Robustness Certification of Visual Perception Models via Camera Motion Smoothing

Anonymous Author(s)
Affiliation
Address
email

Abstract: A vast literature shows that the learning-based visual perception model is sensitive to adversarial noises but few works consider the robustness of robotic perception models under widely-existing camera motion perturbations. To this end, we study the robustness of the visual perception model under camera motion perturbations to investigate the influence of camera motion on robotic perception. Specifically, we propose a motion smoothing technique for arbitrary image classification models, whose robustness under camera motion perturbations could be certified. The proposed robustness certification framework based on camera motion smoothing provides tight and scalable robustness guarantees for visual perception modules so that they are applicable to wide robotic applications. As far as we are aware, this is the first work to provide the robustness certification for the deep perception module against camera motions, which improve the trustworthiness of robotic perception. A realistic indoor robotic dataset with the dense point cloud map for the entire room, *MetaRoom*, is introduced for the challenging certifiable robust perception task. We conduct extensive experiments to validate the certification approach via motion smoothing against camera motion perturbations. Our framework guarantees the certified accuracy of 81.7% against camera translation perturbation along depth direction within -0.1m $\sim$ 0.1m.

Keywords: Robustness Certification, Camera Motion Perturbation, Robotic Perception

1 Introduction

Visual perception has achieved great success in recent years by leveraging the powerful representation capability of neural networks and the diversity of large datasets [1, 2]. Deep learning models have been dominating a wide range of computer vision tasks, such as image classification [3], object detection, [4] and segmentation [5]. However, applying deep perception models to real-world robotic applications is still challenging. Since visual perception is the core upstream module of an autonomous robot system, its failure can cause the robot to sense the surrounding environments incorrectly, which may result in catastrophic consequences. Therefore, developing a trustworthy perception system that can guarantee functionality in diverse real-world scenarios is necessary [6, 7, 8].

We study the robustness of a visual perception system against sensing noise due to camera motion perturbation that commonly exists in robotic applications [9, 10, 8], which is important while challenging for trustworthy robotic applications. The difficulty arises from two perspectives: internal model vulnerability and external sensing uncertainty. On the one hand, rich literature suggests that deep visual models are vulnerable to adversarial perturbations that are even stealthy to human eyes: small perturbations added to the input of neural networks can significantly corrupt the perception performance [11, 12, 13, 14]. On the other hand, several recent works indicate that the perception performance is also sensitive to data acquisition, such as sensor placement and sensing perspec-

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

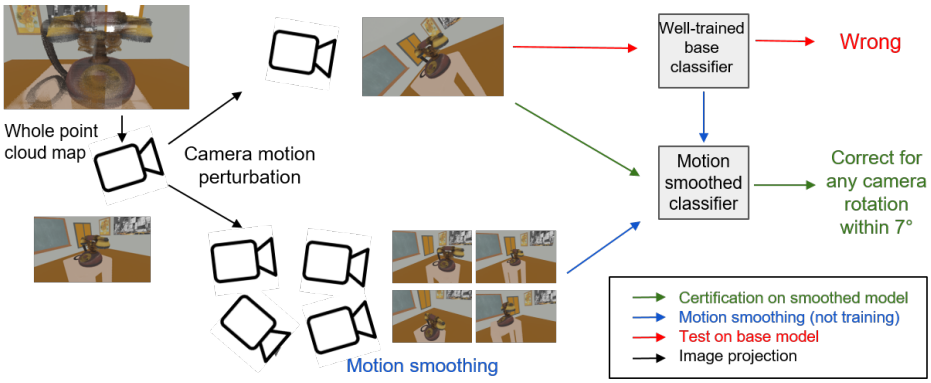

Figure 1: Certification framework via camera motion smoothing

tive [15, 16, 17, 18]. Both internal vulnerability and external sensing uncertainty make it challenging to guarantee the robustness of a visual perception system in real-world robotic applications.

Prior works propose several techniques to improve the visual perception system robustness [19, 20, 21], though most of them are demonstrated to be effective empirically and no theoretical robustness guarantees are given. A recent line of work aims to provide provable robustness certification or verification such that the perception system is guaranteed to function properly under any bounded adversarial perturbations [22, 23, 24] or semantic image transformations and deformations [25, 26, 27], which improves the trustworthiness of the perception model. However, most of them are studied under static 2D image datasets without viewpoint changes, while the robotic visual perception systems process 2D images projected from the 3D physical world, which may not be robust under different perspectives of the moving camera and make it challenging to certify the robustness for real-world robotic applications.

To tackle the challenge, we first study the robustness of the robotic perception model against the camera motion perturbation. Next, we propose the *first* framework with a certified robustness guarantee for robotic perception models against arbitrary bounded camera motion perturbations based on a novel motion smoothing strategy as shown in Fig. 1. Extensive experiments have been conducted on a realistic indoor robotic dataset *MetaRoom* with the dense point cloud map, which is collected from the Webots simulation environment [28] to show the effectiveness of the proposed robustness certification method against camera motion perturbations. To the best of our knowledge, this is the first work to study the certifiable robustness of robotic perception under camera ego-motion based on motion smoothing. The contributions are summarized as follows.

