A Black-Box Evaluation Framework for Semantic Robustness in Bird's Eye View Detection

Fu Wang[1, 2], **Yanghao Zhang**[2], **Xiangyu Yin**[2], **Guangliang Cheng**[2], **Zeyu Fu**[1], **Xiaowei Huang**[2], **Wenjie Ruan**[1, 2]

[1] Department of Computer Science, University of Exeter, Exeter, EX4 4QF, UK
[2] Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK
{fw377, z.fu, wruan}@exeter.ac.uk
{yanghao.zhang, x.yin22, guangliang.cheng, xiaowei.huang}@liverpool.ac.uk

Abstract

Camera-based Bird's Eye View (BEV) perception models receive increasing attention for their crucial role in autonomous driving, a domain where concerns about the robustness and reliability of deep learning have been raised. While only a few works have investigated the effects of randomly generated semantic perturbations, aka natural corruptions, on the multi-view BEV detection task, we develop a black-box robustness evaluation framework that adversarially optimises three common semantic perturbations: geometric transformation, colour shifting, and motion blur, to deceive BEV models, serving as the first approach in this emerging field. To address the challenge posed by optimising the semantic perturbation, we design a smoothed, distance-based surrogate function to replace the mAP metric and introduce SimpleDIRECT, a deterministic optimisation algorithm that utilises observed slopes to guide the optimisation process. By comparing with randomised perturbation and two optimisation baselines, we demonstrate the effectiveness of the proposed framework. Additionally, we provide a benchmark on the semantic robustness of ten recent BEV models. The results reveal that PolarFormer, which emphasises geometric information from multi-view images, exhibits the highest robustness, whereas BEVDet is fully compromised, with its precision reduced to zero.

Code — https://github.com/TrustAI/RobustBEV

Introduction

The landscape of autonomous driving is undergoing a significant transformation, propelled by the rapid advancement of deep neural network-based methods (Chitta, Prakash, and Geiger 2021; Prakash, Chitta, and Geiger 2021; Zhang et al. 2022c; Hu et al. 2022a,b; Chen and Krähenbühl 2022; Hu et al. 2023). One of the central elements in this evolution is the Bird's Eye View (BEV) perception. As highlighted by recent studies (Hu et al. 2023; Xie et al. 2023a), BEV perception is a critical intermediary in boosting the overall efficacy of the full autonomy stack. Significant research efforts have been directed towards developing BEV perception (Ma et al. 2022). Particularly, camera-based BEV perception models have gained increasing attention over the LiDAR-based approaches in recent development. This shift is primarily attributed to the cost-effectiveness of camera systems compared to LiDAR systems. Furthermore, camera-based perception models can provide rich semantic information that is not only meaningful for human observers but also enhances the performance in downstream tasks (Hu et al. 2023).

BEV detection is a fundamental perception task that identifies objects from multi-view input and maps them onto the BEV planform. BEV detection models can extract useful features from the images obtained, which are then utilised by downstream tasks such as object tracking (Yin, Ruan, and Fieldsend 2024), occupancy prediction, and route planning (Hu et al. 2023). It is indisputable that autonomous driving is a safety-critical scenario (Chowdhury et al. 2020) that requires robust and reliable BEV detection models (Cao et al. 2021). Recently, a number of studies (Yang et al. 2022; Zhu et al. 2023; Xie et al. 2023b,a) have been conducted to evaluate the robustness of camera-based BEV detection models against adversarial attacks (Carlini et al. 2019) and natural corruptions (Hendrycks and Dietterich 2019). The findings from these studies, unfortunately, indicate that BEV models also suffer from adversarial threats. In particular, many works investigated the robustness of the BEV detection models against random natural corruption (Zhu et al. 2023; Xie et al. 2023a). These perturbations, crafted to mimic natural corruptions, account for potential disruptions stemming from environmental conditions and hardware constraints. Given that the generation of natural corruption is random and independent of the target models, these studies primarily shed light on the 'average-case' robustness of BEV detection models against semantic perturbations (Xie et al. 2023a). However, these approaches may not fully expose the underlying vulnerabilities of BEV models if the 'worst-case' robustness is not taken into account. Similar to the spirit of the adversarial examiner framework (Shu et al. 2020), we move beyond the natural corruptions and study the adversarial robustness of BEV detection models when subjected to semantic perturbations and develop a gradient-free, query-based evaluation framework. By leveraging advances in global optimisation techniques, the proposed evaluation framework is compatible with most existing camera-based BEV models. It can identify more competitive adversarial perturbations than natural corruption methods, thereby providing a more reliable measurement of the 'worst-case' performance of BEV models.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

This work begins by formulating the semantic perturbation threat model. We notice that adopting the mean Average Precision (mAP) metric as an objective function for optimisation presents challenges due to its discontinuous nature. To address this issue, we design a distance-based surrogate objective function that incorporates the bounding box matching mechanism used in BEV detection tasks. As defined in Eq. (3), our surrogate function is continuous for matched bounding boxes while exhibiting a strong negative correlation with detection precision. These characteristics enable the function to serve as an effective proxy for optimising semantic perturbations aimed at degrading the performance of targeted BEV models. Furthermore, we propose a query-based global optimisation algorithm called SimpleDIRECT, to solve the optimisation problem by finding the 'worst-case' perturbation that compromises the target BEV models. In this algorithm, we boost the optimisation performance and computational efficiency of the DIRECT optimisation (Piyavskii 1972) by reducing redundancies in identifying potential optimal solutions and leveraging slope information to guide the optimisation process. To evaluate the semantic robustness of camera-based BEV detection models, we introduce an evaluation framework focusing on three common semantic perturbations: geometric transformation, colour shift, and motion blur. These perturbations represent typical challenges posed by camera-system anomalies and real-world driving scenarios. We conduct extensive experiments on the nuScenes dataset (Caesar et al. 2020) to assess the performance of recent BEV models under these challenging conditions and empirically demonstrate the superiority of our proposed evaluation framework over randomised perturbations used in previous literature and two optimisation-based baseline methods. In our robustness benchmark on ten BEV models, our evaluation compromised the BEVDet (Huang et al. 2021) that shows considerable resilience against natural corruption. These findings highlight shortcomings in previous literature (Zhu et al. 2023; Xie et al. 2023a) and the critical need for our work.

