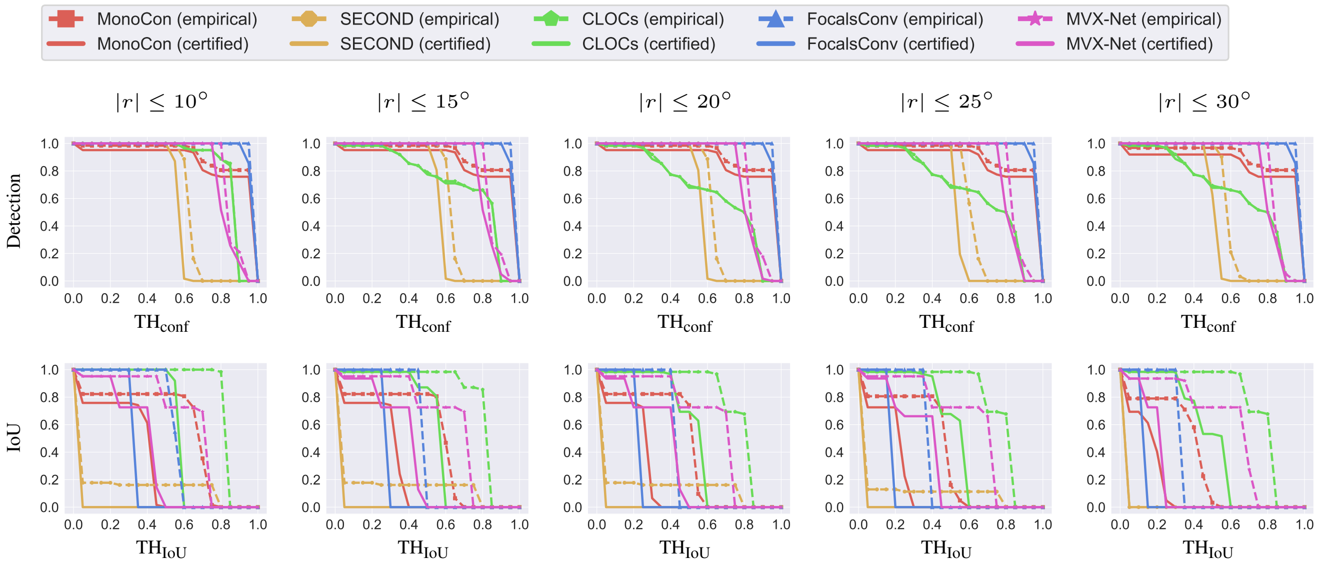


Figure 3: Certified and empirical robustness on detection rate and IoU against rotation transformation (smoothing $\sigma = 0.25$) under different thresholds. Solid lines represent the certified bounds, and dashed lines show the empirical performance under PGD attacks. x-axis represents the threshold for confidence score ($\mathrm{TH}_{\mathrm{conf}}$) and IoU score ($\mathrm{TH}_{\mathrm{IoU}}$), and y-axis represents the ratio of detection whose confidence / IoU score is larger than the confidence / IoU threshold.

The appendices are organized as follows:

- In Appendix A, we formally present the finite partition assumption (from Section 3.1) and provide an empirical verification of the assumption.
- In Appendix B, we present the detailed proofs for lemmas and theorems in Sections 3.3 and 3.4 of the main paper.
- In Appendix C, we present some additional details of our two certification strategies: Appendix C.1 for certifying the detection rate, and Appendix C.2 for certifying the IoU between the detection and the ground truth. We include the detailed algorithm description and the complete pseudocode for each algorithm. We will make our implementation public upon acceptance.
- In Appendix D, we show dataset details (Appendix D.1), detailed experimental evaluation (Appendix D.2), some ablation studies on sample strategies (Appendix D.3) and smoothing parameter σ (Appendix D.4), and failure case analysis (Appendix D.5).
- In Appendix E, we do some side discussions, which includes potential limitations of our method (Appendix E.1) and the connection between our method and other trustworthy research directions (Appendix E.2).

A Details of Fine Partition Assumption

As introduced in Section 3.1, we impose the following assumption for the transformation.

