COMMIT: Certifying Robustness of Multi-Sensor Fusion Systems Against Semantic Attacks

Zijian Huang[1]*, Wenda Chu[2], Linyi Li[3], Chejian Xu[4], Bo Li[4]
[1]University of Michigan
[2]California Institute of Technology
[3]Simon Fraser University
[4]University of Illinois Urbana-Champaign
zijianh@umich.edu, wchu@caltech.edu, linyi_li@sfu.ca, chejian2@illinois.edu, lbo@illinois.edu

Abstract

Multi-sensor fusion systems (MSFs) play a vital role as the perception module in modern autonomous vehicles (AVs). Therefore, ensuring their robustness against common and realistic adversarial semantic transformations, such as rotation and shifting in the physical world, is crucial for the safety of AVs. While empirical evidence suggests that MSFs exhibit improved robustness compared to single-modal models, they are still vulnerable to adversarial semantic transformations. In addition, although many empirical defenses have been proposed, several works show that these defenses can be further attacked by new adaptive attacks. So far, there is no certified defense proposed for MSFs. In this work, we propose the first robustness certification framework COMMIT to certify the robustness of multi-sensor fusion systems against semantic attacks. In particular, we propose a practical anisotropic noise mechanism that leverages randomized smoothing on multi-modal data and performs a grid-based splitting method to characterize complex semantic transformations. We also propose efficient algorithms to compute the certification in terms of object detection accuracy and IoU for large-scale MSF models. Empirically, we evaluate the efficacy of COMMIT in different settings and provide a comprehensive benchmark of certified robustness for different MSF models using the CARLA simulation platform. We show that the certification for MSF models is at most 48.39% higher than that of single-modal models, which validates the advantages of MSF models. We believe our certification framework and benchmark will contribute an important step towards certifiably robust AVs in practice.

1 Introduction

Autonomous driving (AD) has achieved significant advances in recent years (Redmon et al. 2016; Law and Deng 2018; Badrinarayanan, Kendall, and Cipolla 2017; Zhao et al. 2018; Zhou and Tuzel 2018; Luo, Yang, and Urtasun 2018; Qi et al. 2018; Chen et al. 2017), and deep neural networks (DNNs) have been largely deployed as the perception module for AD to process inputs from multiple sources (e.g., camera and LiDAR) to detect objects such as road signs, vehicles, and pedestrians. To make full use of multi-modal inputs, modern AD systems usually adopt the multi-sensor fusion systems (MSFs) as the perception module (Pang, Morris, and Radha 2020; Chen et al. 2022).

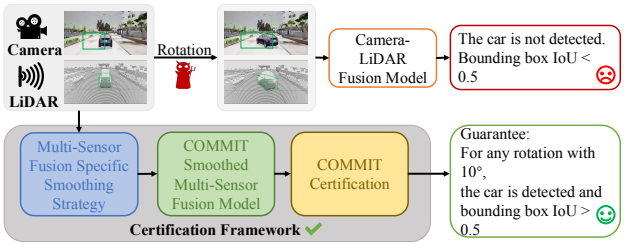

Figure 1: Overview of COMMIT, the first framework that provides certified robustness for multi-sensor fusion systems against semantic transformations.

Along with the wide deployment of AD systems, the safety of AD systems in the physical world has raised serious concerns (Hendrycks et al. 2021; Cao et al. 2021). A rich body of research has shown that both adversarial perturbations and natural semantic transformations can mislead the DNN-based perception modules in AD systems with a high success rate (Pei et al. 2017; Hosseini and Poovendran 2018; Xiao et al. 2018; Guo, Kurup, and Shah 2019; Hendrycks and Dietterich 2018; Engstrom et al. 2019), so that the AD system may ignore the pedestrians, the traffic signs, or other vehicles with high confidence when the object is slightly rotated or shifted, which can lead to severe consequences such as fatal traffic accidents (McCausland 2019). Moreover, even though the multi-sensor fusion systems may be intuitively more robust, assuming that the input transformations/perturbations are not adversarial to multiple input modalities at the same time, existing work (Cao et al. 2021; Hallyburton et al. 2022) has falsified such intuition by proposing feasible and highly efficient attacks against multi-sensor fusion systems (Cao et al. 2021). In other words, serious robustness issues still exist in existing MSFs, resulting in practical safety vulnerabilities in AD.

To mitigate such safety threats, several empirical defenses have been proposed for both single-modal models (Madry et al. 2018) and multi-sensor fusion systems (Zhong et al. 2022). However, certified defenses exist only for single-modal models (Wong and Kolter 2018; Cohen, Rosenfeld, and Kolter 2019; Li et al. 2021; Chu, Li, and Li 2022). Recent works have shown that the empirical defenses for MSFs

*This work is done during study in UIUC.

can be adaptively attacked again by stealthy perturbations or transformations (Zhong et al. 2022; Huang et al. 2022). In this paper, we aim to provide **the first robustness certification and enhancement framework for multi-sensor fusion systems** in AD against various semantic transformations in the physical world.

Our framework leverages the *randomized smoothing* technique (Cohen, Rosenfeld, and Kolter 2019), while randomized smoothing cannot directly provide a robustness guarantee against semantic transformations (e.g., object rotation and shifting) for multi-sensor fusion systems due to three main reasons: (1) **Heterogeneous input dimensions**: In randomized smoothing, the isotropic noise is added to all input dimensions, which is sub-optimal for multi-sensor fusion systems since different input modalities need different noises. (2) **Intractable perturbation spaces**: Semantic transformations incur large ℓ_p that cannot be handled by classical randomized smoothing (Yang et al. 2020), and the transformation function does not have a closed-form expression that is required for semantic-smoothing-based certification (Li et al. 2021). (3) **Unsupported certification criterion**: Existing randomized smoothing techniques are designed for certifying output consistency for classification (Cohen, Rosenfeld, and Kolter 2019) and regression (Kumar et al. 2020) tasks. However, multi-sensor fusion systems output 3D bounding boxes for detected objects and use IoU as the evaluation criterion, while randomized smoothing cannot provide worst-case certification for IoU.

To solve these challenges, our framework `COMMIT` provides the following techniques: (1) We derive an anisotropic noise mechanism that is practical (agnostic to the transformations to be certified) and efficient for randomized smoothing over multi-modal data. (2) Under a mild assumption, we propose a grid-based splitting method to integrate small ℓ_p certifications and form a holistic certification against complex semantic transformations. (3) We derive the first rigorous lower bounds of detection confidence and lower bounds of IoU for MSF models.

We leverage our framework to certify the state-of-the-art large-scale camera and LiDAR fusion 3D object detection models (CLOCs (Pang, Morris, and Radha 2020), FocalsConv (Chen et al. 2022), and MVX-Net (Sindagi, Zhou, and Tuzel 2019)) and compare them with a camera-based 3D object detector (MonoCon (Liu, Xue, and Wu 2022)) as well as one LiDAR-based 3D object detector (SECOND (Yan, Mao, and Li 2018)) in CARLA simulator. We consider transformations such as rotation and shifting, which correspond to the turning around and sudden brake cases in the real world. In particular, among the 62 scenarios randomly sampled from CARLA Town01 map, under arbitrary rotation transformations with 30%, we are able to certify the robustness of confident detection ($\geq$ 80% confidence) for all scenarios and high bounding box IoU ($\geq$ 0.5 IoU) for 53.23% scenarios. Compared to single-modal models, the certification improvements are 25.19% and 53.23%, respectively. We demonstrate that the certified robustness depends on both the input and the fusion pipeline structure.