Related Works

Robustness of BEV detection In this work, we focus on the object detection task in BEV perception based on multi-view images. As a fundamental task in autonomous driving, the robustness of object detection has attracted notable research interest (Kong et al. 2023a,b; Xie et al. 2023a). To assess the robustness of camera-based BEV detection models, several efforts have been made to study the pixel-level and patch-level adversarial attacks (Abdelfattah et al. 2021; Park et al. 2021; Zhang et al. 2022a; Xie et al. 2023b). Unlike adversarial attacks (Carlini et al. 2019), which often exploit model vulnerabilities with unrealistic inputs, semantic perturbations (Hendrycks and Dietterich 2019) mimic naturally occurring corruptions. Therefore, semantic perturbations have been widely considered in robustness evaluations to assess model performance under real-world conditions (Mirza et al. 2021; Yu et al. 2023). Most recently, Kong et al. (2023a) and Xie et al. (2023a) introduced natural corruption respectively on LiDAR and camera-based BEV models, and Zhu et al. (2023) studied the influence of common visual corruptions. While these approaches provide valuable insights, they may not comprehensively reveal the underlying vulnerabilities of BEV models. Their limitation lies in relying on randomised perturbations to assess the semantic robustness of these models.

Deterministic Global Optimisation Originating from the Lipschitz optimisation (Piyavskii 1972), DIRECT is a gradient-free deterministic optimisation algorithm that works on Lipschitz continuous objective functions but does not rely on the Lipschitz constant (Jones et al. 1993; Gablonsky 2001; Zhang et al. 2023). Recent studies have applied DIRECT to assess the robustness of deep neural network models against geometric transformations (Wang et al. 2023) and adversarial patches (Xu et al. 2023), proving its efficiency in approximating the global optimum and highlighting its effectiveness.

Semantic Adversarial Threat Model

Problem Formulation In the domain of 3D object detection utilising multi-monocular views, an input frame $\mathbf{x} \in \mathbb{R}^{N\times H\times W\times 3}$ encapsulates N images, each with dimensions of height H, width W, and three colour channels. Given an end-to-end detection model $\mathcal{F}$, we have $\mathcal{F}(\mathbf{x}) = \hat{\mathbf{y}}$, where $\hat{\mathbf{y}} \in \mathbb{R}^{U\times k}$ is the model's prediction that comprises U bounding boxes, each described by a k-dimensional vector. Considering a s-dimensional semantic perturbation, we apply the perturbation with different setups on each image in the input frame, resulting in a sN-dimensional parameter space. We denote by $\mathcal{S}_\theta$ a semantic perturbation with parameter $\theta \in \mathbb{R}^{sN}$. Given the annotation $\mathbf{y} \in \mathbb{R}^{V\times k}$ that contains V ground-truth bounding boxes and $\hat{\mathbf{y}}_\theta$ the predicted bounding boxes from the perturbed example $\mathcal{S}_\theta(\mathbf{x})$, we aim to find the optimal θ such that

$$\arg\max_{\theta\in\Theta} \mathcal{L}\big(\mathcal{F}(\mathcal{S}_\theta(\mathbf{x})), \mathbf{y}\big), \tag{1}$$

where $\mathcal{L}$ is the detection loss and Θ is a bounded adversary space. Furthermore, we assess the target model's robustness by determining the optimal perturbation for each frame in a video clip, aiming to measure its *worst-case* performance in the face of semantic perturbations within the autonomous driving scenario.

Distance-based Objective Function Building upon the framework established by the nuScenes dataset (Caesar et al. 2020), a majority of existing works (Xie et al. 2023a; Zhu et al. 2023; Li et al. 2022) employs the NDS to measure the detection performance in BEV perception. The NDS is a composite measurement that combines five different metrics, with mean Average Precision (mAP) playing the most significant role. In the nuScenes detection task, we observe that mAP calculation differs from the convention. Instead of using the intersection over union metric, it is based on the 2D centre distances between predicted and annotated bounding boxes on the ground plane. This implementation introduces a box-matching mechanism based on a distance threshold τ (Caesar et al. 2020). Let $\mathcal{D} : \mathbb{R}^{U\times k} \times \mathbb{R}^k \to \mathbb{R}^U$ be a function that returns the 2D centre distances between predicted bounding boxes and annotated boxes. Given a ground

truth detection box $\mathbf{y}_v$ and predicted detection boxes $\hat{\mathbf{y}}$, the matched prediction exists if $\min \mathcal{D}(\hat{\mathbf{y}}^v, \mathbf{y}_v) < \tau$ and is given by $\arg\min \mathcal{D}(\hat{\mathbf{y}}^v, \mathbf{y}_v)$. Once a match is identified, the corresponding detection box is recorded and subsequently excluded from further matching processes. The box-matching mechanism on an annotation $\mathbf{y}$ can be formulated as:

$$\sum_{v=1}^{V} \mathbb{1}\big(\min \mathcal{D}(\hat{\mathbf{y}}^v, \mathbf{y}_v) \leq \tau\big), \tag{2}$$

where $\mathbb{1}$ is the indicator function and $\hat{\mathbf{y}}^v$ is a subset of $\hat{\mathbf{y}}$ containing the predicted boxes with the same classification label as the ground truth box $\mathbf{y}_v$.

Apparently, Eq. (2) is not an ideal objective function for optimisation due to its discontinuity, so we propose the following loss function as a surrogate of the mAP metric:

$$\mathcal{L}\big(F(S_\theta(\mathbf{x})), \mathbf{y}\big) = \sum_{v=1}^{V} \min\big(\min \mathcal{D}(\hat{\mathbf{y}}_\theta^v, \mathbf{y}_v), \tau\big), \tag{3}$$

which ensures that the function value remains continuous for pairs of predicted and ground truth boxes that meet the matching criteria given in Eq. (2). Maximising Eq. (3) would force the semantic perturbation to increase the distance between the matched bounding boxes and simultaneously avoid the emergence of new matches.

Simplified DIRECT Optimisation

DIRECT (Jones et al. 1993) is a Deterministic Optimisation (DO) algorithm that has been adopted to identify the optimal semantic perturbation setup in the adversary space and achieved considerable performance (Wang et al. 2023). In this section, we delve into its core mechanisms and propose a simplified strategy to enhance its performance.