- We demonstrate that the camera motion perturbation can significantly influence perception performance and introduce motion smoothing to improve robustness over camera motion perturbation.
- We propose a smoothing algorithm for any black-box image classification model such that its robustness against camera motion perturbations can be certified by our certification framework.
- We conduct extensive experiments on the realistic indoor robotic dataset *MetaRoom* to validate the effectiveness of the motion smoothing certification, achieving over 80% certified accuracy against any perturbations within radii of 0.1m or 7° for camera translation or rotation along the depth axis.

2 Related Work

Robust Robotic Perception. The robustness of robotic perception has been studied from different viewpoints. A rich literature in the robust machine learning community shows that deep learning-based robotic perception models are vulnerable and can be easily fooled by adversarial samples [11, 12, 13, 14]. Another perspective for robotic perception is the external sensing uncertainty, which is caused by sensor placement [15, 17], camera distortions [18], geometric outliers [8], sim-to-real adaptation [29], etc. However, the robustness of deep perception models given the perturbed projected images from 3D physical world with a moving camera sensor is relatively understudied.

Certifiable Robustness under Perturbations. In recent years, a significant number of approaches has been proposed to provide robustness certification for deep neural networks [30, 31]. In con-

trast to empirical robustness approaches [32, 19, 20, 21], i.e., which train robust models against adversarial perturbations, the robustness certification approaches aim to guarantee the accuracy of the perception model as long as the perturbation magnitude is bounded by some threshold. Such robustness certification approaches have been proposed against both ℓ_p-bounded pixel-wise perturbations [22, 23, 24, 33, 34, 35] and semantic transformation or deformations [36, 25, 26, 27, 37], providing either deterministic guarantees based on function relaxations [37, 38, 39, 26] or high-confident probabilistic guarantees based on random smoothing [40, 25, 27, 41, 36]. However, they consider either 2D or 3D transformations. To the best of our knowledge, no prior work studies certifiable robustness associated with the movement of sensors and 3D-2D projected images, despite the fact that it is commonly seen in trustworthy robotic applications. Therefore, we aim to bridge the camera motion perturbation with deep learning robustness certification for robotic perception systems.

3 Methodology

In this section, we introduce the robustness certification framework against camera motion perturbations through the motion-smoothed perception model. We first define the image projection in terms of camera motion. Then, we clarify the certification goal and define the camera motion smoothed classifier using image projection. Finally, we present the robustness certification for each decomposed translation and rotation translation.

3.1 Image Projection with Camera Motion

We first define the positive projection in Def. 1 based on the camera imaging concept in geometric computer vision [42]. The follow-up Def. 2 defines the relative projective transformation $\phi(x, \alpha)$ parameterized by relative camera motion α with respect to the image x at motion origin.

Definition 1 (Position projective function)**.** *For any 3D point $P = (X, Y, Z) \in \mathbb{P} \subset \mathbb{R}^3$ under the camera coordinate frame with the camera intrinsic matrix K, based on the camera motion $\alpha = (\theta, t) \in \mathcal{Z}_\alpha \subset \mathbb{R}^6$ with rotation matrix $R = \exp(\theta^\wedge) \in SO(3)$ and translation vector $t \in \mathbb{R}^3$, we define the position projective function $\rho : \mathbb{P} \times \mathcal{Z}_\alpha \to \mathbb{R}^2$ and the depth function $D : \mathbb{P} \times \mathcal{Z}_\alpha \to \mathbb{R}$ for point P as*

$$[\rho(P, \alpha), 1]^\top = \frac{1}{D(P, \alpha)} K R^{-1}(P - t), \quad D(P, \alpha) = [0, 0, 1]^\top R^{-1}(P - t) \tag{1}$$

Definition 2 (Channel-wise projective transformation)**.** *Given the position projection function $\rho : \mathbb{P} \times \mathcal{Z}_\alpha \to \mathbb{R}^2$ and the depth function $D : \mathbb{P} \times \mathcal{Z}_\alpha \to \mathbb{R}$ over dense 3D point cloud $\mathbb{P}$, define the 3D-2D global channel-wise projective transformation from C-channel colored point cloud $\mathbb{V} = (\mathbb{R}^C, \mathbb{P}) \subset \mathbb{R}^6$ to $H \times W$ image gird $\mathcal{X} \subset \mathbb{R}^{C \times H \times W}$ as $O : \mathbb{V} \times \mathcal{Z}_\alpha \to \mathcal{X}$ parameterized by camera motion $\alpha \in \mathcal{Z}_\alpha$ using Floor function $\lfloor \cdot \rfloor$,*

$$x_{c,r,s} = O(V, \alpha)_{c,r,s} = V_{c,P^*_\alpha}, \textit{where } P^*_\alpha = \underset{\{P \in \mathbb{P} | \lfloor \rho(P,\alpha) \rfloor = (r,s)\}}{\operatorname{argmin}} D(P, \alpha) \tag{2}$$

