess our method's performance under both minimal and severe domain shifts. As shown in Table 3, The experimental results demonstrate that our method exhibits robust performance across all severity levels, with an average improvement in performance from 0.3563 to 0.4035 in comparison to the baseline. Consistent with the experiment results using BEVFormer as the baseline model, our method proves to be more stable than the methods that adjust BN statistics. On the other hand, to further validate the performance of the LearnableBN method in different real-world scenarios and tasks, we tested various TTA methods (BN adaptation (Schneider et al. 2020), TENT (Wang et al. 2020), EATA (Niu et al. 2022), MonoTTA (Lin et al. 2024)) on the KITTI-C dataset using the monocular 3D object detection task, we compared the experimental results presented in the MonoTTA paper (Lin et al. 2024). As shown in Table 2, the experimental results show that under real-world corruptions, the pre-trained model suffers from significant performance degradation due to data distribution shifts. While our proposed LearnableBN method brings a substantial average performance improvement on MonoFlex and maintains the best performance in detecting pedestrians in the KITTI-C dataset. These experiments demonstrate that the LearnableBN method is adaptable to a wide range of base models, tasks, and real-world scenarios, highlighting its broad applicability and generalizability.

Ablation Studies

Entropy Minimization. As shown in Table 4, The LearnableBN based on EM loss significantly improves model performance under Snow, Motion Blur, and Low Light corruptions, but results in degradation in Fog, Color Quantization, and Brightness scenarios. This indicates that relying solely on EM to boost classification confidence is insufficient for achieving robust object detection under diverse corruptions.
Geometric Confidence Maximization. As shown in Table 4, compared to EM loss, after applying GCM to LearnableBN significantly improves performance in all corruption

LearnableBN	EM	GCM	Dual-stage	Snow	Motion blur	Brightness	Low Light	Fog	Color Quant
				0.2297	0.3330	**0.4908**	0.3212	0.4796	0.4184
✓	✓			0.2509	0.3712	0.4583	0.3473	0.4625	0.4004
✓	✓	✓		0.2687	0.3918	0.4680	0.3655	0.4809	0.4099
✓	✓	✓	✓	**0.2718**	**0.3938**	0.4835	**0.3671**	**0.4816**	**0.4192**

Table 4: Ablation study of our method with LearnableBN, GCM(Geometric Confidence Maximization), EM(Entropy Minimization) and Dual-stage (Semantic-Consistency based Dual-Stage-Adaptation). Comparison of the **average performance** across three levels of severity for six types of corruption.



(a) BEVformer (b) LearnableBN

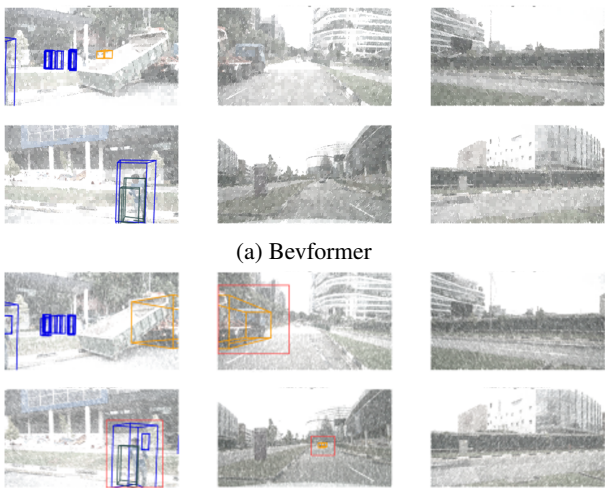
Figure 4: BEV visualization results: green box is ground truth, blue box is prediction, and red box highlight the difference before and after using LearnableBN.

scenarios. However, the experimental result also demonstrates that even with the introduction of the GCM Loss, the performance degradation in the Brightness and Color Quant corruption scenario remains unresolved (performance declined from 0.4908 to 0.4680, 0.4184 to 0.4099). This is due to the challenges of learning from unstable samples.

Semantic-Consistency based Dual-Stage-Adaptation. After introducing the semantic-consistency based dual-stage-adaptation method, compared to the results of only applying EM and GCM to the baseline, we resolved the degradation in the Brightness, Color Quant and Fog corruption scenarios. This improvement is attributed to the dual-stage training, which filtered out unstable samples and further enhanced the stability of the training process. Additionally, performance improvements were also observed in the Snow, Motion Blur, and Low Light corruption scenarios. This is attributed to the adjustment of the learning rate during the dual-stage-adaptation, which encourages the model to converge to a globally optimal solution.

Qualitative Results

To verify the effectiveness of our LearnableBN method, we utilized BEVFormer as the baseline and focused on snow scenario for visualization analysis. Fig. 4 presents the detection results in the BEV perspective, where Fig. 4a shows results of BEVFormer, while Fig. 4b shows results of BEVFormer after applying the LearnableBN method. It is clear from the figure that more objects can be detected after applying our proposed LearnableBN method. Furthermore, Fig. 5 provides visualization results from six different perspectives, where Fig. 5a represents the detection results of BEVFormer, and Fig. 5b illustrates the results of BEVFormer after applying LearnableBN. It can be observed that extreme weather conditions significantly degrade the detection ability of BEVFormer. By applying LearnableBN, not only were more objects detected, but also erroneous predictions were corrected compared to the baseline.



(a) Bevformer

(b) LearnableBN

Figure 5: Visualization results from six perspectives, with red boxes highlighting the difference before and after applying our LearnableBN. Cars are in yellow, pedestrians in blue, and cyclists in green.