Algorithm Overview To locate the global optimum $\theta^\star$ in the adversarial space Θ, the DO starts by projecting Θ into a unit hypercube. The whole adversarial space Θ is then treated as the root node and divided into a tree-like partition throughout the optimisation process (Munos 2011). At every depth of this partition tree, we can form a set of subspaces ϑ_h, where $0 \leq h \leq H$ and H is the maximum depth. Let the diameter of a node be $\delta(\Theta) = ||\Theta||_\infty$, the subspaces at the same depth possess an identical diameter δ, *i.e.*, $\delta(\Theta_a)$ = $\delta(\Theta_b)$ holds for any $\Theta_a, \Theta_b \in \vartheta_h$. The deterministic optimisation operates by selectively dividing specific leaf nodes at different depths, thereby creating new nodes, and it evaluates the objective function at the central point, denoted as θ of each new node. The PO node selection and partitioning are conducted repeatedly during the optimisation loop.

As a specific implementation of DO, DIRECT divides the PO nodes following a space trisection strategy. Without loss of generality, consider a d-dimensional node Θ_r at depth h that is poised for further exploration. Θ_r is shaped as a hyperrectangle characterised by m dimensions each with longer sides of length 3^{-h} and $d-m$ dimensions each with shorter sides of length 3^{-h-1}, where $0 < m \leq d$. DIRECT performs query sampling within those m dimensions that correspond to long edges. The locations of sampled points are given by $\theta_r \pm 3^{-h-1}\mathbf{e}_i$, for $i \in \{1, \ldots, m\}$, where $\mathbf{e}_i$ is a unit vector along i-th dimension. After evaluating the objective function at all newly sampled points, DIRECT divides each long side into three equal parts, creating new subspaces. These subspaces are centred at the sampled points.

Algorithm 1: Deterministic Optimisation Pipeline

Input: The objective function $\mathcal{L}$, the parameter space Θ, the number of iterations T, the maximum depth H
Output: The optimal perturbation factor $\tilde{\theta}$

Initialise Θ as the root node and let $\mathcal{P} = \{\Theta\}$.
$t \leftarrow 0, q \leftarrow 0, \tilde{\theta} \rightarrow 0$
while $(t \leq T)\&(\mathcal{P} \neq \varnothing)$ **do**
  Initialise $\mathcal{X} = \{\}$
  for *each potential optimal node* Θ_p *in* $\mathcal{P}$ **do**
    if $\delta_p > 3^{-H}$ **then**
      for *dimension* i *with long edge of* Θ_p **do**
        Append$(\mathcal{X}, \theta_p \pm \frac{1}{3}\delta_p \mathbf{e}_i)$
  $\mathcal{Y} = \mathcal{L}(\mathcal{X})$
  if $\max\ \mathcal{Y} > \mathcal{L}(\tilde{\theta})$ **then**
    $\tilde{\theta} = \arg\max\ \mathcal{Y}$
  for *each potential optimal nodes* Θ_p *in* $\mathcal{P}$ **do**
    Trisect node Θ_p based on query results in $\mathcal{Y}$
    Update Θ_p's size δ_p and local slope $\hat{K}_p$
  $\mathcal{P}$ = NodeSelection()
  t = t+1

Crucially, points yielding optimal results are allocated proportionally larger subspaces within Θ_r during partition. This principle ensures that areas of the search space with more promising results are explored more thoroughly (Jones et al. 1993). Besides, Wang et al. (2023) proposed to track the slopes between the queried points during space trisection. To achieve this, they enabled DIRECT to compute the slopes along each long edge as follows:

$$\hat{K}_{r,i}^{+} = \frac{|\mathcal{L}(\theta_r) - \mathcal{L}(\theta_r^{i+})|}{3^{-h-1}} \quad \text{and} \quad \hat{K}_{r,i}^{-} = \frac{|\mathcal{L}(\theta_r) - \mathcal{L}(\theta_r^{i-})|}{3^{-h-1}},$$

where θ_r^{i+} and θ_r^{i-} denote $\theta_r \pm 3^{-h-1}\mathbf{e}_i$, respectively. Then the largest slope at node Θ_r given by

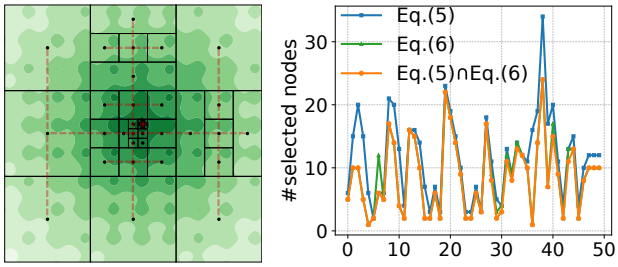
$$\hat{K}_r = \max\{\hat{K}_{r,1}^{+}, \hat{K}_{r,1}^{-} \ldots, \hat{K}_{r,m}^{+}, \hat{K}_{r,m}^{-}\}.$$

We illustrate the space transition in Fig. 1a, highlighting potential slope computation points with dashed red lines. The recorded slopes can be viewed as an approximate to the lower bound of local Lipschitz constants that assist for estimating the potential improvement at each node, so we also adopt this approach in our practice, which will be elaborated in the following sections. We present the pseudocode of this deterministic optimisation pipeline in Alg. 1 and refer the readers to (Jones et al. 1993; Munos 2011) for more details.

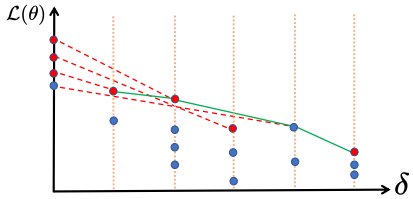
Redundancy in the PO Node Selection

Lemma 1 (Potential Optimal nodes (Gablonsky 2001))**.** *Let* $\mathcal{L}_{\max}$ *denote the current best query result, and* H *represent the depth of the partition tree. Given a node set* $\vartheta = \bigcup_{h=1}^{H} \vartheta_h$ *and a positive tolerance* $\epsilon > 0$*. For any node* Θ_p *at depth* h*, we can define three sets as follows:* $\mathcal{I}_1^h = \{\Theta_q \in \vartheta : \delta_q < \delta_p\}$, $\mathcal{I}_2^h = \{\Theta_q \in \vartheta : \delta_q > \delta_p\}$ *and* $\mathcal{I}_3^h = \{\Theta_q \in \vartheta : \delta_q = \delta_p\}$*, where* $\delta_p = \delta(\Theta_p)$*. Then the node* Θ_p *is said to be potentially optimal if*

$$\mathcal{L}(\theta_p) \geq \mathcal{L}(\theta_q), \forall \Theta_q \in \mathcal{I}_3^h, \tag{4}$$