Specifically, if $x = O(V, 0)$, we define the relative projective transformation $\phi : \mathcal{X} \times \mathcal{Z}_\alpha \to \mathcal{X}$ as,

$$\phi(x, \alpha) = O(V, \alpha). \tag{3}$$

3.2 Certification Goal and Motion Smoothed Classifier

Certification Goal. We consider the classification task as the most fundamental perception task. For any deep learning-based classification model h, given the projected image x at the origin of camera motion in the motion space $\mathcal{Z}_\alpha$, the certification goal is to find a set within a radius $\mathcal{Z}_{\text{radius}} \subseteq \mathcal{Z}_\alpha$ such that, for the relative projective transformation ϕ, with high confidence we have

$$h(x) = h(\phi(x, \alpha)), \forall \alpha \in \mathcal{Z}_{\text{radius}}. \tag{4}$$

Based on the relative projective transformation ϕ over camera motion space $\mathcal{Z}_\alpha$ defined above, we define the camera motion smoothed classification model as follows.

Definition 3 (Camera motion ε-smoothed classifier)**.** *Let $\phi : \mathcal{X} \times \mathcal{Z}_\alpha \to \mathcal{X}$ be a relative projective transformation given the projected image x at the origin of camera motion in the motion space $\mathcal{Z}_\alpha$, and let $\varepsilon \sim \mathcal{P}_\varepsilon$ be a random camera motion taking values in $\mathcal{Z}_\alpha$. Let $h : \mathcal{X} \to \mathcal{Y}$ be a base classifier $h(x) = \operatorname{argmax}_{y \in \mathcal{Y}} p(y \mid x)$, the expectation of projected image predictions $\phi(x, \varepsilon)$ over camera motion distribution $\mathcal{P}_\varepsilon$ is $q(y \mid x; \varepsilon) := \mathbb{E}_{\varepsilon \sim \mathcal{P}_\varepsilon} p(y \mid \phi(x, \varepsilon))$. We define the ε-smoothed classifier $g : \mathcal{X} \to \mathcal{Y}$ as*

$$g(x; \varepsilon) := \operatorname*{argmax}_{y \in \mathcal{Y}} q(y \mid x; \varepsilon) = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbb{E}_{\varepsilon \sim \mathcal{P}_\varepsilon} p(y \mid \phi(x, \varepsilon)). \tag{5}$$

3.3 Certifying Motion-parameterized Projection with Smoothed Classifier

We first show that the relative projection is compatible with the global projection as follows.

Lemma 1 (Compatible Relative Projection with Global Projection)**.** *With a global projective transformation $O : \mathbb{V} \times \mathcal{Z}_\alpha \to \mathcal{X}$ from 3D point cloud and a relative projective transformation $\phi : \mathcal{X} \times \mathcal{Z}_\alpha \to \mathcal{X}$ given some original camera motions, for any $\alpha_1 \in \mathcal{Z}_\alpha$ there exists an injective, continuously differentiable and non-vanishing-Jacobian function $\gamma_{\alpha_1} : \mathcal{Z}_\alpha \to \mathcal{Z}_\alpha$ such that*

$$\phi(O(V, \alpha_1), \alpha_2) = O(V, \gamma_{\alpha_1}(\alpha_2)), V \in \mathbb{V}, \alpha_2 \in \mathcal{Z}_\alpha. \tag{6}$$

The high-level idea for the proof is that $SO(3)$ has the rules of multiplication and inverse operation which can be transferred to the γ function in (6), and min-pooling defined in the projective transformation in (2) does not break these rules as well. The full proof can be found in the supplementary materials. Furthermore, if the camera motion is with translation and fix-axis rotation, the γ function in (6) will be additive, and then we have the following certification theorem.

Theorem 1 (Robustness certification under camera motion with fixed-axis rotation)**.** *Let $\alpha \in \mathcal{Z}_\alpha \subset \mathbb{R}^6$ be the parameters of projective transformation ϕ with translation $(t_x, t_y, t_z)^T \in \mathbb{R}^3$ and fixed-axis rotation $(\theta n_1, \theta n_2, \theta n_3)^T \in \mathbb{R}^3, \sum_{i=1}^3 n_i^2 = 1$, suppose the composed camera motion $\varepsilon_1 \in \mathcal{Z}_\alpha$ satisfies $\phi(x, \varepsilon_1) = \phi(\phi(x, \varepsilon_0), \alpha)$ given some $\alpha \in \mathcal{Z}_\alpha$ and zero-mean Gaussian motion ε_0 with variance $\sigma_x^2, \sigma_y^2, \sigma_z^2, \sigma_\theta^2$ for t_x, t_y, t_z, θ respectively, let $p_A, p_B \in [0, 1]$ be bounds of the top-2 class probabilities for the motion smoothed model, i.e.,*

$$q(y_A \mid x, \varepsilon_0) \geq p_A > p_B \geq \max_{y \neq y_A} q(y \mid \varepsilon_0). \tag{7}$$