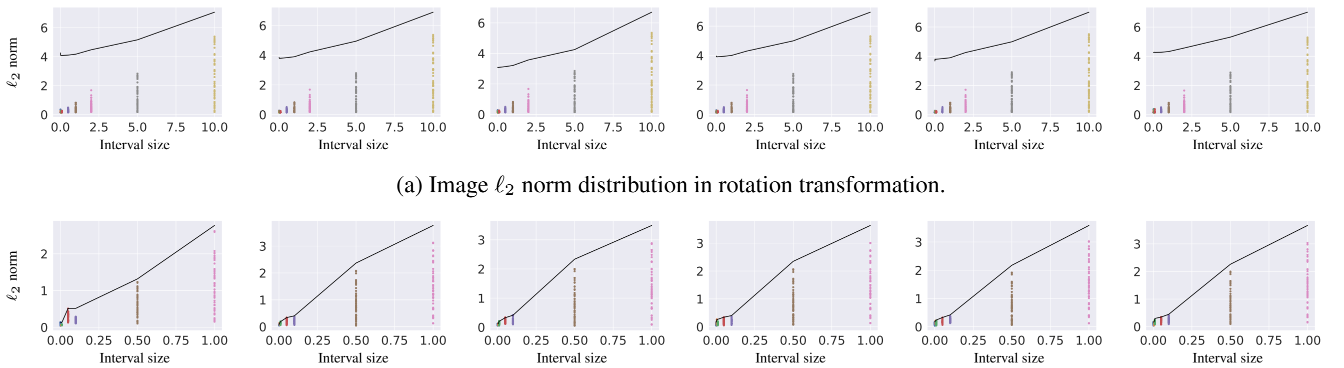
Assumption 5. For given transformation $T = \{T_x, T_p\}$ with parameter space $\mathcal{Z} \subseteq \mathbb{R}^m$, there exists a small threshold $\tau > 0$, for any polytope of the parameter space $\mathcal{Z}_{\mathrm{sub}} \subseteq \mathcal{Z}$ whose ℓ_∞ diameter is smaller than τ, i.e., $\mathrm{diam}_\infty(\mathcal{Z}_{\mathrm{sub}}) < \tau$, when parameters are picked from the subspace, the pairwise ℓ_2 distance between transformed outputs is upper bounded by maximum pairwise ℓ_2 distance with extreme points picked as parameters. Formally, let $E(\mathcal{Z}_{\mathrm{sub}})$ be the set of extreme points of $\mathcal{Z}_{\mathrm{sub}}$, then $\forall \boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}_{\mathrm{sub}}, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{p} \in \mathcal{P}$,

$$\begin{aligned} \|T_x(\boldsymbol{x}, \boldsymbol{z}_1) - T_x(\boldsymbol{x}, \boldsymbol{z}_2)\|_2 &\leq \max_{\boldsymbol{z}_1', \boldsymbol{z}_2' \in E(\mathcal{Z}_{\mathrm{sub}})} \|T_x(\boldsymbol{x}, \boldsymbol{z}_1') - T_x(\boldsymbol{x}, \boldsymbol{z}_2')\|_2, \\ \|T_p(\boldsymbol{p}, \boldsymbol{z}_1) - T_p(\boldsymbol{p}, \boldsymbol{z}_2)\|_2 &\leq \max_{\boldsymbol{z}_1', \boldsymbol{z}_2' \in E(\mathcal{Z}_{\mathrm{sub}})} \|T_p(\boldsymbol{p}, \boldsymbol{z}_1') - T_p(\boldsymbol{p}, \boldsymbol{z}_2')\|_2. \end{aligned} \tag{11}$$

Remark 3. Intuitively, the assumption states that, within a tiny subspace of the parameter space, the displacement incurred by the transformation, when measured by Euclidean distance, is proportional to the magnitude between parameters, so the maximum displacement can be upper bounded by the displacement incurred by choosing extreme points as the transformation parameters. Taking the rotation as an example, within a sufficiently small range of rotation angle $[r - \Delta, r + \Delta]$ where $2\Delta < \tau$, the assumption means that, the difference between rotated images $\|T_x(\boldsymbol{x}, \delta_1) - T_x(\boldsymbol{x}, \delta_2)\|_2$ and point clouds $\|T_p(\boldsymbol{p}, \delta_1) - T_p(\boldsymbol{p}, \delta_2)\|_2$ is no larger than $\|T_x(\boldsymbol{x}, r - \Delta) - T_x(\boldsymbol{x}, r + \Delta)\|_2$ and $\|T_p(\boldsymbol{p}, r - \Delta) - T_p(\boldsymbol{p}, r + \Delta)\|_2$ respectively.

While it is hard to prove the Assumption 5, we empirically evaluate the Assumption 5 by plotting the distribution of image ℓ_2 norm with different interval sizes ($0.001°, 0.01°, 0.02°, 0.03°, 0.04°, 0.05°$ for rotation and $0.001, 0.01, 0.02, 0.03, 0.04, 0.05$ for shifting) in randomly selected big intervals ($0.06°$ for rotation and 0.07 for shifting) in Figure 4.

From Figure 4a and Figure 4b, we can notice that with larger rotation and shifting intervals, the image ℓ_2 norm becomes larger and larger, and the ℓ_2 distance between the endpoints of each big interval can bound the ℓ_2 distance between randomly chosen points in that big interval, which means that the pairwise ℓ_2 distance picks the maximum value with extreme points when the transformation intervals are sufficiently small, and thus Assumption 5 is empirically confirmed.



(a) Image ℓ_2 norm distribution in rotation transformation.

(b) Image ℓ_2 norm distribution in shifting transformation.

Figure 4: Image ℓ_2 norm distribution in rotation and shift transformation. Images are from spawn point 15, 30, 43, 46, 57 and 86 from dataset with building and without pedestrian. The scatter plots show the ℓ_2 distance of randomly chosen pairs in randomly chosen big intervals. The black line is the ℓ_2 distance between the endpoints of big intervals.