(a) Space partition (b) Counting the PO nodes



(c) The difference between Eq. (5) and (7)

Figure 1: Fig. (a) visualises the space trisection strategy, where the dashed red lines represent where the slopes can be obtained. Fig. (b) demonstrates the redundancy introduced by Eq. (5), which often qualifies notably more nodes than Eq. (6). Fig. (c) illustrates the difference between Eq. (5) (green lines) and Eq. (7) (dashed red lines).

where θ_a denotes the centre of node Θ_a, and the following inequations hold:

$$\min_{\Theta_q \in \mathcal{I}_1^h} \frac{\mathcal{L}(\theta_p)-\mathcal{L}(\theta_q)}{\delta_p-\delta_q} \geq \max_{\Theta_q \in \mathcal{I}_2^h} \frac{\mathcal{L}(\theta_q)-\mathcal{L}(\theta_p)}{\delta_q-\delta_p}; \quad (5)$$

$$|\mathcal{L}_{\max}| \cdot \epsilon \leq \mathcal{L}(\theta_p) - \mathcal{L}_{\max} - \delta_p \max_{\Theta_q \in \mathcal{I}_2^h} \frac{\mathcal{L}(\theta_q)-\mathcal{L}(\theta_p)}{\delta_q-\delta_p}. \quad (6)$$

The PO node selection is a core mechanism of DIRECT. According to Lemma 1, DIRECT employs three conditions to select the PO nodes. While Eq. (4) identifies nodes yielding the highest query result at each depth, Eq. (5) and (6) refine the selection based on the potential for improvement over the current optimal results (Gablonsky 2001) Although both Eq. (5) and (6) play a critical role in selecting PO nodes, our analysis reveals that Eq. (5) has less impact in identifying PO nodes compared to Eq. (6). To illustrate, we applied DIRECT optimisation to the 6-dimensional Schwefel test function (Yang 2010) and tracked the number of nodes satisfying each condition. As shown in Fig. 1b, we can see that the PO nodes selection is mostly dominated by Eq. (6), while Eq. (5) qualifies a notably larger number of nodes at each iteration, implying its relative inefficiency.

Simplifying the PO Node Selection

Querying BEV perception models is computationally expensive, and introducing less qualified and redundant PO nodes would slow down the optimisation process. Due to the impractical selection condition given in Eq. (5), we propose to select the R highest-ranking nodes in terms of potential improvement. Specifically, consider $\tilde{\vartheta}$ as a set of lead nodes qualified by Eq. (4) and (6), if its cardinality $|\tilde{\vartheta}| > R$, we

Algorithm 2: NodeSelection() in SimpleDIRECT

Input: The maximum number of PO nodes R
Output: A set of PO nodes $\mathcal{P}$

Set $\tilde{\vartheta} = \{\}$
for *each depth* $h \in H$ **do**
  Append$(\tilde{\vartheta}, \Theta_p)$ **if** Θ_p satisfies Eq. (4) and (6)
if $|\tilde{\vartheta}| > R$ **then**
  Rank $\tilde{\vartheta}$ based on Eq. (7)
  $\mathcal{P} = \{\text{Top } R-1 \text{ nodes from } \tilde{\vartheta}\} \cup \{\arg\max \delta(\tilde{\vartheta})\}$
else
  $\mathcal{P} = \tilde{\vartheta}$

assign a score I for each selected node. For any $\Theta_j \in \tilde{\vartheta}$, the score is calculated as

$$I(\Theta_j) = \mathcal{L}(\theta_j) + 0.5 \cdot \delta(\Theta_j) \cdot \hat{K}_j, \quad (7)$$

where $I(\Theta_j)$ estimates the potential improvement at Θ_j, taking into account its diameter $\delta(\Theta_j)$ and the largest slope $\hat{K}_j$ observed in its closest vicinity. We then further refine $\tilde{\vartheta}$ by picking the top $R-1$ leaf nodes based on this score given in Eq. (7) and the leaf node with the largest diameter, determining the final PO nodes. The leaf node with the largest diameter is selected because it always satisfies Eq. (5) and Eq. (6) and plays an important role in the convergence guarantee (Gablonsky 2001). The pseudocode of this simplified PO node selection strategy is presented in Alg. 2, and, by integrating Alg. 1 and Alg. 2 together, we propose an improvement of DIRECT, namely **SimpleDIRECT**. An illustration of the difference between our strategy and its counterpart given in Eq. (5) is shown in Fig. 1c, where the nodes selected by Eq. (5) is connected by green lines and our strategy is visualised by red dashed lines. We highlight that Eq. (5) selects the PO nodes based on an analytic comparison between nodes at different depths, while our strategy gives nodes with larger slopes higher priority and could reveal optimal nodes earlier. Moreover, our simplified condition does not require the left-hand side of Eq. (5), whose computational complexity is $\mathcal{O}(H^2)$. As the complexity of estimating and ranking the potential improvement is $\mathcal{O}(H \log H)$, our approach also advances Eq. (5) in terms of computational efficiency.

Applying Semantic Perturbations

When deploying a trained BEV perception model in a vehicle equipped with cameras different from those used during the training data collection, variations in camera systems may result in **geometric transformations** in the captured monocular images. Similarly, due to sensor variations among different camera systems, **colour shift** in recorded images is a common issue (Xie et al. 2023a). On the other hand, as a natural and inevitable distortion in autonomous driving contexts, **motion blur** could significantly affect image clarity and, consequently, the accuracy of BEV perception models. Therefore, we apply these three semantic perturbations to examine the robustness of BEV perception derived from multi-view images.