Technical Contributions. In this paper, we provide the first certification framework for multi-sensor fusion systems against adversarial semantic transformations. We make contributions on both theoretical and empirical fronts.

- We propose the *first* generic framework for certifying the robustness of multi-sensor fusion systems against practical semantic transformations in the physical world.
- We propose a practical anisotropic noise mechanism to leverage randomized smoothing given multi-modal data, a grid-based splitting method to characterize complex semantic transformations, and efficient algorithms to compute the certification for object detection and IoU lower bounds for large-scale MSF models.
- We construct extensive experiments and provide a benchmark of certified robustness for multi-sensor fusion systems based on `COMMIT`. We certify several state-of-the-art camera-LiDAR fusion models and compare them with single-modal models. We show that the multi-sensor fusion systems provide nontrivial gains on certified robustness, e.g., achieving 53.23% improvement against the rotation transformation. In addition, we present several interesting observations which would further inspire the development of robust sensor fusion algorithms. The benchmark will be open source upon acceptance and will be continuously expanding to evaluate more AD systems.

2 Related Work

Multi-sensor fusion systems. Multi-sensor fusion DNN systems leverage data of multiple modalities to predict 3D bounding boxes for object detection. In this work, we consider multi-sensor fusion systems that take both image (from a camera) and point clouds (from a LiDAR sensor) for object detection (Pang, Morris, and Radha 2020; Chen et al. 2022), which is one of the most common forms of AD perception module (Shen et al. 2022). These fusion systems typically integrate outputs from sub-models for each modality via learning-based methods or aggregation rules. Note that our framework is architecture-agnostic — applicable for any fusion system regardless of their internal architectures.

Adversarial attacks for DNNs. The robustness vulnerabilities of DNNs are manifested by adversarial attacks. A rich body of research shows that DNNs can be attacked by pixel-wise perturbations bounded by small ℓ_p norm with even 100% success rate (Szegedy et al. 2013; Goodfellow, Shlens, and Szegedy 2014; Carlini and Wagner 2017). Besides ℓ_p-bounded perturbations, subsequent research shows that spatial transformations (Xiao et al. 2018), occlusions (Sun et al. 2020), and semantic transformations that naturally exist (Pei et al. 2017; Hendrycks and Dietterich 2018; Ghiasi, Shafahi, and Goldstein 2020) can also mislead DNNs to make severe incorrect predictions. In particular, several physically realizable and effective adversarial attacks have been proposed against multi-sensor fusion systems (Eykholt et al. 2018; Cao et al. 2021), posing serious safety threats to modern AD.

Certified robustness for DNNs. To mitigate the robustness vulnerabilities, several defenses are proposed, which

can be roughly categorized into empirical and certified defenses. The *empirical defenses* (Madry et al. 2018; Samangouei, Kabkab, and Chellappa 2018; Shafahi et al. 2019) train DNNs with heuristic approaches, e.g., adversarial training, to defend against adversarial attacks. However, they cannot provide rigorous robustness guarantees against possible future attacks. In contrast, *certified defenses* can prove that the trained DNNs are certifiably robust against any possible attacks under some perturbation constraints (Li, Xie, and Li 2023). Certified defenses are mainly based on verification methods like linear relaxations with branch-and-bound (Wong and Kolter 2018; Zhang et al. 2022b), Lipschitz DNN architectures (Zhang et al. 2022a; Singla and Feizi 2021; Xu, Li, and Li 2022), and randomized smoothing (Cohen, Rosenfeld, and Kolter 2019; Yang et al. 2020; Chiang et al. 2020; Li et al. 2022; Sun et al. 2022; Carlini et al. 2023).

For multi-sensor fusion systems, although some empirical defenses have been proposed (Liu and Lei 2022; Zhong et al. 2022), there is no certified defense that provides robustness guarantees to our best knowledge. Thus, here we aim to provide the first robustness certification and enhancement framework for MSF models.

3 Robustness Certification for Multi-Sensor Fusion Systems

In this section, we introduce our framework `COMMIT` for certifying the robustness of multi-sensor fusion systems against semantic transformations in detail.

3.1 Threat Model and Certification Goal

Notation. We consider a multi-sensor fusion system that takes an image from a camera and a point cloud from a LiDAR sensor as the input and outputs several labeled 3D bounding boxes for its detected objects. In particular, the image input $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ has d dimensions, and the point cloud input $\boldsymbol{p} \in \mathcal{P} \subseteq \mathbb{R}^{3\times N}$ contains N (un-ordered) 3D point coordinates. Note that our framework can be easily extended to handle point clouds with intensity. In the output, each labeled 3D bounding box is a tuple of box coordinates $B = (x, y, z, w, h, l, r) \in \mathcal{B} \subseteq \mathbb{R}^6 \times [0, 2\pi]$ (where x, y, z, w, h, l are 3D center coordinates and width, height, length respectively, and r is the rotation angle in the $x - z$ plane), label $c \in \mathcal{C}$, and confidence score $s \in [0, 1]$. Hence, a multi-sensor fusion system can be modeled by a function $g : \mathcal{X} \times \mathcal{P} \to (\mathcal{B} \times \mathcal{C} \times [0, 1])^n$ where n is of variable length and stands for the number of output bounding boxes.

Threat model. An adversary can apply a certain parameterized transformation that may alter both the image and point clouds to mislead the model. We formulate a transformation by two functions $T = \{T_x, T_p\}$ where $T_x : \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ transforms images and $T_p : \mathcal{P} \times \mathcal{Z} \to \mathcal{P}$ transforms point clouds respectively. Note that $\mathcal{Z} \subseteq \mathbb{R}^m$ is the set of valid and continuous parameters of the transformation, which is usually in a low-dimensional space, i.e., m is small. We consider the strongest adversary that can pick an *arbitrary* parameter $\boldsymbol{z} \in \mathcal{Z}$ to transform the input $\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{p}\end{pmatrix} \mapsto \begin{pmatrix}T_x(\boldsymbol{x}, \boldsymbol{z})\\ T_p(\boldsymbol{p}, \boldsymbol{z})\end{pmatrix}$ and feed into the system.

In particular, we will instantiate our robustness certification framework for two common transformations: **rotation** and **shifting**. The rotation transformation T_{rot} takes a scalar rotation angle r as the parameter and rotates the front car in the $x - z$ plane clockwise. The angle can be negative, meaning a counterclockwise rotation. The shifting transformation T_{sft} takes a scalar distance $\delta \in \mathbb{R}_+$ as the parameter and places the front car in δ meters away, i.e., imposes a δ displacement along the z-axis. Figure 2 illustrates these two types of transformations. Note that our framework will require only oracle access to the output of the transformation function to derive robustness certification. Hence, our framework can be readily extended to other transformations, as long as the transformation is measurable, i.e., can be deterministically parameterized.