Then, it holds that $g(\phi(x, \alpha); \varepsilon_0) = g(x; \varepsilon_0)$ if $\alpha = (t_x, t_y, t_z, \theta n_1, \theta n_2, \theta n_3)^T$ satisfies

$$\sqrt{\left(\frac{\theta}{\sigma_\theta}\right)^2 + \left(\frac{t_x}{\sigma_x}\right)^2 + \left(\frac{t_y}{\sigma_y}\right)^2 + \left(\frac{t_z}{\sigma_z}\right)^2} < \frac{1}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right). \tag{8}$$

The full proof can be found in the supplementary materials following previous certification work [22, 25]. We remark that for the general rotation without a fixed axis, although $\varepsilon_1 = \alpha + \varepsilon_0$ does not hold, the γ function can also be found based on multiplication in $SO(3)$.
Since all the camera motions can be regarded as the composition of each 1-axis translation or rotation, we make the following corollary to show the certification for any 1-axis translation or rotation.

Corollary 1 (Certification of 1-axis motion perturbation)**.** *For any 1-axis camera motion perturbation α with non-zero entry α_i satisfying $\alpha_i < \frac{\sigma_i}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right)$ under motion smoothing $\varepsilon_0 \sim \mathcal{N}(0, \sigma_i)$, it holds that $g(\phi(x, \alpha); \varepsilon_0) = g(x; \varepsilon_0)$ for the motion smoothed classifier g.*

Remark 1. *Cor. 1 directly follows Thm. 1 by taking only one non-zero entry in $\alpha \in \mathcal{Z}_\alpha \subset \mathbb{R}^6$, where each entry of in $\alpha = (t_x, t_y, t_z, \theta n_1, \theta n_2, \theta n_3)^T$ means T_x: translation along depth-orthogonal horizontal axis, T_y: translation along depth-orthogonal vertical axis, T_z: translation along depth axis, R_x: rotation around depth-orthogonal pitch axis, R_y: rotation around depth-orthogonal yaw axis and R_z: rotation around depth roll axis, as shown in Fig. 3.*

Therefore, we conduct comprehensive experiments for each 1-axis camera motion based on the Cor. 1 and Remark 1 for robustness certification.

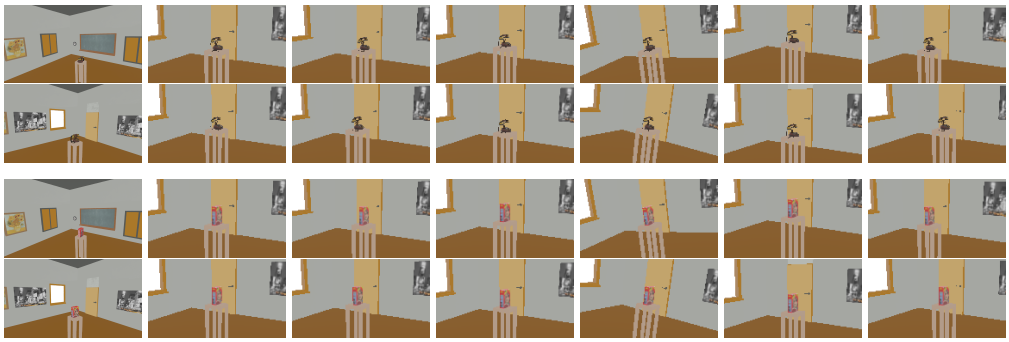

Figure 2: *MetaRoom* dataset with global point cloud map and projected images under 6-axis perturbations. From left to right: point cloud maps, perturbations over T_z, T_x, T_y, R_z, R_x and R_y.

4 Experiments

In this section, we aim to address two questions: 1) How does the perturbation of camera motion influence the perception performance empirically? 2) How can we certify the accuracy using the motion smoothed model under the camera perturbation within some radius? To answer these questions, we first set up a simulated indoor environment *MetaRoom* with dense point cloud maps and conduct extensive experiments based on it. Supplementary materials contain more experiment details.