Geometric Transformation We evaluate the models' robustness against scaling and translation transformation. For each monocular image, $\mathbf{x}_n \in \mathbb{R}^{H\times W\times C}$, and its corresponding variant, $\mathbf{x}'_i$, the relationship between pixel coordinates in the perturbed image and the original image is governed by the intrinsic matrix. Specifically, a pixel in $\mathbf{x}'_i$ at location $(\mathtt{x}'_i, \mathtt{y}'_i)$ corresponds to a pixel in $\mathbf{x}_n$ with index $(\mathtt{x}_j, \mathtt{y}_j)$. Such a mapping can be written as

$$\begin{bmatrix} \mathtt{x}_j \\ \mathtt{y}_j \end{bmatrix} = \begin{bmatrix} \theta_s^{hor} & 0 & \theta_t^{hor} \\ 0 & \theta_s^{vrt} & \theta_t^{vrt} \end{bmatrix} \begin{bmatrix} \mathtt{x}'_i \\ \mathtt{y}'_i \\ 1 \end{bmatrix}, \tag{8}$$

where θ_s^{hor}, θ_s^{vrt}, θ_t^{hor}, and θ_t^{vrt} are four intrinsic perimeters. By manipulating θ_s^{hor} and θ_s^{vrt}, we can respectively simulate horizontal and vertical scaling effects, while altering θ_t^{hor} and θ_t^{vrt} introduces horizontal and vertical translation in the image. Furthermore, the geometric transformation can be formulated as a Lipschitz continuous operation, which has been proved by Li et al. (2021) and Wang et al. (2023).

Colour Shift Echoing the previous practices (Mohapatra et al. 2020; Zhang et al. 2022b) in perturbing the colour space, we manipulate image colours in the HSB (Hue, Saturation, and Brightness) space rather than RGB. Given an image, we first obtain its representation in the HSB colour space and then directly change the hue, saturation, and brightness values to achieve colour shifting. Let $\mathbf{x}^{hue}$, $\mathbf{x}^{sat}$, and $\mathbf{x}^{brt}$ represent the colour channels in the HSB space. The hue, saturation, and brightness perturbations can formulated as follows:

$$S^{hue}(\mathbf{x}^{hue}, \theta^{hue}) = (\mathbf{x}^{hue} + \theta^{hue}) \bmod 2\pi, \tag{9}$$

$$S^{sat}(\mathbf{x}^{sat}, \theta^{sat}) = \min(\max(0, \theta^{sat} \cdot \mathbf{x}^{sat}), 1), \tag{10}$$

$$S^{brt}(\mathbf{x}^{brt}, \theta^{brt}) = \min(\max(\mathbf{x}^{brt} + \theta^{brt}, 0), 1). \tag{11}$$

The perturbed image is then converted back to the RGB space to evaluate the robustness of the BEV perception models. A rigorous proof of continuity for colour projection is beyond the scope of this paper. We assume that colour shifting is Lipschitz continuous for small perturbations, given the inherent continuity of colour mapping from HSB to RGB space (Levkowitz et al. 1993).

Motion Blur To integrate motion blur into our robustness evaluation, we keep the size of the blur kernel a constant and optimise the blur angle θ^{ang} and direction θ^{dir} to maximise its impact on BEV perception models (Riba et al. 2020). In this case, motion blur can be divided into two procedures: rotating the blur kernel at a given direction and applying the kernel in a convolution manner. The continuity of rotating a 2-D metric has been proved by Wang et al. (2023), while Liang and Huang (2021) demonstrated the convolution operation without bias is a linear transformation. Therefore, the overall process of performing motion blur is continuous.

Experiments

In this section, we first demonstrate the effectiveness of the designed objective function and make initial observations on the impact of different semantic perturbations on the BEV perception models' performance. Then, we evaluate the effectiveness of the proposed method in locating the optimal setup for different perturbations and compare its performance to baselines. Furthermore, we benchmark the robustness of recent BEV perception models, providing a comprehensive assessment of the robustness of the models under varied adversarial image corruptions. Due to limited GPU resources, we adopt the validation set of the mini-subset of nuScenes (Caesar et al. 2020) and expand our approach to the whole validation set. The implementation details are deferred to **Appendix**.

Perturbation	Parameters for each image
Geometric Transformation	$\theta_s^{vrt} \in [-\gamma H, \gamma H],\ \theta_s^{hor} \in [-\gamma W, \gamma W]$ $\theta_t^{vrt}, \theta_t^{hor} \in [1-\gamma, 1+\gamma]$
Colour Shift	$\theta^{brt} \in [-\gamma, \gamma],\ \theta^{sat} \in [1-\gamma, 1+\gamma]$ $\theta^{hue} \in [-\pi\cdot\gamma, \pi\cdot\gamma]$
Motion Blur	$\theta^{ang} \in [-\pi, \pi], \theta^{dir} \in [-1, 1]$

Table 1: The bound of each perturbation factor.

Evaluating Semantic Perturbations

In this experiment, we evaluate the effectiveness of the proposed objective function over different semantic perturbations and BEV models. As summarised in Tab. 1, we introduce a parameter γ to control the bound of perturbation factors. We constrain γ to the range $[0.1, 0.4]$ for colour perturbations and use $\gamma \in [0.04, 0.1]$ for geometric perturbations, where $\gamma = 0.1$ allowing the perturbation to scale or translate the image by up to 10% of its size. The severity of motion blur is controlled by the size of the blur kernel (Xie et al. 2023a), and we set the kernel size to $\{5, 7, 9, 11\}$. Additionally, we bound the angle $\theta^{ang} \in [-\pi, \pi]$ and the direction $\theta^{dir} \in [-1, 1]$ to explore the optimal kernel at different scales (Riba et al. 2020).

We uniformly sample 5 frames from the mini-validation set in nuScenes for evaluation and conduct 50 iterations of the SimpleDIRECT algorithm and record the obtained optimum results for the initial assessment of the impact of various perturbations. In Fig. 2, we present a visualisation of the proposed objective function's value, namely the distance, and the corresponding number of matched bounding boxes under different perturbation setups. While all types of perturbations could significantly reduce the target models' performance, the most notable observation is the strong negative correlation between the distance and the number of matches. As the distance increases, the number of matched boxes decreases substantially. This phenomenon demonstrates that by maximising our proposed objective function, the generated adversarial semantic perturbations will effectively degrade the precision of the targeted BEV models.