Fine partition assumption. For common transformations, we find that when the parameter space is partitioned into tiny subspaces with ℓ_∞ diameter smaller than some threshold τ, in each subspace bounded by ℓ_2 norm, the distortion incurred by the transformation is upper bounded by the distortion with extreme points as transformation parameter. We formally state such partition assumption and empirically verify it in Appendix A.

Robustness certification goal. Our goal is to certify that, no matter what transformation parameter is chosen by the adversary within a bounded constraint or what transformation strategy is used, the multi-sensor fusion system can always detect the object and locate the object precisely. Here we mainly focus on the task that the multi-sensor fusion system aims to detect the front vehicle when it is present. Extensions to other tasks such as multi-object detection are straightforward via box alignment (Chiang et al. 2020). Now, we formalize this certification goal by two criteria: *Given an input $(\boldsymbol{x}, \boldsymbol{p})$ containing a front vehicle, a transformation T, and a constrained parameter space $\mathcal{S}$ for any transformed input $(T_x(\boldsymbol{x}, \boldsymbol{z}), T_p(\boldsymbol{p}, \boldsymbol{z}))$ with $\boldsymbol{z} \in \mathcal{S}$,*

- (Detection Certification) *the multi-sensor fusion system always outputs a bounding box for the vehicle with confidence $\geq \eta$, where η is a pre-defined threshold*;
- (IoU Certification) *the multi-sensor fusion system always outputs a bounding box for the vehicle whose volume IoU (intersection over union) with the ground-truth bounding box $\geq$ some value v.*

In the above criteria, η determines whether the confidence is high enough to report "vehicle detected", which is usually set to 0.8; the IoU is the standard for evaluating bounding box precision (i.e., given two 3D bounding boxes $B_1, B_2 \in \mathcal{B}$, $\text{IoU}(B_1, B_2) = \frac{\text{Vol}(B_1 \cap B_2)}{\text{Vol}(B_1 \cup B_2)}$ denotes the ratio of intersection volume over the union volume).

3.2 Constructing Certifiably Robust MSFs via Smoothing

Common multi-sensor fusion systems are challenging to be certified due to complex DNN architectures and fusion rules. Hence, we leverage the randomized smoothing (Cohen, Rosenfeld, and Kolter 2019), in particular, median

smoothing (Chiang et al. 2020), as the post-processing protocol to construct a *smoothed* multi-sensor system. Formally, for each coordinate of the multi-sensor fusion system $g_i : \mathcal{X} \times \mathcal{P} \to ((\mathcal{B} \times \mathcal{C} \times [0,1])^n)_i$, we add *anisotropic Gaussian noise* to the input and define q-percentile of the resulting distribution of g_i:

$$h_{iq}(\boldsymbol{x}, \boldsymbol{p}) = \sup\{y \in \mathbb{R} \mid \Pr[g_i(\boldsymbol{x} + \delta_x, \boldsymbol{p} + \delta_p) \leq y] \leq q\}, \tag{1}$$

where $\delta_x \sim \mathcal{N}(0, \sigma_x^2 \boldsymbol{I}_d)$ and $\delta_p \sim \mathcal{N}(0, \sigma_p^2 \boldsymbol{I}_{3\times N})$. We define the resulting smoothed multi-sensor fusion system $h_q := (h_{1q}, h_{2q}, \dots)$. In practice, we use finite δ_x and δ_p samples to approximate h_q by $\hat{h}_q$ with high probability and deploy (q is usually set to 0.5 so it is called median smoothing). For any q, we can obtain high-confidence intervals for h_q via Monte-Carlo sampling (Chiang et al. 2020).

Though existing work provides robustness certification for smoothed models (Cohen, Rosenfeld, and Kolter 2019; Chiang et al. 2020; Li et al. 2021; Chu, Li, and Li 2022), such certification is limited to single-modal classification or regression against ℓ_p-bounded perturbations. In contrast, our goal is to certify the robustness of multi-sensor fusion systems against semantic transformations under the two aforementioned criteria, where direct applications of prior work are infeasible due to heterogeneous input dimensions, intractable perturbation spaces, and unsupported certification criteria. In the following text, we introduce theoretical results that fulfill our robustness certification goal.

3.3 General Detection Certification

For detection certification, we locate the vehicle bounding box with the highest confidence and consider the confidence of this box as the detection confidence. Hence, for notation simplicity, we let $g : \mathcal{X} \times \mathcal{P} \to [0,1]$ to represent this detection confidence of the multi-sensor fusion system.

Theorem 1. *Let $T = \{T_x, T_p\}$ be a transformation with parameter space $\mathcal{Z}$. Suppose $\mathcal{S} \subseteq \mathcal{Z}$ and $\{\alpha_i\}_{i=1}^M \subseteq \mathcal{S}$. For detection confidence $g : \mathcal{X} \times \mathcal{P} \to [0,1]$, let $h_q(\boldsymbol{x}, \boldsymbol{p})$ be the median smoothing of g as defined in Eq.* (1)*. Then for all transformations $\boldsymbol{z} \in \mathcal{S}$, the confidence score of the smoothed detector satisfies:*

$$h_q(T_x(\boldsymbol{x}, \boldsymbol{z}), T_p(\boldsymbol{p}, \boldsymbol{z})) \geq \min_{1\leq i\leq M} h_{\underline{q}}(T_x(\boldsymbol{x}, \alpha_i), T_p(\boldsymbol{p}, \alpha_i)) \tag{2}$$

$$\textit{where} \quad \underline{q} = \Phi\left(\Phi^{-1}(q) - \sqrt{\frac{M_x^2}{\sigma_x^2} + \frac{M_p^2}{\sigma_p^2}}\right), \tag{3}$$

$$M_x = \max_{\alpha\in\mathcal{S}} \min_{1\leq i\leq M} \|T_x(\boldsymbol{x}, \alpha) - T_x(\boldsymbol{x}, \alpha_i)\|_2, \tag{4}$$

$$M_p = \max_{\alpha\in\mathcal{S}} \min_{1\leq i\leq M} \|T_p(\boldsymbol{p}, \alpha) - T_p(\boldsymbol{p}, \alpha_i)\|_2. \tag{5}$$

Remark 1. A full proof for Theorem 1 is in Appendix B.1. Suppose we have upper bounds for M_x and M_p (to be given in Lemma 2), we can compute a lower bound of $\underline{q}$, and a high-confidence lower bound of $h_{\underline{q}}(T_x(\boldsymbol{x}, \alpha_i), T_p(\underline{\boldsymbol{p}}, \alpha_i))$ via Monte-Carlo sampling. As a result, we can compute a high-confidence lower bound of detection confidence h_q. By comparing it with η in Section 3.1, we can derive the detection certification.