4.1 Experimental Setup

MetaRoom Dataset. To implement the 3D-2D projection from point clouds to images through the camera, we first introduce the *MetaRoom* dataset based on Webots [28]. The *MetaRoom* dataset contains 20 commonly-seen indoor objects and for each object, we place it on a small table located in the center of an empty room with size of $3\,\mathrm{m} \times 3\,\mathrm{m}$. For each object, the collected data is associated with the point cloud of entire room, all the camera poses, and camera intrinsic parameters. We first reconstruct the $0.0025\,\mathrm{m}$-density point clouds of the room together with each object on the table through random snapshots from an RGBD camera with 500 camera poses for training and 120 camera poses for testing. The objects are $0.1\,\mathrm{m}$ higher than the origin of the global coordinate. For the training set, the camera is oriented toward the origin and uniformly randomly moves within a semi-spherical area of $2.3\,\mathrm{m} \sim 2.9\,\mathrm{m}$ radius with the yaw range of $360°$, the pitch range of $0° \sim 90°$ and roll range of $-60° \sim 60°$. The poses in the test set are $2.8\,\mathrm{m}$ away from the original with 6 fixed yaw angles uniformly distributed in $0° \sim 360°$, the pitch angle of $10°$, and the roll angle of $0°$. Note that the poses in the training set are at least $15°$ away from any yaw, pitch, or roll angle of every pose in the test set to avoid overlapping. The coordinate is illustrated in Fig. 3.

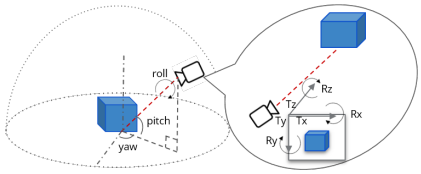

Figure 3: The illustration of coordinates of *MetaRoom* dataset and camera motion.

Model Training. We adopt two off-the-shelf representative architectures, ResNet-18 and ResNet-50 [3], to train base classifiers on *MetaRoom* for 100 epochs. For baselines, we adopt motion-specific augmentation as a defense to train the **Motion Augmented** models while **Vanilla** models are not trained with the augmentation since the augmented classifier could generalize well to the images with corresponding perturbation following [25, 41]. The motion augmentations are consistent with smoothing Gaussian distributions with σ shown in Tab. 2. Here motion augmented and vanilla are different strategies to train a **Base** classifier for **Smoothed** model via motion smoothing in Def. 3.

Evaluation Metrics. The **Benign Accuracy** is calculated as the ratio of correctly classified projected images without any perturbation, which is used to show the robustness/accuracy trade-off [22]. The robustness can be evaluated as the classification accuracy under worst-case perturbations. We uniformly sample 5 perturbed camera motions within the radius and consider the model is not robust if any of them is wrongly classified. We report the ratio of robust ones over the whole test set as **Empirical Robust Accuracy**, which can be viewed as the model's accuracy under the uniform-sampling-based perturbation. To provide a rigorous lower bound on the accuracy against any possible perturbations, we report **Certified Accuracy** [25, 41] to evaluate the certification re-

Camera Motion Types	Motion Augmented ResNet18	Motion Augmented ResNet50
T_z, radius [-0.1m, 0.1m]	Base / Smoothed	Base / Smoothed
Benign Accuracy	**0.858** / 0.842	**0.875** / 0.867
Empirical Robust Accuracy	0.817 / **0.833**	0.783 / **0.850**
T_x, radius [-0.05m, 0.05m]	Base / Smoothed	Base / Smoothed
Benign Accuracy	**0.925** / 0.900	**0.883** / 0.867
Empirical Robust Accuracy	0.817 / **0.875**	**0.825** / **0.825**
T_y, radius [-0.05m, 0.05m]	Base / Smoothed	Base / Smoothed
Benign Accuracy	**0.925** / 0.892	0.917 / **0.942**
Empirical Robust Accuracy	0.825 / **0.875**	0.792 / **0.925**
R_z, radius [-7°, 7°]	Base / Smoothed	Base / Smoothed
Benign Accuracy	0.933 / **0.958**	0.933 / **0.950**
Empirical Robust Accuracy	0.867 / **0.933**	0.875 / **0.917**
R_x, radius [-2.5°, 2.5°]	Base / Smoothed	Base / Smoothed
Benign Accuracy	**0.975** / 0.950	0.925 / **0.942**
Empirical Robust Accuracy	0.917 / **0.942**	0.858 / **0.933**
R_y, radius [-2.5°, 2.5°]	Base / Smoothed	Base / Smoothed
Benign Accuracy	0.917 / **0.925**	0.975 / **0.992**
Empirical Robust Accuracy	0.842 / **0.925**	0.892 / **0.992**

Table 1: The comparison between base and motion smoothed models which are both trained with motion augmentations in terms of benign and empirical robust accuracy for all camera motions. The higher one between each base and motion smoothed model is in **bold**.

sults, which is the fraction of test images that are both correctly classified and satisfy the certification condition of Equ. (8) within motion perturbation radius through smoothing, meaning models will predict correctly with at least this certified accuracy for any camera perturbation within the given motion radius with high confidence. Following convention [22], we use the confidence of 99% and 1000 smoothing camera motions under zero-mean Gaussian distribution for each test image.