B Proofs

B.1 Proof of Theorem 1: Detection Certification

In this section, we present the full proof of Theorem 1, which provides generic certification for multi-sensor fusion detection against an abstract transformation. We first restate this theorem from the main text.

Theorem 1 (restated). *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 median smoothed detector satisfies:*

$$h_q(T_x(\boldsymbol{x}, \boldsymbol{z}), T_p(\boldsymbol{p}, \boldsymbol{z})) \geq \min_{1\leq i\leq M} h_{\underline{q}}(T_x(\boldsymbol{x}, \alpha_i), T_p(\boldsymbol{p}, \alpha_i)) \tag{12}$$

where

$$\underline{q} = \Phi\left(\Phi^{-1}(q) - \sqrt{\frac{M_x^2}{\sigma_x^2} + \frac{M_p^2}{\sigma_p^2}}\right), \tag{13}$$

$$M_x = \max_{\alpha\in\mathcal{S}} \min_{1\leq i\leq M} \|T_x(\boldsymbol{x}, \alpha) - T_x(\boldsymbol{x}, \alpha_i)\|_2, \tag{14}$$

$$M_p = \max_{\alpha\in\mathcal{S}} \min_{1\leq i\leq M} \|T_p(\boldsymbol{p}, \alpha) - T_p(\boldsymbol{p}, \alpha_i)\|_2. \tag{15}$$

Proof. We first recall the median smoothed classifier $h_q(\boldsymbol{x}, \boldsymbol{p}) = \sup\{y \in \mathbb{R} \mid \Pr[g(\boldsymbol{x} + \delta_x, \boldsymbol{p} + \delta_p) \leq y] \leq q\}$, where $\delta_x \sim \mathcal{N}(0, \sigma_x^2\mathbf{I}_d)$ and $\delta_p \sim \mathcal{N}(0, \sigma_p^2\mathbf{I}_{3\times N})$.

Consider a function $f : \mathbb{R}^d \times \mathbb{R}^{3\times N} \to [0, 1]$ with $f(\boldsymbol{x}, \boldsymbol{p}) = \Pr[g(\boldsymbol{x}+\delta_x, \boldsymbol{p}+\delta_p) \leq h_{\underline{q}}(\boldsymbol{x}, \boldsymbol{p})]$. We define $\tilde{g} : \mathcal{X} \times \mathcal{P} \to [0, 1]$ as $\tilde{g}(\boldsymbol{x}, \boldsymbol{p}) = g(\boldsymbol{x}, \frac{\sigma_p}{\sigma_x}\boldsymbol{p})$. Then

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{p}) &= \Pr[\tilde{g}(\boldsymbol{x} + \delta_x, \boldsymbol{p}' + \delta_p') \leq h_{\underline{q}}(\boldsymbol{x}, \boldsymbol{p})], \\ \text{where } \boldsymbol{p}' &= \frac{\sigma_x}{\sigma_p}\boldsymbol{p},\ \delta_p' \sim \mathcal{N}(0, \sigma_x^2\mathbf{I}_{3\times N}). \end{aligned} \tag{16}$$

We now introduce a lemma from (Salman et al. 2019)(Lemma 2) and (Chiang et al. 2020)(Corollary 1).

Lemma 6. *For any $g : \mathbb{R}^d \to [l, u]$, let $f(x) = \mathbb{E}[g(x + G)]$ where $G \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. Then the map $\eta(x) = \sigma \cdot \Phi^{-1}(\frac{f(x)-l}{u-l})$ is 1-Lipschitz.*

Note that $f(\boldsymbol{x}, \boldsymbol{p}') = \mathbb{E}\left(\mathbb{1}[\tilde{g}(\boldsymbol{x} + \delta_x, \boldsymbol{p}' + \delta_p') \le h_{\underline{q}}(\boldsymbol{x}, \boldsymbol{p})]\right)$, which means $\sigma_x \cdot \Phi^{-1}(f(\boldsymbol{x}, \boldsymbol{p}'))$ is 1-Lipschitz.
Now let us consider an arbitrary transformation $\boldsymbol{z} \in \mathcal{S}$, by the definition of M_x and M_p we have

$$\begin{aligned} \forall \boldsymbol{z} \in \mathcal{S},\ \exists \alpha_i,\ &\|T_x(\boldsymbol{x}, \boldsymbol{z}) - T_x(\boldsymbol{x}, \alpha_i)\|_2 \le M_x, \\ &\|T_p(\boldsymbol{p}, \boldsymbol{z}) - T_p(\boldsymbol{p}, \alpha_i)\|_2 \le M_p. \end{aligned} \tag{17}$$