Lemma 2. *If the parameter space to certify $\mathcal{S} = [l_1, u_1] \times \cdots \times [l_m, u_m]$ is a hypercube satisfying the finite partition assumption (Assumption 5) with threshold τ, and $\{\alpha_i\}_{i=1}^M = \{\frac{K_1-k_1}{K_1} l_1 + \frac{k_1}{K_1} u_1 : k_1 = 0, 1, \dots, K_1\} \times \cdots \times \{\frac{K_m-k_m}{K_m} l_m + \frac{k_m}{K_m} u_m : k_m = 0, 1, \dots, K_m\}$, where $K_i \geq \frac{u_i - l_i}{\tau}$, then*

$$M_x \leq \sum_{i=1}^m \max_{\boldsymbol{k}\in\Delta} \left\| T_x(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{k})) - T_x(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{k}) + w_i) \right\|_2,$$

$$M_p \leq \sum_{i=1}^m \max_{\boldsymbol{k}\in\Delta} \left\| T_p(\boldsymbol{p}, \boldsymbol{w}(\boldsymbol{k})) - T_p(\boldsymbol{p}, \boldsymbol{w}(\boldsymbol{k}) + w_i) \right\|_2$$

where $\Delta = \{(k_1, \dots, k_m) \in \mathbb{Z}^m \mid 0 \leq k_i < K_i\}$ and $\boldsymbol{w}(\boldsymbol{k}) = (\frac{K_1-k_1}{K_1} l_1 + \frac{k_1}{K_1} u_1, \cdots, \frac{K_m-k_m}{K_m} l_m + \frac{k_m}{K_m} u_m)$. $w_i = \frac{u_i - l_i}{K_i} \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is a unit vector at coordinate i.

Remark 2. This lemma splits each dimension of $\mathcal{S}$ by a τ-cover: $\{\frac{K_i-k_i}{K_i} l_i + \frac{k_i}{K_i} u_i : k_i = 0, 1, \dots, K_i\}$. Hence, for each tiny subspace defined by $[\boldsymbol{w}(\boldsymbol{k}), \boldsymbol{w}(\boldsymbol{k}) + (w_1, \dots, w_m)]$, we can apply the finite partition assumption (Assumption 5) and the lemma follows. A full proof is in Appendix B.2. The lemma provides feasible upper bounds (via computing maximum of finite terms) for M_x and M_p, so a lower bound of $\underline{q}$ is computable, and hence the robustness certification in Theorem 1 is computationally feasible.

3.4 General IoU Certification for 3D Bounding Boxes

Median smoothing for 3D bounding boxes. Given a base 3D bounding box predictor for the front vehicle $g : \mathcal{X} \times \mathcal{P} \to \mathcal{B}$ with $\mathcal{B} \subseteq \mathbb{R}^6 \times [0, 2\pi]$ describing the geometric shape of the bounding box (details in Section 3.1), we denote by $h_q(\boldsymbol{x}, \boldsymbol{p})$ the coordinate-wise median smoothing on the outputs of g following Equation (1).

First, by applying Theorem 1 on each coordinate of the bounding box from two sides, we obtain the intervals of bounding box coordinates after any possible transformation.

Theorem 3. *Let $T = \{T_x, T_p\}$ be a transformation with parameter space $\mathcal{Z}$. Suppose $\mathcal{S} \subseteq \mathcal{Z}$ and $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$. Let $g_i : \mathcal{X} \times \mathcal{P} \to (\mathcal{B})_i$ be the i-th coordinate of a predicted bounding box of a multi-sensor fusion system, and $h_{iq}(\boldsymbol{x}, \boldsymbol{p})$ be the median smoothing of g_i as defined in Eq.* (1)*. Then for all transformations $z \in \mathcal{S}$, the i-th coordinate of the median smoothed bounding box predictor satisfies:*

$$\min_{1\leq i\leq M} h_{i\underline{q}}(T_x(\boldsymbol{x}, \alpha_i), T_p(\boldsymbol{p}, \alpha_i)) \leq h_{iq}(T_x(\boldsymbol{x}, \boldsymbol{z}), T_p(\boldsymbol{p}, \boldsymbol{z})) \leq \max_{1\leq i\leq M} h_{i\overline{q}}(T_x(\boldsymbol{x}, \alpha_i), T_p(\boldsymbol{p}, \alpha_i)) \tag{6}$$

where

$$\underline{q} = \Phi\left(\Phi^{-1}(q) - \sqrt{\frac{M_x^2}{\sigma_x^2} + \frac{M_p^2}{\sigma_p^2}}\right), \tag{7}$$

$$\overline{q} = \Phi\left(\Phi^{-1}(q) + \sqrt{\frac{M_x^2}{\sigma_x^2} + \frac{M_p^2}{\sigma_p^2}}\right). \tag{8}$$

with M_x, M_p defined as Eq. (5)*.*

With the intervals of bounding box coordinates, we propose the following theorem for computing the lower bound of IoU between the output bounding box and the ground truth.

Theorem 4. *Let* $\mathbf{B}$ *be a set of bounding boxes whose coordinates are bounded. We denote the lower bound of each coordinate by* $(\underline{x}, \underline{y}, \underline{z}, \underline{w}, \underline{h}, \underline{l}, \underline{r})$ *and upper bound by* $(\bar{x}, \bar{y}, \bar{z}, \bar{w}, \bar{h}, \bar{l}, \bar{r})$. *Let* $B_{gt} = (x, y, z, w, h, l, r)$ *be the ground truth bounding box. Then for any* $B_i \in \mathbf{B}$,

$$\mathrm{IoU}(B_i, B_{gt}) \geq \frac{h_1 \cdot (\underline{l}\underline{w} - \mathrm{Vol}(\underline{S} \backslash S_{gt}))}{hwl + \bar{h}\bar{w}\bar{l} - h_2 \cdot \big(\bar{l}\bar{w} - \mathrm{Vol}(\bar{S} \backslash S_{gt})\big)} \tag{9}$$

where $\underline{S}, \bar{S}$ *are convex hulls formed by* $(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r})$ *with respect to* $(\underline{w}, \underline{l})$ *and* $(\bar{w}, \bar{l})$ *(details in Appendix B.3), and* $S_{gt} = (x, z, w, l, r)_{gt}$ *is the projection of* B_{gt} *to the* $x - z$ *plane.*

$$h_1 = \max\big(\min_{y' \in [\underline{y}, \bar{y}]} \min\{h, \underline{h}, \frac{h + \underline{h}}{2} - |y' - y|\}, 0\big),$$
$$h_2 = \max\big(\min_{y' \in [\underline{y}, \bar{y}]} \min\{h, \bar{h}, \frac{h + \bar{h}}{2} - |y' - y|\}, 0\big). \tag{10}$$

Proof Sketch.. To prove a lower bound for the IoU between B_i and the ground truth B_{gt}, we lower bound the volume of the intersection $B_i \cap B_{gt}$ and upper bound the volume of the union $B_i \cup B_{gt}$ separately. We estimate the upper bound of union by $\mathrm{Vol}(B_{gt}) + \mathrm{Vol}(B_{\max}) - \min_{(x,y,z,r)} \mathrm{Vol}(B(x, y, z, \bar{w}, \bar{h}, \bar{l}, r) \cap B_{gt})$. We calculate h_1 and h_2 as the smallest possible intersection between B_i and B_{gt} along y axis given height $\underline{h}$ and $\bar{h}$, respectively. We then prove the lower bound of their intersection on the $x - z$ plane. We leverage the fact that $\mathrm{Vol}(S \cap S_{gt}) = \mathrm{Vol}(S) - \mathrm{Vol}(S \backslash S_{gt})$ and upper bound the volume of $S \backslash S_{gt}$ by considering the convex hull that contains all possible bounding boxes with bounded (x, z, r). The full proof is in Appendix B.3. □

We illustrate the computing procedures for both detection and IoU certification in Appendix C.