4.2 Empirical Robustness against Camera Motion with Smoothing

First, we answer the question about the influence of camera motion perturbation on the perception performance empirically. For models trained with motion augmented defense, we compare the base and smoothed classifiers over benign and empirical robust accuracy in Tab. 1. It can be seen that under each motion perturbation, the empirical robust accuracy is lower than the benign accuracy for the base and smoothed models, showing that the perception models are not robust under motion perturbations even with the motion augmented defense when training. However, the gap between empirical robust accuracy and benign accuracy of motion smoothed models become less than those of base models due to the effective motion smoothing that improves robustness. Specifically, the motion smoothed models perform better in terms of empirical robust accuracy than the base models for camera motion perturbation along all axis. Interestingly, under a similar magnitude of radius, rotational smoothing will mostly increase the benign accuracy while translational smoothing tends to decrease the benign accuracy due to the robustness/accuracy trade-off [22].

4.3 Comparison of Certified Accuracy

Here we aim to show that how we certify the robust accuracy using the motion smoothing strategy against motion perturbation, so we compare the smoothed augmented models with the smoothed vanilla models in Tab. 2. For the motion perturbations on each axis, under the same smoothing strategy, most smoothed augmented models have better performance in benign, empirical robust and certified accuracy compared to smoothed vanilla models, showing that the motion augmentation improves the overall perception performance, empirical and certifiable robustness. Besides, the certified accuracy is only slightly lower than empirical robust accuracy under the same perturbation radius, which shows that our certification guarantees are tight and scalable against all camera motion perturbations for different baseline models.

4.4 Ablation Study

Certified accuracy under different radius. Fig. 4 shows the certified accuracy with respect to the motion perturbation radius on each axis. We find that as the certified radius increases, the

σ_z	T_z within radius [-0.1m, 0.1m]	Benign Acc	Emp. Robust Acc	Certified Acc
0.1m	Smoothed Vanilla ResNet-18	**0.858**	0.817	0.792
	Smoothed Augmented ResNet-18	0.842	**0.833**	**0.817**
	Smoothed Vanilla ResNet-50	0.675	0.617	0.558
	Smoothed Augmented ResNet-50	**0.867**	**0.850**	**0.817**
σ_x	T_x within radius [-0.05m, 0.05m]	Benign Acc	Emp. Robust Acc	Certified Acc
0.05m	Smoothed Vanilla ResNet-18	0.825	0.783	0.700
	Smoothed Augmented ResNet-18	**0.900**	**0.875**	**0.833**
	Smoothed Vanilla ResNet-50	0.767	0.675	0.508
	Smoothed Augmented ResNet-50	**0.867**	**0.825**	**0.708**
σ_y	T_y within radius [-0.05m, 0.05m]	Benign Acc	Emp. Robust Acc	Certified Acc
0.05m	Smoothed Vanilla ResNet-18	0.850	0.825	0.767
	Smoothed Augmented ResNet-18	**0.892**	**0.875**	**0.817**
	Smoothed Vanilla ResNet-50	0.792	0.767	0.683
	Smoothed Augmented ResNet-50	**0.942**	**0.925**	**0.892**
σ_θ	R_z within radius [-7°, 7°]	Benign Acc	Emp. Robust Acc	Certified Acc
7°	Smoothed Vanilla ResNet-18	0.817	0.742	0.608
	Smoothed Augmented ResNet-18	**0.958**	**0.933**	**0.883**
	Smoothed Vanilla ResNet-50	0.758	0.717	0.633
	Smoothed Augmented ResNet-50	**0.950**	**0.917**	**0.883**
σ_θ	R_x within radius [-2.5°, 2.5°]	Benign Acc	Emp. Robust Acc	Certified Acc
2.5°	Smoothed Vanilla ResNet-18	0.842	0.800	0.633
	Smoothed Augmented ResNet-18	**0.950**	**0.942**	**0.867**
	Smoothed Vanilla ResNet-50	0.767	0.742	0..567
	Smoothed Augmented ResNet-50	**0.942**	**0.933**	**0.867**
σ_θ	R_y within radius [-2.5°, 2.5°]	Benign Acc	Emp. Robust Acc	Certified Acc
2.5°	Smoothed Vanilla ResNet-18	0.892	0.875	0.708
	Smoothed Augmented ResNet-18	**0.925**	**0.925**	**0.917**
	Smoothed Vanilla ResNet-50	0.808	0.783	0.717
	Smoothed Augmented ResNet-50	**0.992**	**0.992**	**0.967**

Table 2: The comparison between smoothed vanilla and smoothed augmented models in terms of benign, empirical robust (Emp. Robust), and certified accuracy for with all camera motions. The higher one between vanilla and motion augmented models is in **bold**.

certified accuracy decreases and comes to 0 at some radius. Smoothed vanilla models have slowly decaying certified accuracy. For motion augmented models, the certified accuracies beyond the certified radii of *Translation z* and *Rotation y* still remain high until two times larger than σ in motion smoothing, while *Translation y* and *Rotation x* have quickly decayed certified accuracies, showing that *Translation z* and *Rotation y* can be better certified over larger perturbation radii.