Then

$$\begin{aligned} &\sigma_x \Phi^{-1}\Big(\Pr\Big[g(T_x(\boldsymbol{x}, \boldsymbol{z}) + \delta_x, T_p(\boldsymbol{p}, \boldsymbol{z}) + \delta_p) \le h_{\underline{q}}(\boldsymbol{x}, \boldsymbol{p}) \Big] \Big) \\ &= \sigma_x \Phi^{-1}\left(f(T_x(\boldsymbol{x}, \boldsymbol{z}), \frac{\sigma_x}{\sigma_p} T_p(\boldsymbol{p}, \boldsymbol{z})) \right) \\ &\ge \sigma_x \Phi^{-1}\left(f\Big(T_x(\boldsymbol{x}, \alpha_i), \frac{\sigma_x}{\sigma_p} T_p(\boldsymbol{p}, \alpha_i)\Big) \right) \\ &\qquad - \sqrt{M_x^2 + \frac{\sigma_x^2}{\sigma_p^2} M_p^2} \\ &\ge \sigma_x \Phi^{-1}\Big(\Pr\Big[g(T_x(\boldsymbol{x}, \alpha_i) + \delta_x, T_p(\boldsymbol{p}, \alpha_i) + \delta_p) \le h_{\underline{q}}(\boldsymbol{x}, \boldsymbol{p}) \Big] \Big) \\ &\qquad - \sqrt{M_x^2 + \frac{\sigma_x^2}{\sigma_p^2} M_p^2} \\ &= \sigma_x \Phi^{-1}(\underline{q}) - \sqrt{M_x^2 + \frac{\sigma_x^2}{\sigma_p^2} M_p^2} \\ &= \sigma_x \Phi^{-1}(q). \end{aligned} \tag{18}$$

Since that $\Phi(\cdot)$ is monotonic, we conclude that

$$\forall \boldsymbol{z} \in \mathcal{S},\ \exists \alpha_i, \text{s.t. } h_q(T_x(\boldsymbol{x}, \boldsymbol{z}), T_p(\boldsymbol{p}, \boldsymbol{z})) \ge h_{\underline{q}}(\boldsymbol{x}, \boldsymbol{p}). \tag{19}$$

□

B.2 Proof of Lemma 2: Upper Bound for the Interpolation Error

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

$$M_x \le \sum_{i=1}^m \max_{\boldsymbol{k} \in \Delta} \Big\| T_x(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{k})) - T_x(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{k}) + w_i) \Big\|_2, \tag{20}$$

$$M_p \le \sum_{i=1}^m \max_{\boldsymbol{k} \in \Delta} \Big\| T_p(\boldsymbol{p}, \boldsymbol{w}(\boldsymbol{k})) - T_p(\boldsymbol{p}, \boldsymbol{w}(\boldsymbol{k}) + w_i) \Big\|_2 \tag{21}$$

where $\Delta = \{(k_1, \ldots, k_m) \in \mathbb{Z}^m \mid 0 \le k_i < K_i\}$ 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.

Proof. Let $\boldsymbol{z} \in \mathcal{S} \subseteq \mathbb{R}^m$ be a parameter in the parameter space to certify and $\boldsymbol{z} = (z_1, \ldots, z_m)$. There must be $(k_1, k_2, \ldots, k_m)$ with $k_i \in \{0, 1, \ldots, K_i - 1\}$, such that

$$\begin{aligned} \frac{K_i - k_i}{K_i} l_i + \frac{k_i}{K_i} u_i \le z_i \le \frac{K_i - k_i - 1}{K_i} l_i + \frac{k_i + 1}{K_i} u_i, \\ \forall i = 1, 2, \ldots, m. \end{aligned} \tag{22}$$

Let $\boldsymbol{k} = (k_1, k_2, \ldots, k_m)$. Consider the small polytope $\mathcal{Z}_{\text{sub}} = \boldsymbol{k} + [0, 1]^m \cdot (w_1, \ldots, w_m)$. By Assumption 5,

$$\|T_x(\boldsymbol{x}, \boldsymbol{z}) - T_x(\boldsymbol{x}, \boldsymbol{w}(k))\|_2 \tag{23}$$

$$\le \max_{\alpha, \beta \in E(\mathcal{Z}_{\text{sub}})} \|T_x(\boldsymbol{x}, \alpha) - T_x(\boldsymbol{x}, \beta)\|_2 \tag{24}$$

$$= \|T_x(\boldsymbol{x}, \boldsymbol{z}_1) - T_x(\boldsymbol{x}, \boldsymbol{z}_2)\|_2 \tag{25}$$

where $\boldsymbol{z}_1 = \boldsymbol{k} + I_1 \cdot (w_1, \dots, w_m)$ and $\boldsymbol{z}_2 = \boldsymbol{k} + I_2 \cdot (w_1, \dots, w_m)$ for some $I_1, I_2 \in \{0,1\}^m$. Let $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ be a shortest path from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$ such that $\alpha_i \in \{0,1\}^m \cdot (w_1, \dots, w_m) + \boldsymbol{k}$, $\alpha_1 = \boldsymbol{k}_1$, and $\alpha_n = \boldsymbol{k}_2$. Moreover, α_j and α_{j+1} differ on exactly 1 non-repeating coordinate c_j. Then