3.5 Instantiating Certification for Rotation and Shifting

In this section, we demonstrate how our certification framework works for concrete transformations. Specifically, we discuss two of the most common transformations for vehicles—rotation and shifting. For rotation transformation, we consider a vehicle rotating around the vertical axis based on bounded angle z within a radius r, i.e., $z \in [-r, r]$. For shifting transformation, we consider a vehicle moving along the road based on bounded distance $z \in [a, a + 2r]$ where a is the original distance.

We instantiate Theorem 1 on certifying detection and Theorem 4 on certifying IoU against both transformations by computing their interpolation errors M_x and M_p as defined in Equation (5). We choose $\{\alpha_i\}_{i=1}^K = \{\frac{2i-K}{K} r\}_{i=1}^K$ according to Lemma 2 and compute the interpolation errors M_x and M_p for both transformations. We then leverage M_x and M_p to derive the lower bound for the detecting confidence score and the IoU regarding the ground-truth bounding box based on Theorems 1 and 4.

Although the certification procedure can be time-consuming due to space partitioning, the certification cost usually happens before deployment (i.e., pre-deployment verification). After the model is deployed, the inference of smoothed inference is efficient (Cohen, Rosenfeld, and Kolter 2019). It is an active field to further reduce the inference cost (Horváth et al. 2022), and our framework can seamlessly integrate these advances.

4 Experimental Evaluation

We first construct a benchmark for evaluating certified robustness, then systematically evaluate our certification framework `COMMIT` on several state-of-the-art MSFs.

Dataset. There is no established benchmark for certified robustness evaluation for multi-sensor fusion systems to our knowledge. Hence, we construct a diverse dataset leveraging the CARLA simulator (Dosovitskiy et al. 2017). We consider two types of transformation: 1) Rotation transformation, which is common in the real world since the relative orientation of the car in front of the ego vehicle frequently changes. 2) Shifting transformation, which simulates the scenario where the distance between the front and the ego vehicle changes drastically within a short time.

We provide details of training, testing, and certification data construction as below.

- **Training and testing data.** We generate our KITTI-CARLA dataset (Deschaud 2021) with $5,000$ frames in CARLA Town01 with 50 pedestrians and 100 vehicles randomly spawned, in which $3,500$ frames are used for training and $1,500$ frames are used for testing.
- **Rotation certification data.** We spawn our ego vehicle at 15 spawn points randomly chosen in CARLA Town01, and we then spawn a leading vehicle in front of the ego vehicle within rotation interval $[-30°, 30°]$. This is because 30 degrees of car rotation largely cover the car rotation happening in the real world (cars changing lanes usually incur less than 30 degrees of rotation) and previous work (Li et al. 2021) mainly focuses on rotation within 30 degrees. Our framework can be further extended to a larger range easily. To study the effect of car color and surrounding objects on the rotation robustness, we collect our rotation certification data with 3 different colors of the leading vehicle in 4 different settings (combination of with or without buildings + with or without pedestrians), which is summarized in Table 2 in Appendix D.1.
- **Shifting certification data.** We spawn our ego vehicle at the same 15 spawn points as above and then spawn a leading vehicle facing forward in front of the ego vehicle. We choose $[10, 15]$ for the shifting intervals according to our empirical experiments. Specifically, we test the performance of our four models (MonoCon, SECOND, CLOCs, FocalsConv and MVX-Net) when the distance is smaller than 10 meters and we find that they always predict the car in front of the ego vehicle accurately enough. Therefore, we choose 10 meters as the starting point of our testing distance. We notice that the performance of all models drops tremendously when the distance reaches 15 meters. Thus, we choose 15 meters as our evaluation ending distance. Similar to the rotation certification data, we use the

same environment settings to study the effect of vehicle color, buildings, and pedestrians.

In total, in our benchmark dataset, the certification data contains 62 scenarios for rotation and 62 scenarios for shifting. We set the size of the image input to 64×87 following the standard setting.

Models. We choose two fusion models based on image and point clouds, which are highly ranked on the KITTI leaderboard: FocalsConv (Chen et al. 2022), CLOCs (Pang, Morris, and Radha 2020) and MVX-Net (Sindagi, Zhou, and Tuzel 2019). FocalsConv (Voxel R-CNN (Car) + multimodal) achieves 85.22% 3D Average Precision (AP) on the moderate KITTI Car detection task and 100% 3D AP on our KITTI-CARLA dataset. CLOCs (Faster RCNN (Ren et al. 2015) + SECOND (Yan, Mao, and Li 2018)) achieves 80.67% 3D AP on moderate KITTI Car detection task and 100% 3D AP on our KITTI-CARLA dataset. MVX-Net achieves 85.9% 3D AP on moderate KITTI Car detection task and 100% 3D AP on our KITTI-CARLA dataset.

To compare the performance between fusion models and single-modal models, we select a camera-based model–MonoCon (Liu, Xue, and Wu 2022), and a LiDAR-based model–SECOND (Yan, Mao, and Li 2018), which achieve 19.03% and 78.43% 3D AP respectively on the moderate KITTI Car detection task and 100% on our KITTI-CARLA moderate car detection task.

Metrics. We consider two metrics: detection rate and IoU. In detection certification, attackers aim to reduce the object detection confidence score to fool the detectors to detect nothing. We aim to certify the lower bound of the detection rate under a detection threshold, where **Det@80** means the ratio of detected bounding boxes with confidence score larger than 0.8. In IoU certification, we aim to lower bound the IoU between the detected bounding box and the ground truth bounding box when attackers are allowed to attack the IoU in a transformation space. As for the notation in all tables, **AP@50** means the ratio of detected bounding boxes whose IoU with the ground truth bounding boxes is larger than 0.5. In Figure 3 and Figure 7, we show corresponding results by choosing different detection or IoU thresholds.

Certification details. To make the models adapt with Gaussian noise smoothed data, we train two sets of models with Gaussian augmentation (Cohen, Rosenfeld, and Kolter 2019) using noise variance $\sigma = 0.25$ and $\sigma = 0.5$. For the ease of robustness certification, for rotation certification, we use models trained with $\sigma = 0.25$ to construct smoothed models; for shifting certification, we use models trained with $\sigma = 0.5$. Note that our framework allows using different σ and sample strategies for image and point cloud data.