Ablation and Comparison

The second part of our empirical study focuses on the effectiveness of the proposed SimpleDIRECT. The evaluation was done on the same sampled frames in Fig. 2. As targeted

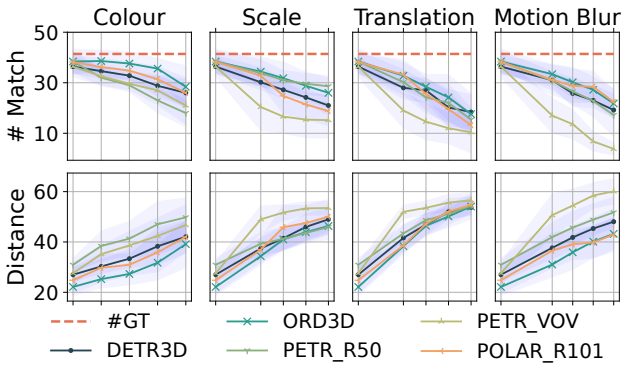


Figure 2: A illustration of the impact of semantic perturbations operated at different strengths on five BEV perception models. The effects are quantified regarding the number of matches (# Match) and the distance metric defined in Eq. (3). Each figure starts with clean input frames, while we set the following perturbations to $\gamma \in [0.1, 0.4]$ for colour, $\gamma \in [0.04, 0.1]$ for geometry, and kernel sizes $5, 7, 9, 11$ for motion blur. The shaded area indicates the standard deviations of the models' performance.

models, we select PETR (Liu et al. 2022) with two backbones: ResNet50 and VoVNet 800. In terms of the number of matched bounding boxes, their average performance on the sampled frames is 38.4 and 36.4, respectively. We perform three types of semantic perturbations, *i.e.*, colour shift, motion blur, and the combination of scaling and translation. Colour shift and motion blur are applied with $\gamma = 0.3$ and a kernel size of 9, respectively, aligning with moderate severity levels in natural corruption (Xie et al. 2023a). The combined geometric perturbation is conducted at $\gamma = 0.1$ following Wang et al. (2023)'s practice. Regarding baseline methods, as there are currently no other studies addressing the same scenario as ours, we adopt DIRECT (Wang et al. 2023), which is methodologically the most relevant work to ours, and Bayesian Optimisation with Expected Improvement acquisition function (**BO-EI**), the most popular black-box optimisation algorithm, for comparison. Additionally, inspired by the concept of natural corruption (Xie et al. 2023a), we include the random search and evaluate each perturbation's performance at its maximum strength in both directions, using these as baselines denoted by **Natural**$+/-$. All optimisation-based methods and the random search are allowed a maximum of 2000 queries for a fair comparison.

As shown in Tab. 2, SimpleDIRECT was tested at $R \in \{1, 2, 3, 4\}$, with the best performance at $R = 3$. There is a performance drop at $R = 1$ and $R = 4$, indicating that choosing too many or too few PO nodes affects efficiency. Compared to baselines, optimisation-based approaches outperform random and natural corruptions, which highlights the limitation of relying on randomised and fixed perturbations for robustness evaluation, as such methods may not fully expose the vulnerabilities of the models. SimpleDIRECT achieves the **best** or *second-best* results across all perturbations, generally outperforming DIRECT when $R > 1$.

Methods	Colour		Scale & Transl.		Motion Blur		Time
	#Match↓	Dist.↑	#Match↓	Dist.↑	#Match↓	Dist.↑	(min.)
PETR RestNet50							
Random	24.4	37.1	19.6	41.6	23.6	39.4	8.2
Natural+	28.8	40.3	19.2	52.6	36.0	34.3	-
Natural-	23.4	43.4	21.4	51.4	36.2	34.9	-
DIRECT	24.0	45.9	**13.8**	56.9	*22.4*	48.3	9.5
BO-EI	**19.8**	48.5	15.6	54.1	23.2	46.4	55.1
$R = 1$	*23.2*	46.9	14.6	55.4	23.4	46.6	9.0
$R = 2$	*23.2*	46.9	14.6	55.4	**21.4**	48.5	9.2
$R = 3$	*23.2*	47.1	*14.4*	55.3	**21.4**	48.5	9.6
$R = 4$	*23.2*	47.0	*14.4*	55.3	*22.4*	48.3	8.6
PETR VoVNet 800							
Random	27.6	42.6	10.8	47.5	9.8	51.0	7.6
Natural+	31.0	38.5	13.4	55.3	30.4	40.6	-
Natural-	27.2	40.4	23.4	49.2	31.0	39.6	-
DIRECT	27.2	41.5	9.4	57.9	*7.0*	58.3	7.8
BO-EI	**24.8**	44.4	*7.6*	58.7	9.6	56.6	29.0
$R = 1$	27.4	39.7	10	57.2	9.2	57.4	7.6
$R = 2$	26.2	43.3	**7.2**	59.0	**6.8**	58.3	7.7
$R = 3$	*26.0*	43.7	**7.2**	59.0	*7.0*	58.3	7.8
$R = 4$	26.8	42.4	*7.6*	58.8	*7.0*	58.3	7.6

Table 2: Comparison of SimpleDIRECT with baseline methods in terms of the number of matches (# Matches), the distance metric in Eq. (3), and the average run time. **SimpleDIRECT** is carried out with $R \in \{1, 2, 3, 4\}$ to assess its performance. The best results are marked in **bold** and the second-best in *italic*.

Although BO-EI, as a strong baseline, achieves the best performance on the colour shift perturbation, it was outperformed by DO methods elsewhere. Regarding the runtime, DO methods take slightly longer than random search, where the runtime is mainly spent on querying the target model, while BO-EI is significantly slower than DO methods due to fitting the Gaussian process surrogate model in high-dimensional space (Kawaguchi et al. 2015).

Benchmarking the Semantic Robustness

In the previous sections, we develop a framework that can efficiently evaluate the adversarial robustness of BEV models against semantic perturbations, which enables us to benchmark the semantic robustness of recent camera-based BEV models. Using the same perturbation setup as in the previous section, we conduct SimpleDIRECT at $R = 3$ with a maximum of 2500 queries to find the optimal perturbation. Following previous works (Kong et al. 2023a; Xie et al. 2023a), we evaluate robustness based on a model's performance on perturbed inputs, with the highest-performing model considered the most robust.