$$
\begin{aligned}
&\|T_x(\boldsymbol{x}, \boldsymbol{z}_1) - T_x(\boldsymbol{x}, \boldsymbol{z}_2)\|_2 && (26)\\
=&\|\sum_{j=1}^{n-1} T_x(\boldsymbol{x}, \alpha_j) - T_x(x, \alpha_{j+1})\|_2 && (27)\\
\leq&\sum_{j=1}^{n-1} \|T_x(\boldsymbol{x}, \alpha_j) - T_x(\boldsymbol{x}, \alpha_{j+1})\|_2 && (28)\\
\leq&\sum_{j=1}^{n-1} \max_{\boldsymbol{k} \in \Delta} \|T_x(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{k})) - T_x(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{k}) + w_{c_j})\|_2, && (29)\\
\leq&\sum_{i=1}^{m} \max_{\boldsymbol{k} \in \Delta} \|T_x(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{k})) - T_x(\boldsymbol{x}, \boldsymbol{w}(\boldsymbol{k}) + w_i)\|_2. && (30)
\end{aligned}
$$

which implies Eq. (20). Eq. (21) also holds, following exactly the same argument for point clouds. □

B.3 Proof of Theorem 4: General IoU Certification for 3D Bounding Boxes

We first recall Theorem 4 from the main paper. Note that we omit details for the convex hulls $\underline{S}, \bar{S}$ in the main paper version for simplicity. Here we provide a complete version with a formal description for $\underline{S}, \bar{S}$.

Theorem 4 (restated)**.** *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 $B_i \in \mathbf{B}$,*

$$
\mathrm{IoU}(B_i, B_{gt}) \geq \frac{h_1 \cdot (\underline{l}\underline{w} - \mathrm{Vol}(\underline{S} \backslash S_{gt}))}{hwl + \bar{h}\bar{w}\bar{l} - h_2 \cdot \big(\bar{l}\bar{w} - \mathrm{Vol}(\bar{S} \backslash S_{gt})\big)} \tag{31}
$$

where $S_{gt} = (x, z, w, l, r)_{gt}$ is the projection of B_{gt} to the $x - z$ plane.

$$
\begin{aligned}
h_1 &= \max\big(\min_{y' \in [\underline{y}, \bar{y}]} \min\{h, \underline{h}, \frac{h + \underline{h}}{2} - |y' - y|\}, 0\big),\\
h_2 &= \max\big(\min_{y' \in [\underline{y}, \bar{y}]} \min\{h, \bar{h}, \frac{h + \bar{h}}{2} - |y' - y|\}, 0\big).
\end{aligned} \tag{32}
$$

$\underline{S}, \bar{S}$ 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 $(\bar{w}, \bar{l})$. Here we formally define $C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, w, l)$. Let $\varphi = \arctan(\frac{l}{w})$, we first define

$$
\Delta x_{\max,k} = \max_{\theta \in [\underline{r}, \bar{r}]} \sqrt{w^2 + l^2} \cos(\theta + k\varphi), \tag{33}
$$

$$
\Delta x_{\min,k} = \min_{\theta \in [\underline{r}, \bar{r}]} \sqrt{w^2 + l^2} \cos(\theta + k\varphi), \tag{34}
$$

$$
\Delta z_{\max,k} = \max_{\theta \in [\underline{r}, \bar{r}]} \sqrt{w^2 + l^2} \sin(\theta + k\varphi), \tag{35}
$$

$$
\Delta z_{\min,k} = \min_{\theta \in [\underline{r}, \bar{r}]} \sqrt{w^2 + l^2} \sin(\theta + k\varphi). \tag{36}
$$

where $k \in \{-1, 1\}$. The range for each of the four coordinates can be expressed as

$$
\begin{aligned}
P_{1,1} = \{&\underline{x} + \frac{\Delta x_{\min,1}}{2}, \bar{x} + \frac{\Delta x_{\max,1}}{2}\} && (37)\\
&\otimes \{\underline{z} + \frac{\Delta z_{\min,1}}{2}, \bar{z} + \frac{\Delta z_{\max,1}}{2}\}, && (38)\\
P_{1,-1} = \{&\underline{x} + \frac{\Delta x_{\min,-1}}{2}, \bar{x} + \frac{\Delta x_{\max,-1}}{2}\} && (39)\\
&\otimes \{\underline{z} + \frac{\Delta z_{\min,-1}}{2}, \bar{z} + \frac{\Delta z_{\max,-1}}{2}\}, && (40)\\
P_{-1,1} = \{&\underline{x} - \frac{\Delta x_{\max,-1}}{2}, \bar{x} - \frac{\Delta x_{\min,-1}}{2}\} && (41)\\
&\otimes \{\underline{z} - \frac{\Delta z_{\max,-1}}{2}, \bar{z} - \frac{\Delta z_{\min,-1}}{2}\}, && (42)\\
P_{-1,-1} = \{&\underline{x} - \frac{\Delta x_{\max,1}}{2}, \bar{x} - \frac{\Delta x_{\min,1}}{2}\} && (43)\\
&\otimes \{\underline{z} - \frac{\Delta z_{\max,1}}{2}, \bar{z} - \frac{\Delta z_{\min,1}}{2}\}. && (44)
\end{aligned}
$$