4.1 Certification against Rotation Transformation

In this section, we present the evaluations for the certified and empirical results of our framework `COMMIT` against rotation transformation. In terms of certification, we use small intervals of rotation angles 0.1° and samples 1000 times with $\sigma_x = 0.25, \sigma_p = 0.25$ Gaussian noises for each interval (in total 600×1000 Gaussian noises with certification confidence 95%) to estimate h_q (see definition in Section 3.2). Empirically, we split the rotation intervals into small 0.01° and use the models' worst empirical performance in these samples as the empirical robustness against rotation attacks, which is equivalent to the PGD attack with step 0.01°. We set the overall confidence of certification to be 95%, aligning with the setting in (Kang et al. 2022).

Table 1a shows the results in the $[-30^\circ, 30^\circ]$ rotation interval. We can see that in terms of detection ability, the robustness order is FocalsConv $>$ MVX-Net $>$ MonoCon $>$ CLOCs $>$ SECOND, according to both the empirical and certification results. Hence, FocalsConv and MonoCon may be more likely to predict the existence of the object when it exists. Furthermore, we observe that multi-sensor fusion models have better detection robustness compared with single-modal models under the same threshold (e.g., FocalsConv $>$ MonoCon, CLOCs $>$ SECOND). In addition, we find that our certification is pretty tight in many cases. In particular, row "Certification" serves as the lower bound of row "Adv (Smoothed)", and in most cases they are very close, indicating the tight robustness certification.

Now we study the IoU metric since we expect that models can predict not only with high confidence but also precise bounding boxes. By comparing the empirical and certified results in IoU metric, CLOCs outperforms all other models. It is easy to understand that CLOCs outperforms single-modal models since it combines the information from both images and point clouds. However, FocalsConv is not as robust as CLOCs even though it is a also camera-LiDAR fusion model, which could be due to the fusion mechanism, and thus more robust fusion algorithms will help improve the model prediction robustness.

We also present some interesting findings of rotation transformation in Appendix D.2 (e.g., the effect of threshold, the effect of attack radius) and possible reasons for detection failure cases in Appendix D.5.

4.2 Certification against Shifting Transformation

Here we present the evaluations for the certified and empirical results of our framework `COMMIT` against shifting transformation. For robustness certification, we use small shifting intervals of size 0.01 and samples 1000 times with $\sigma_x = 0.5, \sigma_p = 0.5$ Gaussian noises in each interval to estimate h_q (see definition in Section 3.2). In the empirical experiments, we divide the shifting intervals into small intervals of size 0.001 and use the worst empirical performance of the model among these samples as the empirical robustness against PGD attacks. We set the overall confidence of certification to be 95% following the standard setting.

Table 1b shows the certified and empirical robustness of different models against shifting transformation. The robustness in terms of both detection ability and IoU is CLOCs $>$ MVX-Net $\approx$ SECOND $\approx$ MonoCon $>$ FocalsConv. We also notice that SECOND outperforms MonoCon when the distance is larger while they have a similar performance within short distances, which could be due to the accurate estimation of large distances by LiDAR sensors and the lack of depth information in the 2D camera images. However, this does not mean that image data do not have meaningful features because CLOCs is always more robust than SEC-

Model	Input Modality	Attack Radius	Benign		Adv (Vanilla)		Adv (Smoothed)		Certification	
			Det@80	**AP@50**	**Det@80**	**AP@50**	**Det@80**	**AP@50**	**Det@80**	**AP@50**
MonoCon (Liu, Xue, and Wu 2022)	Image	$\lvert r\rvert \leqslant 10°$	100.00%	100.00%	58.06%	56.45%	80.65%	82.26%	75.81%	0.00%
		$\lvert r\rvert \leqslant 15°$			58.06%	54.84%	80.65%	82.26%	75.81%	0.00%
		$\lvert r\rvert \leqslant 20°$			58.06%	53.23%	80.65%	74.19%	75.81%	0.00%
		$\lvert r\rvert \leqslant 25°$			45.16%	35.48%	80.65%	16.13%	75.81%	0.00%
		$\lvert r\rvert \leqslant 30°$			32.26%	0.00%	80.65%	3.23%	75.81%	0.00%
SECOND (Yan, Mao, and Li 2018)	Point Cloud	$\lvert r\rvert \leqslant 10°$	100.00%	100.00%	19.35%	96.77%	0.00%	100.00%	0.00%	100.00%
		$\lvert r\rvert \leqslant 15°$			19.35%	96.77%	0.00%	100.00%	0.00%	100.00%
		$\lvert r\rvert \leqslant 20°$			19.35%	96.77%	0.00%	100.00%	0.00%	100.00%
		$\lvert r\rvert \leqslant 25°$			1.61%	83.87%	0.00%	96.77%	0.00%	0.00%
		$\lvert r\rvert \leqslant 30°$			1.61%	51.61%	0.00%	54.84%	0.00%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	Image + Point Cloud	$\lvert r\rvert \leqslant 10°$	100.00%	100.00%	100.00%	90.32%	88.71%	100.00%	88.71%	100.00%
		$\lvert r\rvert \leqslant 15°$			100.00%	90.32%	66.13%	98.39%	66.13%	87.10%
		$\lvert r\rvert \leqslant 20°$			100.00%	88.71%	50.00%	98.39%	50.00%	69.35%
		$\lvert r\rvert \leqslant 25°$			20.97%	87.10%	50.00%	98.39%	50.00%	67.74%
		$\lvert r\rvert \leqslant 30°$			3.23%	80.65%	50.00%	98.39%	50.00%	53.23%
FocalsConv (Chen et al. 2022)	Image + Point Cloud	$\lvert r\rvert \leqslant 10°$	100.00%	100.00%	100.00%	96.77%	100.00%	100.00%	100.00%	0.00%
		$\lvert r\rvert \leqslant 15°$			100.00%	0.00%	100.00%	0.00%	100.00%	0.00%
		$\lvert r\rvert \leqslant 20°$			100.00%	0.00%	100.00%	0.00%	100.00%	0.00%
		$\lvert r\rvert \leqslant 25°$			100.00%	0.00%	100.00%	0.00%	100.00%	0.00%
		$\lvert r\rvert \leqslant 30°$			98.39%	0.00%	100.00%	0.00%	100.00%	0.00%
MVX-Net (Sindagi, Zhou, and Tuzel 2019)	Image + Point Cloud	$\lvert r\rvert \leqslant 10°$	100.00%	100.00%	100.00%	96.77%	100.00%	100.00%	100.00%	0.00%
		$\lvert r\rvert \leqslant 15°$			100.00%	96.77%	100.00%	100.00%	100.00%	0.00%
		$\lvert r\rvert \leqslant 20°$			90.32%	96.77%	100.00%	100.00%	100.00%	0.00%
		$\lvert r\rvert \leqslant 25°$			3.23%	75.81%	100.00%	100.00%	100.00%	0.00%
		$\lvert r\rvert \leqslant 30°$			3.23%	75.81%	93.55%	100.00%	99.71%	0.00%