Comparison of different model complexity. Smoothed vanilla ResNet-18 performs better in certified accuracy compared to ResNet-50 within smoothing radii σ. With motion augmented defense, although most empirical robust accuracy of base ResNet-50 is worse than base ResNet-18 in Tab. 1 and 3, ResNet-50 has better certified accuracy after motion smoothing under larger perturbation radii for most 1-axis camera motions in Fig. 4, which indicates that motion-augmented smoothed classifiers with larger complexity tend to be more certifiably robust to camera motion perturbations, although they may suffer from lower empirical robust accuracy due to overfitting.

Smoothed v.s. base vanilla model. For the undefended vanilla models, Tab. 3 presents that smoothed models have higher empirical robust accuracy and the gap between benign and empirical robust accuracy becomes less after motion smoothing compared to the base models, showing that smoothing strategy works not only for well-defended motion augmented models, but also for undefended vanilla models. Compared to Tab. 1, there is more robustness/accuracy trade-off in rotation for the vanilla models, which implies that motion augmentation as a defense in model training helps to improve the robustness better against rotational perturbations than translational perturbations.

4.5 Limitation and Discussion

One limitation of this work is that our current certification framework is only evaluated on image classification tasks, although it is fundamental for other robot applications. In addition, the robust-

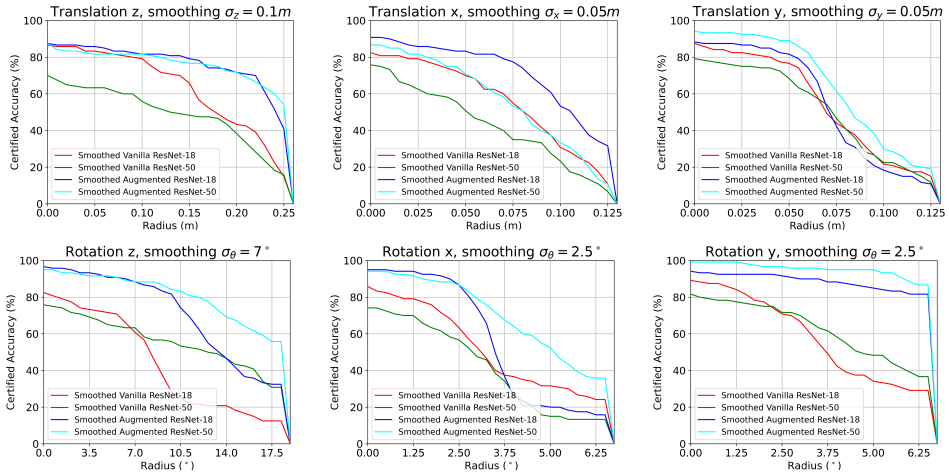

Figure 4: Certified accuracy with respect to radius for all ego-motions

Camera Motion Types	Vanilla ResNet18	Vanilla ResNet50
T_z, radius [-0.1m, 0.1m]	Base / Smoothed	Base / Smoothed
Benign Accuracy	0.800 / **0.858**	**0.708** / 0.675
Empirical Robust Accuracy	0.700 / **0.817**	0.558 / **0.617**
T_x, radius [-0.05m, 0.05m]	Base / Smoothed	Base / Smoothed
Benign Accuracy	0.817 / **0.825**	0.717 / **0.767**
Empirical Robust Accuracy	0.675 / **0.783**	0.567 / **0.675**
T_y, radius [-0.05m, 0.05m]	Base / Smoothed	Base / Smoothed
Benign Accuracy	0.825 / **0.850**	**0.817** / 0.792
Empirical Robust Accuracy	0.717 / **0.825**	0.667 / **0.767**
R_z, radius [-7°, 7°]	Base / Smoothed	Base / Smoothed
Benign Accuracy	0.800 / **0.817**	**0.783** / 0.758
Empirical Robust Accuracy	0.650 / **0.742**	0.600 / **0.717**
R_x, radius [-2.5°, 2.5°]	Base / Smoothed	Base / Smoothed
Benign Accuracy	**0.867** / 0.842	**0.767** / **0.767**
Empirical Robust Accuracy	0.667 / **0.800**	0.585 / **0.742**
R_y, radius [-2.5°, 2.5°]	Base / Smoothed	Base / Smoothed
Benign Accuracy	**0.917** / 0.892	0.800 / **0.808**
Empirical Robust Accuracy	0.700 / **0.875**	0.642 / **0.783**

Table 3: The comparison between base vanilla models and smoothed vanilla models in benign and empirical robust accuracy for all camera motions. The higher one between each base and smoothed model is in **bold**.

ness certification procedure costs many computational resources to obtain a tight guarantee and can be addressed through differential certifications [30, 25] as future work. Finally, the current certification framework is built upon the camera sensor, and it would be interesting to extend the current work to other perception sensors in robotics, such as 3D LiDAR.