The convex hull $C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, w, l)$ is

$$C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, w, l) = \mathrm{Conv}\left(P_{1,1} \cup P_{1,-1} \cup P_{-1,1} \cup P_{-1,-1}\right). \tag{45}$$

The convex hull $\underline{S} = C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, \underline{w}, \underline{l})$, and $\bar{S} = C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, \bar{w}, \bar{l})$.

Proof. Let $B_i \in \mathbf{B}$ be a bounding box whose coordinates are lower bounded by $(\underline{x}, \underline{y}, \underline{z}, \underline{w}, \underline{h}, \underline{l}, \underline{r})$ and upper bounded 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.

$$\mathrm{IoU}(B_i, B_{gt}) = \frac{\mathrm{Vol}(B_i \cap B_{gt})}{\mathrm{Vol}(B_i \cup B_{gt})} \tag{46}$$

Given a fixed center $(x, y, z)_i$ and a rotation angle r_i for B_i, the volumes $\mathrm{Vol}(B_i \cap B_{gt})$ and $\mathrm{Vol}(B_i \cup B_{gt})$ are both monotonic in terms of the size of B_i, $(w, h, l)_i$. Hence

$$\mathrm{IoU}(B_i, B_{gt}) \geq \frac{\min_{x,y,z,r} \mathrm{Vol}(B_i(x, y, z, \underline{w}, \underline{h}, \underline{l}, r) \cap B_{gt})}{\max_{x,y,z,r} \mathrm{Vol}(B_i(x, y, z, \bar{w}, \bar{h}, \bar{l}, r) \cup B_{gt})} \tag{47}$$

Note that

$$
\begin{aligned}
&\mathrm{Vol}(B_i(x_i, y_i, z_i, \bar{w}, \bar{h}, \bar{l}, r_i) \cup B_{gt}) && (48)\\
=&\mathrm{Vol}(B_i(x_i, y_i, z_i, \bar{w}, \bar{h}, \bar{l}, r_i)) + \mathrm{Vol}(B_{gt})\\
&\quad - \mathrm{Vol}(B_i(x_i, y_i, z_i, \bar{w}, \bar{h}, \bar{l}, r_i) \cap B_{gt})\\
=&\bar{w}\bar{h}\bar{l} + whl - \mathrm{Vol}(B_i(x_i, y_i, z_i, \bar{w}, \bar{h}, \bar{l}, r_i) \cap B_{gt}). && (49)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\max_{x,y,z,r} \mathrm{Vol}\big(B_i(x, y, z, \bar{w}, \bar{h}, \bar{l}, r) \cup B_{gt}\big)\\
=&\bar{w}\bar{h}\bar{l} + whl - \min_{x,y,z,r} \mathrm{Vol}(B_i(x, y, z, \bar{w}, \bar{h}, \bar{l}, r) \cap B_{gt}).
\end{aligned}
\tag{50}
$$

Combine Eqs. (47) and (50), we are left with the work of estimating $\min_{x,y,z,r} \mathrm{Vol}$
$(B_i(x, y, z, w_i, h_i, l_i, r) \cap B_{gt})$ for some fixed $(w_i, h_i, l_i) = (\underline{w}, \underline{h}, \underline{l})$ or $(\bar{w}, \bar{h}, \bar{l})$. Notice that 3D bounding boxes can be arbitrarily rotated along the y-axis, we consider the intersection on the y-axis and on the x-z plane separately.

Intersection on the y-axis. Projecting B_i and B_{gt} to the y-axis, we want to lower bound the intersection between an interval I_1 with length h_i centered at $y_i \in [\underline{y}, \bar{y}]$ and the ground truth interval $I_2 = [y - \frac{h}{2}, y + \frac{h}{2}]$.

Suppose $h_i < h$. If $|y_i - y| < \frac{h - h_i}{2}$, $|I_1(y_i) \cap I_2| = h_i$; otherwise $|I_1(y_i) \cap I_2| = \max\{\frac{h + h_i}{2} - |y_i - y|, 0\}$. In this case we conclude that $|I_i(y_i) \cap I_2| = \max\{\min\{h_i, \frac{h+h_i}{2} - |y' - y|\}, 0\}$. By the exact same argument, when $h_i \geq h$, $|I_1(y_i) \cap I_2| = \max\{\min\{h, \frac{h+h_i}{2} - |y_i - y|\}, 0\}$. Thus,

$$|I_1(y_i) \cap I_2| \geq \max\{\min_{y_i \in [\underline{y}, \bar{y}]} \min\{h, h_i, \frac{h + h_i}{2} - |y_i - y|\}, 0\}. \tag{51}$$

In particular, when $h_i = \underline{h}$ and $h_i = \bar{h}$, the intersection between B_i and B_{gt} on y-axis is larger than h_1 and h_2, respectively, where

$$h_1 = \max\big(\min_{y' \in [\underline{y}, \bar{y}]} \min\{h, \underline{h}, \frac{h + \underline{h}}{2} - |y' - y|\}, 0\big),$$
$$h_2 = \max\big(\min_{y' \in [\underline{y}, \bar{y}]} \min\{h, \bar{h}, \frac{h + \bar{h}}{2} - |y' - y|\}, 0\big). \tag{52}$$

Note that both h_1 and h_2 can be precisely numerically computed, where the pseudocode is in Algorithm 3.