Based on the mini-validation set of nuScenes, our benchmark includes ten BEV models, varying in scale, resolution, and backbones. As summarised in Tab. 3, the PolarFormer demonstrates the best robustness across different semantic perturbations. The outstanding performance

Perturbation	Metric	BEVFormer				PETR			PolarFormer		DETR3D	ORA3D	BEVDet
		small	small+tem.	base	base+tem.	R50	V800	V1600	R101	VoV			
None	NDS↑	0.279	0.399	0.357	0.426	0.321	0.363	0.402	0.408	**0.471**	0.375	0.410	0.356
	mAP↑	0.189	0.355	0.319	0.294	0.298	0.341	0.380	0.367	**0.429**	0.313	0.359	0.300
Colour	NDS↑	0.223	0.291	0.261	0.223	0.162	0.234	0.263	0.323	**0.409**	0.306	0.350	0.000
	mAP↑	0.108	0.231	0.187	0.187	0.095	0.155	0.184	0.282	**0.329**	0.227	0.276	0.000
Scale&Shift	NDS↑	0.206	0.254	0.224	0.237	0.194	0.177	0.197	0.261	**0.273**	0.242	0.249	0.000
	mAP↑	0.085	**0.182**	0.131	0.131	0.113	0.069	0.105	0.149	0.119	0.129	0.127	0.000
Motion Blur	NDS↑	0.179	0.231	0.246	0.229	0.170	0.081	0.189	0.326	**0.367**	0.268	0.283	0.000
	mAP↑	0.100	0.159	0.180	0.160	0.104	0.010	0.093	0.229	**0.272**	0.169	0.199	0.000

Table 3: Benchmarking BEV models against Semantic perturbations.

may be attributed to the polar coordinate system (Jiang et al. 2023), which potentially enhances the model's capability to comprehend geometric information from multi-view input. Note that higher precision on clean frames does not always result in better robustness. We can observe that ORA3D, PETR with VoVNet 1600, and PolarFormer with ResNet101 show similar performance on clean frames, but PolarFormer is notably more robust than the others. In contrast, BEVDet demonstrated considerable resistance to natural corruptions (Xie et al. 2023a), but it fails to defend against semantic perturbations. This vulnerability may be attributed to the inherent data augmentation in BEVDet (Huang et al. 2021), which potentially weakens the model's reliability. On the other hand, as the temporal information is not considered during the optimisation, we reuse the semantically perturbed images to evaluate the temporal version of BEVFormer. Both small and base versions show boosted performance on clean images with temporal information. However, the impact on robustness differs between versions: the small version benefits from temporal information and achieves improved robustness, while the robustness of the base version marginally decreases.

A Case Study on the Full Validation Set

Limited by the GPU resources, we could not afford to conduct the benchmark on the full validation set that contains 150 scenes (6019 frames). In this case study, we extend our method to the full validation set by perturbing selected frames and subsequently applying the perturbation to other frames. Specifically, given that each scene contains approximately 40 frames, we perturb the initial frame of a scene and apply the resulting perturbation to the following nineteen frames. We then update the perturbation at the middle (21^{th}) frame and apply it to the remaining frames in the scene. We adopt the small version of BEVFormer with and without using temporal information as two target models to study the impact of temporal information on the robustness performance. Additionally, we conduct the random search with five attempts on all frames and report the best perturbation result on each frame as a baseline here to approximate the natural corruption in existing literature (Zhu et al. 2023; Xie et al. 2023a). As evidenced in Tab. 4, despite the perturbations being optimised on only a small fraction (5%) of the validation set, the proposed framework consistently outperforms randomised perturbation applied on the entire dataset. While both models could resist random perturbations, the model with temporal information has a notably smaller performance drop under our optimised perturbation, compared to when the temporal information is disabled.

Temp.	Methods	Colour		Scale & Transl.		Motion Blur	
		NDS↓	mAP↓	NDS↓	mAP↓	NDS↓	mAP↓
W/O[1]	Rnd.	0.252	0.122	0.244	0.112	0.247	0.120
	Ours	**0.101**	**0.007**	**0.074**	**0.006**	**0.103**	**0.007**
W/[2]	Rnd.	0.373	0.307	0.338	0.258	0.370	0.299
	Ours	**0.347**	**0.267**	**0.286**	**0.181**	**0.340**	**0.255**

[1] Performance on clean frames: NDS=0.263, mAP=0.132
[2] Performance on clean frames: NDS=0.479, mAP=0.370

Table 4: A case study on extending our framework to the full validation set. We perturb the 1^{st} and 21^{th} frames in each scene (300 frames in total) and subsequently apply the perturbation to the following frames.

Conclusion

In this work, we propose a query-based black-box framework for evaluating the 'worst-case' robustness of BEV detection models against semantic perturbations. In our experiment, this method outperforms natural perturbation and strong optimisation-based baselines and significantly reduces the precision of camera-based BEV models, demonstrating the effectiveness of the proposed framework for measuring the vulnerability of existing BEV models. As autonomous driving technology continues to evolve, the proposed evaluation framework can be easily deployed on newly developed perception models due to its black-box nature. However, our robustness evaluation is still limited to certain semantic perturbations. Developing more practical and representative perturbations offers a promising avenue for future research.

Acknowledgments

FW is funded by the Faculty of Environment, Science and Economy at the University of Exeter. WR and XH are the corresponding authors. The work was undertaken during FW's internship at the TACPS Lab at the University of Liverpool, supported by the U.K. EPSRC through End-to-End Conceptual Guarding of Neural Architectures [EP/T026995/1]. XH is also partially supported by the same project. We would like to thank the anonymous reviewers for their insightful and constructive comments, and we are also grateful to Jinwei Hu for his assistance at the early stage of this project.

References

Abdelfattah, M.; Yuan, K.; Wang, Z. J.; and Ward, R. 2021. Towards Universal Physical Attacks On Cascaded Camera-Lidar 3d Object Detection Models. In *IEEE International Conference on Image Processing*.

Caesar, H.; Bankiti, V.; Lang, A. H.; et al. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*.

Cao, Y.; Wang, N.; Xiao, C.; et al. 2021. Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks. In *SSP*.

Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; and Kurakin, A. 2019. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*.

Chen, D.; and Krähenbühl, P. 2022. Learning from all vehicles. In *CVPR*.

Chitta, K.; Prakash, A.; and Geiger, A. 2021. Neat: Neural attention fields for end-to-end autonomous driving. In *ICCV*.