5 Conclusion

In this work, we study the robust visual perception against camera motion perturbation as the projective transformation from 3D to 2D. We propose a robustness certification framework via a camera motion smoothing approach to provide robust guarantees for image classification models for real-world robotic applications. We collect a realistic indoor robotic dataset MetaRoom with the dense point cloud map for robustness certification against camera motion perturbation. We conduct extensive experiments to compare the empirical robust accuracy with the certified robust accuracy for the motion smoothed model within large radii of 6-axis camera motion perturbations, i.e., translations and rotations, to guarantee the lower bound of accuracy for trustworthy robotic perception.

References

[1] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

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

[7] H. Hu, B. Yang, Z. Qiao, D. Zhao, and H. Wang. Seasondepth: Cross-season monocular depth prediction dataset and benchmark under multiple environments. *arXiv preprint arXiv:2011.04408*, 2021.

[8] H. Yang and L. Carlone. Certifiably optimal outlier-robust geometric perception: Semidefinite relaxations and scalable global optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[9] J. F. Ferreira and J. M. Dias. *Probabilistic approaches to robotic perception*. Springer, 2014.

[10] G. Liu, A. Amini, M. Takac, and N. Motee. Robustness analysis of classification using recurrent neural networks with perturbed sequential input. *arXiv preprint arXiv:2203.05403*, 2022.

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

[12] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

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

[14] C. Xiao, B. Li, J. Y. Zhu, W. He, M. Liu, and D. Song. Generating adversarial examples with adversarial networks. In *27th International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 3905–3911. International Joint Conferences on Artificial Intelligence, 2018.

[15] Z. Liu, M. Arief, and D. Zhao. Where should we place lidars on the autonomous vehicle?-an optimal design approach. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2793–2799. IEEE, 2019.

[16] J. Feng, J. Lee, M. Durner, and R. Triebel. Bridging the last mile in sim-to-real robot perception via bayesian active learning. *arXiv preprint arXiv:2109.11547*, 2021.

[17] H. Hu, Z. Liu, S. Chitlangia, A. Agnihotri, and D. Zhao. Investigating the impact of multi-lidar placement on object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2550–2559, 2022.

[18] Z. Tang, R. G. Von Gioi, P. Monasse, and J.-M. Morel. A precision analysis of camera distortion models. *IEEE Transactions on Image Processing*, 26(6):2694–2704, 2017.

[19] F. Tramèr, D. Boneh, A. Kurakin, I. Goodfellow, N. Papernot, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, 2018.

[20] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

[21] F. Tramer, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645, 2020.

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

[23] V. Tjeng, K. Y. Xiao, and R. Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2018.

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

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

[26] A. Ruoss, M. Baader, M. Balunović, and M. Vechev. Efficient certification of spatial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2504–2513, 2021.

[27] M. Alfarra, A. Bibi, N. Khan, P. H. Torr, and B. Ghanem. Deformrs: Certifying input deformations with randomized smoothing. *arXiv preprint arXiv:2107.00996*, 2021.

[28] Webots. http://www.cyberbotics.com. URL http://www.cyberbotics.com. Open-source Mobile Robot Simulation Software.

[29] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[30] L. Li, T. Xie, and B. Li. Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*, 2020.

[31] C. Liu, T. Arnon, C. Lazarus, C. Barrett, and M. J. Kochenderfer. Algorithms for verifying deep neural networks. *arXiv preprint arXiv:1903.06758*, 2019.

[32] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[33] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in neural information processing systems*, pages 4939–4948, 2018.

[34] G. Singh, T. Gehr, M. Püschel, and M. Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):41, 2019.

[35] S. Dathathri, K. Dvijotham, A. Kurakin, A. Raghunathan, J. Uesato, R. R. Bunel, S. Shankar, J. Steinhardt, I. Goodfellow, P. S. Liang, and P. Kohli. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5318–5331, 2020.

[36] H. Zhongkai, C. Ying, Y. Dong, H. Su, and J. Zhu. GSmooth: Certified robustness against semantic transformations via generalized randomized smoothing. In *International Conference on Machine Learning*. PMLR, 2022.

[37] M. Balunović, M. Baader, G. Singh, T. Gehr, and M. Vechev. Certifying geometric robustness of neural networks. *Advances in Neural Information Processing Systems 32*, 2019.

[38] J. Mohapatra, T.-W. Weng, P.-Y. Chen, S. Liu, and L. Daniel. Towards verifying robustness of neural networks against a family of semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–252, 2020.

[39] T. Lorenz, A. Ruoss, M. Balunović, G. Singh, and M. Vechev. Robustness certification for point cloud models. *arXiv preprint arXiv:2103.16652*, 2021.

[40] M. Fischer, M. Baader, and M. T. Vechev. Certified defense to image transformations via randomized smoothing. In *NeurIPS*, 2020.

[41] W. Chu, L. Li, and B. Li. Tpc: Transformation-specific smoothing for point cloud models. In *International Conference on Machine Learning*. PMLR, 2022.

[42] R. Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.