Intersection on the x-z plane. Next, we consider the projection of B_i and B_{gt} on the x-z plane, denoted by S_i and S_{gt}, respectively. We have

$$\min_{x,z,r} \mathrm{Vol}(S_i(x, z, w_i, l_i, r) \cap S_{gt}) \tag{53}$$
$$= \mathrm{Vol}(S_i(x, z, w_i, l_i, r)) - \min_{x,z,r} \mathrm{Vol}(S_i(x, z, w_i, l_i, r) \backslash S_{gt}) \tag{54}$$
$$= w_i l_i - \min_{x,z,r} \mathrm{Vol}(S_i(x, z, w_i, l_i, r) \backslash S_{gt}) \tag{55}$$
$$\geq w_i l_i - \mathrm{Vol}(C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, w_i, l_i) \backslash S_{gt}) \tag{56}$$

where $C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, w_i, l_i)$ is an envelop that contains all possible x-z bounding boxes S_i with $(\underline{x}, \underline{z}, \underline{r}) \leq (x, z, r) \leq (\bar{x}, \bar{z}, \bar{r})$ and a fixed size (w_i, l_i).

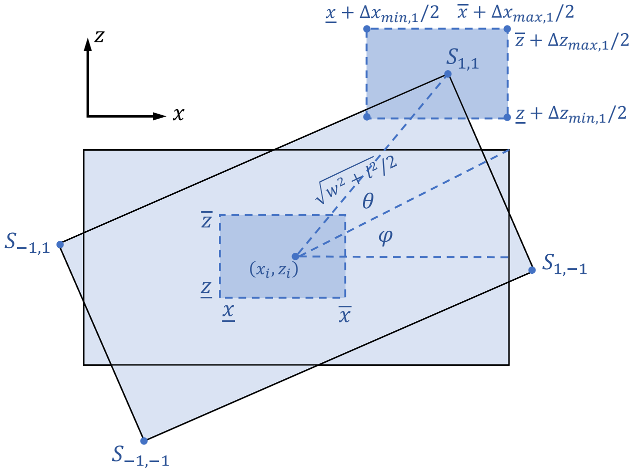


Figure 5: Illustration of $C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, w, l)$ on $x - z$ plane.

As shown in Fig. 5, we calculate the possible range for each of the four vertices of bounding box S_i. For example, Fig. 5 illustrates that the $x - z$ coordinate of the upper-right vertex is $s_{1,1} = (x_i + \frac{\sqrt{w^2+l^2}}{2}\cos(\theta + \varphi), z_i + \frac{\sqrt{w^2+l^2}}{2}\sin(\theta + \varphi))$. Therefore, its x, z coordinate of $s_{1,1}$satisfies

$$\underline{x} + \min_{\theta \in [\underline{r}, \bar{r}]} \frac{1}{2}\sqrt{w^2 + l^2}\cos(\theta + \varphi) \leq x_{1,1}$$
$$\leq \bar{x} + \max_{\theta \in [\underline{r}, \bar{r}]} \frac{1}{2}\sqrt{w^2 + l^2}\cos(\theta + \varphi) \tag{57}$$
$$\underline{z} + \min_{\theta \in [\underline{r}, \bar{r}]} \frac{1}{2}\sqrt{w^2 + l^2}\sin(\theta + \varphi) \leq z_{1,1}$$
$$\leq \bar{z} + \max_{\theta \in [\underline{r}, \bar{r}]} \frac{1}{2}\sqrt{w^2 + l^2}\sin(\theta + \varphi). \tag{58}$$

which means $s_{1,1} = (x_{1,1}, z_{1,1})$ is contained by the rectangle formed by the four points in $S_{1,1}$, i.e., $s_{1,1} \in \mathrm{Conv}(P_{1,1})$. Similar arguments also hold for rest of vertices: $s_{1,-1} \in \mathrm{Conv}(P_{1,-1})$, $s_{-1,1} \in \mathrm{Conv}(P_{-1,1})$, and $s_{-1,-1} \in \mathrm{Conv}(P_{-1,-1})$. Finally, we conclude that

$$\begin{aligned} S_i &= \mathrm{Conv}(s_{1,1}, s_{1,-1}, s_{-1,1}, s_{-1,-1}) \\ &\subseteq \mathrm{Conv}(P_{1,1}, P_{1,-1}, P_{-1,1}, P_{-1,-1}) \end{aligned} \tag{59}$$