Conclusion

In this paper, we presented a LearnableBN to improve the robustness of perception models in real-world autonomous driving, which introduced auxiliary learnable parameters to the BN layer, and adopted the GCM and EM loss function. Additionally, we employed the semantic-consistency based dual-stage adaptation to enhance generalization. Comprehensive experimental results demonstrated the effectiveness and superiority of our proposed methods.

Acknowledgements

This work is partly supported by the Funds for the NSFC Project (Grant No. 62572072, U24B20176, U25A6024), and Beijing Natural Science Foundation (L243027).

References

Benz, P.; Zhang, C.; Karjauv, A.; and Kweon, I. S. 2021. Revisiting batch normalization for improving corruption robustness. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 494–503.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Boudiaf, M.; Mueller, R.; Ben Ayed, I.; and Bertinetto, L. 2022. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8344–8353.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cicek, S.; and Soatto, S. 2019. Unsupervised domain adaptation via regularized conditional alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1416–1425.

Deng, W.; Qi, M.; and Ma, H. 2025. Global-local tree search in vlms for 3d indoor scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8975–8984.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Ding, M.; Huo, Y.; Yi, H.; Wang, Z.; Shi, J.; Lu, Z.; and Luo, P. 2020. Learning depth-guided convolutions for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*, 1000–1001.

Fleuret, F.; et al. 2021. Test time adaptation through perturbation robustness. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Iwasawa, Y.; and Matsuo, Y. 2021. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34: 2427–2440.

Kong, L.; Liu, Y.; Li, X.; Chen, R.; Zhang, W.; Ren, J.; Pan, L.; Chen, K.; and Liu, Z. 2023. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19994–20006.

Li, Y.; Wang, N.; Shi, J.; Hou, X.; and Liu, J. 2018. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80: 109–117.

Li, Y.; Xu, X.; Su, Y.; and Jia, K. 2023. On the robustness of open-world test-time training: Self-training with dynamic prototype expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11836–11846.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Liang, J.; He, R.; and Tan, T. 2024. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 1–34.

Lin, H.; Zhang, Y.; Niu, S.; Cui, S.; and Li, Z. 2024. Fully Test-Time Adaptation for Monocular 3D Object Detection. *arXiv preprint arXiv:2405.19682*.

Lin, X.; Lin, T.; Pei, Z.; Huang, L.; and Su, Z. 2022. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*.

Liu, Y.; Kothari, P.; Van Delft, B.; Bellot-Gurlet, B.; Mordan, T.; and Alahi, A. 2021. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34: 21808–21820.

Lv, C.; Qi, M.; Liu, L.; and Ma, H. 2025. T2sg: Traffic topology scene graph for topology reasoning in autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17197–17206.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, 16888–16905. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*.

Qi, M.; Li, W.; Yang, Z.; Wang, Y.; and Luo, J. 2019. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3957–3966.

Qi, M.; Lv, C.; and Ma, H. 2025. Robust Disentangled Counterfactual Learning for Physical Audiovisual Commonsense Reasoning. *arXiv preprint arXiv:2502.12425*.

Qi, M.; Qin, J.; Yang, Y.; Wang, Y.; and Luo, J. 2021. Semantics-aware spatial-temporal binaries for cross-modal video retrieval. *IEEE Transactions on Image Processing*, 30: 2989–3004.

Qi, M.; Wang, Y.; Li, A.; and Luo, J. 2020. STC-GAN: Spatio-temporally coupled generative adversarial networks for predictive scene parsing. *IEEE Transactions on Image Processing*, 29: 5420–5430.

Qi, M.; Ye, H.; Peng, J.; and Ma, H. 2025. Action Quality Assessment via Hierarchical Pose-guided Multi-stage Contrastive Regression. *arXiv preprint arXiv:2501.03674*.

Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; and Bethge, M. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33: 11539–11551.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, 9229–9248. PMLR.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2517–2526.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Wang, R.; Qi, M.; Shao, Y.; Zhou, A.; and Ma, H. 2024. Adversarial Contrastive Learning Based Physics-Informed Temporal Networks for Cuffless Blood Pressure Estimation. *arXiv e-prints*, arXiv–2408.

Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, 180–191. PMLR.

Xie, Q.; Luong, M.-T.; Hovy, E.; and Le, Q. V. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10687–10698.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17830–17839.

Ye, H.; Qi, M.; Liu, Z.; Liu, L.; and Ma, H. 2025. SafeDriveRAG: Towards Safe Autonomous Driving with Knowledge Graph-based Retrieval-Augmented Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 11170–11178.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6023–6032.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, J.; Qi, L.; Shi, Y.; and Gao, Y. 2023. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18971–18981.

Zhang, M.; Marklund, H.; Dhawan, N.; Gupta, A.; Levine, S.; and Finn, C. 2021. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34: 23664–23678.

Zhang, Y.; Lu, J.; and Zhou, J. 2021. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3289–3298.

Zhu, P.; Qi, M.; Li, X.; Li, W.; and Ma, H. 2023a. Unsupervised self-driving attention prediction via uncertainty mining and knowledge embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8558–8568.

Zhu, Z.; Zhang, Y.; Chen, H.; Dong, Y.; Zhao, S.; Ding, W.; Zhong, J.; and Zheng, S. 2023b. Understanding the Robustness of 3D Object Detection With Bird's-Eye-View Representations in Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21600–21610.

Zou, Z.; Ye, X.; Du, L.; Cheng, X.; Tan, X.; Zhang, L.; Feng, J.; Xue, X.; and Ding, E. 2021. 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*, 2713–2722.