(a) Rotation

Model	Input Modality	Attack Radius	Benign		Adv (Vanilla)		Adv (Smoothed)		Certification	
			Det@80	**AP@50**	**Det@80**	**AP@50**	**Det@80**	**AP@50**	**Det@80**	**AP@50**
MonoCon (Liu, Xue, and Wu 2022)	Image	$10 \leqslant z \leqslant 11$	100.00%	100.00%	66.13%	77.42%	66.13%	77.42%	64.52%	41.94%
		$10 \leqslant z \leqslant 12$			62.90%	74.19%	62.90%	74.19%	61.29%	1.61%
		$10 \leqslant z \leqslant 13$			56.45%	72.58%	56.45%	72.58%	51.61%	0.00%
		$10 \leqslant z \leqslant 14$			46.77%	33.87%	46.77%	33.87%	41.94%	0.00%
		$10 \leqslant z \leqslant 15$			27.42%	1.61%	27.42%	1.61%	27.42%	0.00%
SECOND (Yan, Mao, and Li 2018)	Point Cloud	$10 \leqslant z \leqslant 11$	100.00%	100.00%	0.00%	93.55%	0.00%	100.00%	0.00%	0.00%
		$10 \leqslant z \leqslant 12$			0.00%	80.65%	0.00%	80.65%	0.00%	0.00%
		$10 \leqslant z \leqslant 13$			0.00%	80.65%	0.00%	80.65%	0.00%	0.00%
		$10 \leqslant z \leqslant 14$			0.00%	80.65%	0.00%	80.65%	0.00%	0.00%
		$10 \leqslant z \leqslant 15$			0.00%	80.65%	0.00%	80.65%	0.00%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	Image + Point Cloud	$10 \leqslant z \leqslant 11$	100.00%	100.00%	100.00%	93.55%	93.55%	100.00%	67.74%	79.03%
		$10 \leqslant z \leqslant 12$			100.00%	80.65%	93.55%	80.65%	66.13%	51.61%
		$10 \leqslant z \leqslant 13$			85.48%	80.65%	88.71%	80.65%	64.52%	48.39%
		$10 \leqslant z \leqslant 14$			64.52%	80.65%	85.48%	80.65%	62.90%	48.39%
		$10 \leqslant z \leqslant 15$			64.52%	80.65%	83.87%	80.65%	61.29%	48.39%
FocalsConv (Chen et al. 2022)	Image + Point Cloud	$10 \leqslant z \leqslant 11$	100.00%	100.00%	96.77%	0.00%	96.77%	100.00%	54.84%	0.00%
		$10 \leqslant z \leqslant 12$			96.77%	0.00%	96.77%	100.00%	4.84%	0.00%
		$10 \leqslant z \leqslant 13$			0.00%	0.00%	0.00%	82.26%	0.00%	0.00%
		$10 \leqslant z \leqslant 14$			0.00%	0.00%	0.00%	14.52%	0.00%	0.00%
		$10 \leqslant z \leqslant 15$			0.00%	0.00%	0.00%	8.06%	0.00%	0.00%
MVX-Net (Sindagi, Zhou, and Tuzel 2019)	Image + Point Cloud	$10 \leqslant z \leqslant 11$	100.00%	100.00%	88.71%	96.77%	100.00%	100.00%	100.00%	0.00%
		$10 \leqslant z \leqslant 12$			88.71%	96.77%	100.00%	100.00%	98.39%	0.00%
		$10 \leqslant z \leqslant 13$			88.71%	96.77%	100.00%	100.00%	98.39%	0.00%
		$10 \leqslant z \leqslant 14$			88.71%	96.77%	100.00%	100.00%	96.77%	0.00%
		$10 \leqslant z \leqslant 15$			20.97%	96.77%	95.16%	100.00%	85.48%	0.00%

(b) Shifting

Table 1: Certified and empirical robustness of different models against different semantic transformations. 62 scenarios are evaluated for each transformations and we report the percentage of correct and (certifiably or empirically) robust scenarios. Each row represents the corresponding model and attack radius. "Benign", "Adv (Vanilla)", "Adv (Smoothed)", and "Certification" stand for benign performance, vanilla models' performance under attacks, smoothed models' performance under attacks, and certified lower bound of performance under bounded transformations. **Det@80** and **AP@50** mean that we use 0.8 and 0.5 as the thresholds of confidence score and IoU score. Results under other thresholds are in Appendix D.2.

OND, which could be caused by the fact that more candidates are proposed by 2D detectors (Faster RCNN in our case) which are ignored by the point cloud detectors. On the other hand, there is an interesting finding that FocalsConv performs poorly against shifting transformation. The reason might be that FocalsConv highly depends on the image features, and shifting transformation can attack the image and point cloud spaces at the same time. This implies that the design of the fusion mechanism is also an important factor on the robustness of multi-modal sensor fusion models. More findings and failure case analysis are in Appendix D.2 and Appendix D.5.

5 Conclusion

In this work, we provide the first robustness certification framework COMMIT for multi-sensor fusion systems against semantic transformations. Our theoretical certification framework is flexible for different models and transformations. Our evaluations show that current fusion models are more robust than single-modal models, and the design of the fusion mechanism is an important factor in improving the robustness against semantic transformations.

Acknowledgments

This work is partially supported by the National Science Foundation under grant No. 2046726, NSF AI Institute ACTION No. IIS-2229876, DARPA GARD, the National Aeronautics and Space Administration (NASA) under grant No. 80NSSC20M0229, ARL Grant W911NF-23-2-0137, the Alfred P. Sloan Fellowship, the Meta research award, the AI Safety Fund, and the eBay research award.

References

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12): 2481–2495.

Cao, Y.; Wang, N.; Xiao, C.; Yang, D.; Fang, J.; Yang, R.; Chen, Q.; Liu, M.; and Li, B. 2021. Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*, 1302–1320. Los Alamitos, CA, USA: IEEE Computer Society.

Carlini, N.; Tramer, F.; Dvijotham, K. D.; Rice, L.; Sun, M.; and Kolter, J. Z. 2023. (Certified!!) Adversarial Robustness for Free! In *The Eleventh International Conference on Learning Representations*.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1907–1915.

Chen, Y.; Li, Y.; Zhang, X.; Sun, J.; and Jia, J. 2022. Focal Sparse Convolutional Networks for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5428–5437.

Chiang, P.-y.; Curry, M.; Abdelkader, A.; Kumar, A.; Dickerson, J.; and Goldstein, T. 2020. Detection as regression: Certified object detection with median smoothing. *Advances in Neural Information Processing Systems*, 33: 1275–1286.

Chu, W.; Li, L.; and Li, B. 2022. TPC: Transformation-Specific Smoothing for Point Cloud Models. In *39th International Conference on Machine Learning (ICML 2022)*.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 1310–1320. PMLR.

Deschaud, J.-E. 2021. KITTI-CARLA: a KITTI-like dataset generated by CARLA Simulator. *arXiv e-prints*, arXiv:2109.00892.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, 1–16.