Chowdhury, A.; Karmakar, G.; Kamruzzaman, J.; Jolfaei, A.; and Das, R. 2020. Attacks on self-driving cars and their countermeasures: A survey. *IEEE Access*, 8: 207308–207342.

Gablonsky, J. M. X. 2001. *Modifications of the DIRECT algorithm*. North Carolina state university.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*.

Hu, A.; Corrado, G.; Griffiths, N.; Murez, Z.; Gurau, C.; Yeo, H.; Kendall, A.; Cipolla, R.; and Shotton, J. 2022a. Model-Based Imitation Learning for Urban Driving. In *Adv. Neural Inform. Process. Syst.*

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022b. ST-P3: End-to-end Vision-based Autonomous Driving via Spatial-Temporal Feature Learning. In *ECCV*.

Hu, Y.; Yang, J.; Chen, L.; et al. 2023. Planning-oriented Autonomous Driving. In *CVPR*.

Huang, J.; Huang, G.; Zhu, Z.; et al. 2021. BEVDet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Jiang, Y.; Zhang, L.; Miao, Z.; et al. 2023. PolarFormer: Multi-camera 3D object detection with polar transformer. In *AAAI*.

Jones, D. R.; et al. 1993. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications*, 79: 157–181.

Kawaguchi, K.; et al. 2015. Bayesian optimization with exponential convergence. In *Adv. Neural Inform. Process. Syst.*

Kong, L.; Liu, Y.; Li, X.; Chen, R.; Zhang, W.; Ren, J.; Pan, L.; Chen, K.; and Liu, Z. 2023a. Robo3d: Towards robust and reliable 3d perception against corruptions. In *ICCV*.

Kong, L.; Xie, S.; Hu, H.; Ng, L. X.; Cottereau, B. R.; and Ooi, W. T. 2023b. Robodepth: Robust out-of-distribution depth estimation under corruptions. *arXiv preprint arXiv:2310.15171*.

Levkowitz, H.; et al. 1993. GLHS: A generalized lightness, hue, and saturation color model. *CVGIP: Graphical Models and Image Processing*, 55(4): 271–285.

Li, L.; Weber, M.; Xu, X.; et al. 2021. TSS: Transformation-Specific Smoothing for Robustness Certification. In *ACM CCS*.

Li, Z.; Wang, W.; Li, H.; et al. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*.

Liang, Y.; and Huang, D. 2021. Large Norms of CNN Layers Do Not Hurt Adversarial Robustness. In *AAAI*.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*.

Ma, Y.; Wang, T.; Bai, X.; et al. 2022. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*.

Mirza, M. J.; Bürkle, C.; Jarquin, J.; Opitz, M.; Oboril, F.; Scholl, K.; and Bischof, H. 2021. Robustness of Object Detectors in Degrading Weather Conditions. In *IEEE International Intelligent Transportation Systems Conference*.

Mohapatra, J.; Weng, T.-W.; Chen, P.-Y.; et al. 2020. Towards verifying robustness of neural networks against a family of semantic perturbations. In *CVPR*.

Munos, R. 2011. Optimistic Optimization of a Deterministic Function without the Knowledge of its Smoothness. In *Adv. Neural Inform. Process. Syst.*

Park, W.; Liu, N.; Chen, Q. A.; and Mao, Z. M. 2021. Sensor Adversarial Traits: Analyzing Robustness of 3D Object Detection Sensor Fusion Models. In *IEEE International Conference on Image Processing*.

Piyavskii, S. 1972. An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics*, 12(4): 57–67.

Prakash, A.; Chitta, K.; and Geiger, A. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*.

Riba, E.; Mishkin, D.; Ponsa, D.; Rublee, E.; and Bradski, G. 2020. Kornia: an Open Source Differentiable Computer Vision Library for PyTorch. In *Winter Conference on Applications of Computer Vision*.

Shu, M.; Liu, C.; Qiu, W.; and Yuille, A. 2020. Identifying model weakness with adversarial examiner. In *AAAI*.

Wang, F.; Xu, P.; Ruan, W.; and Huang, X. 2023. Towards Verifying the Geometric Robustness of Large-scale Neural Networks. In *AAAI*.

Xie, S.; Kong, L.; Zhang, W.; Ren, J.; Pan, L.; Chen, K.; and Liu, Z. 2023a. RoboBEV: Towards Robust Bird's Eye View Perception under Corruptions. *arXiv preprint arXiv:2304.06719*.

Xie, S.; Li, Z.; Wang, Z.; and Xie, C. 2023b. On the Adversarial Robustness of Camera-based 3D Object Detection. *arXiv preprint arXiv:2301.10766*.

Xu, P.; Wang, F.; Ruan, W.; Zhang, C.; and Huang, X. 2023. Sora: Scalable Black-Box Reachability Analyser on Neural Networks. In *ICASSP*.

Yang, R.; Laurel, J.; Misailovic, S.; and Singh, G. 2022. Provable Defense Against Geometric Transformations. In *ICLR*.

Yang, X.-S. 2010. *Engineering optimization: an introduction with metaheuristic applications*. John Wiley & Sons.

Yin, X.; Ruan, W.; and Fieldsend, J. 2024. Dimba: discretely masked black-box attack in single object tracking. *Machine Learning*, 113(4): 1705–1723.

Yu, K.; Tao, T.; Xie, H.; Lin, Z.; et al. 2023. Benchmarking the Robustness of LiDAR-Camera Fusion for 3D Object Detection. In *CVPR*.

Zhang, C.; Ruan, W.; Wang, F.; Xu, P.; Min, G.; and Huang, X. 2023. Model-agnostic reachability analysis on deep neural networks. In *PAKDD*.

Zhang, J.; Lou, Y.; Wang, J.; Wu, K.; Lu, K.; and Jia, X. 2022a. Evaluating Adversarial Attacks on Driving Safety in Vision-Based Autonomous Vehicles. *IEEE Internet Things J.*, 9(5): 3443–3456.

Zhang, T.; et al. 2022b. Proa: A probabilistic robustness assessment against functional perturbations. In *ECML PKDD*.

Zhang, Y.; Zhu, Z.; Zheng, W.; et al. 2022c. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*.

Zhu, Z.; Zhang, Y.; Chen, H.; et al. 2023. Understanding the Robustness of 3D Object Detection With Bird's-Eye-View Representations in Autonomous Driving. In *CVPR*.