Engstrom, L.; Tran, B.; Tsipras, D.; Schmidt, L.; and Madry, A. 2019. Exploring the landscape of spatial robustness. In *International conference on machine learning*, 1802–1811. PMLR.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1625–1634.

Ghiasi, A.; Shafahi, A.; and Goldstein, T. 2020. BREAKING CERTIFIED DEFENSES: SEMANTIC ADVERSARIAL EXAMPLES WITH SPOOFED ROBUSTNESS CERTIFICATES. In *International Conference on Learning Representations*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guo, J.; Kurup, U.; and Shah, M. 2019. Is it safe to drive? An overview of factors, metrics, and datasets for driveability assessment in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(8): 3135–3151.

Hallyburton, R. S.; Liu, Y.; Cao, Y.; Mao, Z. M.; and Pajic, M. 2022. Security analysis of camera-lidar fusion against black-box attacks on autonomous vehicles. In *31st USENIX Security Symposium (USENIX SECURITY)*.

Hendrycks, D.; Carlini, N.; Schulman, J.; and Steinhardt, J. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Hendrycks, D.; and Dietterich, T. 2018. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Horváth, M. Z.; Mueller, M. N.; Fischer, M.; and Vechev, M. 2022. Boosting Randomized Smoothing with Variance Reduced Classifiers. In *International Conference on Learning Representations*.

Hosseini, H.; and Poovendran, R. 2018. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1614–1619.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17853–17862.

Huang, Q.; Dong, X.; Chen, D.; Zhou, H.; Zhang, W.; Zhang, K.; Hua, G.; and Yu, N. 2022. PointCAT: Contrastive Adversarial Training for Robust Point Cloud Recognition. *arXiv preprint arXiv:2209.07788*.

Kang, M.; Li, L.; Weber, M.; Liu, Y.; Zhang, C.; and Li, B. 2022. Certifying Some Distributional Fairness with Subpopulation Decomposition. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Kumar, A.; Levine, A.; Feizi, S.; and Goldstein, T. 2020. Certifying confidence via randomized smoothing. *Advances in Neural Information Processing Systems*, 33: 5165–5177.

Law, H.; and Deng, J. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, 734–750.

Li, L.; Weber, M.; Xu, X.; Rimanic, L.; Kailkhura, B.; Xie, T.; Zhang, C.; and Li, B. 2021. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 535–557.

Li, L.; Xie, T.; and Li, B. 2023. SoK: Certified Robustness for Deep Neural Networks. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, 22-26 May 2023*. IEEE.

Li, L.; Zhang, J.; Xie, T.; and Li, B. 2022. Double Sampling Randomized Smoothing. In *39th International Conference on Machine Learning (ICML 2022)*.

Liu, H.; and Lei, W. 2022. Attack Detection of Localization Based on Multi-Sensor Fusion in Autonomous Systems. In *2022 IEEE International Conference on Unmanned Systems (ICUS)*, 1333–1338. IEEE.

Liu, X.; Xue, N.; and Wu, T. 2022. Learning auxiliary monocular contexts helps monocular 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1810–1818.

Luo, W.; Yang, B.; and Urtasun, R. 2018. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 3569–3577.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

McCausland, P. 2019. Self-driving uber car that hit and killed woman did not recognize that pedestrians jaywalk.

Mueller, M. N.; Balunovic, M.; and Vechev, M. 2020. Certify or Predict: Boosting Certified Robustness with Compositional Architectures. In *International Conference on Learning Representations*.

Osinski, B.; Jakubowski, A.; Ziecina, P.; Milos, P.; Galias, C.; Homoceanu, S.; and Michalewski, H. 2020. Simulation-Based Reinforcement Learning for Real-World Autonomous Driving. In *ICRA*, 6411–6418.

Pang, S.; Morris, D.; and Radha, H. 2020. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10386–10393. IEEE.

Pei, K.; Cao, Y.; Yang, J.; and Jana, S. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, 1–18.

Prakash, A.; Chitta, K.; and Geiger, A. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7077–7087.

Qi, C. R.; Liu, W.; Wu, C.; Su, H.; and Guibas, L. J. 2018. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 918–927.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Samangouei, P.; Kabkab, M.; and Chellappa, R. 2018. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. In *International Conference on Learning Representations*.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32.

Shen, J.; Wang, N.; Wan, Z.; Luo, Y.; Sato, T.; Hu, Z.; Zhang, X.; Guo, S.; Zhong, Z.; Li, K.; et al. 2022. SoK: On the Semantic AI Security in Autonomous Driving. *arXiv preprint arXiv:2203.05314*.

Sindagi, V. A.; Zhou, Y.; and Tuzel, O. 2019. MVX-Net: Multimodal VoxelNet for 3D Object Detection. In *2019 International Conference on Robotics and Automation (ICRA)*, 7276–7282. IEEE.

Singla, S.; and Feizi, S. 2021. Skew orthogonal convolutions. In *International Conference on Machine Learning*, 9756–9766. PMLR.

Sun, J.; Cao, Y.; Chen, Q. A.; and Mao, Z. M. 2020. Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *USENIX Security Symposium (Usenix Security'20)*.

Sun, J.; Mehra, A.; Kailkhura, B.; Chen, P.-Y.; Hendrycks, D.; Hamm, J.; and Mao, Z. M. 2022. A Spectral View of Randomized Smoothing Under Common Corruptions: Benchmarking and Improving Certified Robustness. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 654–671. Springer.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Wong, E.; and Kolter, J. Z. 2021. Learning perturbation sets for robust machine learning. In *International Conference on Learning Representations*.

Wong, E.; and Kolter, Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 5286–5295. PMLR.

Xiao, C.; Zhu, J.-Y.; Li, B.; He, W.; Liu, M.; and Song, D. 2018. Spatially Transformed Adversarial Examples. In *International Conference on Learning Representations*.

Xu, C.; Ding, W.; Lyu, W.; Liu, Z.; Wang, S.; He, Y.; Hu, H.; Zhao, D.; and Li, B. 2022. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. *Advances in Neural Information Processing Systems*, 35: 25667–25682.

Xu, X.; Li, L.; and Li, B. 2022. LOT: Layer-wise Orthogonal Training on Improving l2 Certified Robustness. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yang, G.; Duan, T.; Hu, J. E.; Salman, H.; Razenshteyn, I.; and Li, J. 2020. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, 10693–10705. PMLR.

Zhang, B.; Jiang, D.; He, D.; and Wang, L. 2022a. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. In *Advances in Neural Information Processing Systems*.

Zhang, H.; Wang, S.; Xu, K.; Li, L.; Li, B.; Jana, S.; Hsieh, C.-J.; and Kolter, J. Z. 2022b. General Cutting Planes for Bound-Propagation-Based Neural Network Verification. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhao, H.; Qi, X.; Shen, X.; Shi, J.; and Jia, J. 2018. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, 405–420.

Zhong, Z.; Hu, Z.; Guo, S.; Zhang, X.; Zhong, Z.; and Ray, B. 2022. Detecting multi-sensor fusion errors in advanced driver-assistance systems. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, 493–505.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.