Note that $\underline{S} = C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, \underline{w}, \underline{l})$, and $\bar{S} = C(\underline{x}, \underline{z}, \underline{r}, \bar{x}, \bar{z}, \bar{r}, \bar{w}, \bar{l})$. By Eqs. (56) and (59),

$$\min_{x,z,r} \mathrm{Vol}(S_i(x, z, \underline{w}, \underline{l}, r) \cap S_{gt}) \geq \underline{wl} - \mathrm{Vol}(\underline{S} \backslash S_{gt}) \tag{60}$$

and

$$\min_{x,z,r} \mathrm{Vol}(S_i(x, z, \bar{w}, \bar{l}, r) \cap S_{gt}) \geq \bar{w}\bar{l} - \mathrm{Vol}(\bar{S} \backslash S_{gt}). \tag{61}$$

Combining the above, we have

$$\mathrm{IoU}(B_i, B_{gt}) \geq \frac{h_1(\underline{wl} - \mathrm{Vol}(\underline{S} \backslash S_{gt}))}{hwl + \bar{h}\bar{w}\bar{l} - h_2(\bar{w}\bar{l} - \mathrm{Vol}(\bar{S} \backslash S_{gt}))}. \tag{62}$$

□

C Additional Details of Certification Strategies

In this section, we present the details of our detection (Appendix C.1) and IoU (Appendix C.2) certification algorithms for camera and LiDAR fusion models, including the pseudocode and some details of our implementation. Note that our algorithms can be adapted to any fusion framework and single-modality model by doing smoothing inference and certification for corresponding modules.

C.1 Detection Certification

Algorithm 1 presents the pseudocode of our detection certification (**function** CERTIFY) and median smoothing detection (**function** INFERENCE). In our implementation, we directly sample on the left endpoints of each interval and do smoothing inference, which can also be replaced by a random point in each small interval.

C.2 IoU Certification

Algorithm 2 presents the pseudocode of our IoU certification (**function** CERTIFY) and smoothing bounding box detection (**function** INFERENCE). As in the detection certification framework, we also do sampling on the lower endpoint of each small interval. Algorithm 3 presents the pseudocode of IoU lower bound computation given the bounding box parameter intervals and the ground truth bounding box, and Algorithm 4 presents the pseudocode for computing the x, z coordinate intervals for bounding box endpoints. Note that our framework focuses on the 7-parameter representation of bounding boxes (x, y, z, height, width, length, rotation angle), which can be adapted to other representation formats easily (e.g. 8 parameter representation, endpoint representation).

Algorithm 1: 0/1 Detection and Certification

```
Input: clean image input $\boldsymbol{x}_0 \in \mathcal{X} \subseteq \mathbb{R}^d$, clean point cloud input $\boldsymbol{p}_0 \in \mathcal{P} \subseteq \mathbb{R}^{3 \times N}$, multi-sensor fusion pipeline $F : \mathcal{X} \times \mathcal{P} \to ([C] \times [0, 1])_n$, ground truth label $y$, image Gaussian
       noise variance $\sigma_x^2$, point Gaussian noise variance $\sigma_p^2$, transformation space $[\boldsymbol{z}_l, \boldsymbol{z}_u]$, transformation function $T$, detection threshold $\gamma_0$
Output: smoothed confidence score of ground truth label $c_{median}$, lower bound and upper bound of confidence score of ground truth $c_u, c_l$, whether the object is robustly detected as a boolean
        variable $b$
function INFERENCE ($F, \boldsymbol{x}_0, \boldsymbol{p}_0, y_k, \sigma_x, \sigma_p, n, \gamma_0$):
    $\hat{x} \leftarrow$ ADDGAUSSIANNOISE ($\boldsymbol{x}_0, \sigma_x, n$)
    $\hat{p} \leftarrow$ ADDGAUSSIANNOISE ($\boldsymbol{p}_0, \sigma_p, n$)
    $\hat{c} \leftarrow$ COMPUTECONFIDENCESCORES ($F, \hat{x}, \hat{p}, y_k$)
    $\hat{c} \leftarrow$ SORT ($\hat{c}$)
    $c_{median} \leftarrow \hat{c}_{\lfloor 0.5n \rfloor}$
    $b \leftarrow \mathbb{I}[c_{median} \geq \gamma_0]$
    return $c_{median}, b$
function CERTIFY ($F, \boldsymbol{x}_0, \boldsymbol{p}_0, y_k, \sigma_x, \sigma_p, n, [\boldsymbol{z}_l, \boldsymbol{z}_u], c, \alpha, \gamma_0$):
    $interval\_list \leftarrow$ SPLITINTERVAL ($[\boldsymbol{z}_l, \boldsymbol{z}_u]$)
    $c_{median}\_list, c_u\_list, c_l\_list \leftarrow$ [], [], []
    for $[z_{sl}, z_{su}] \in$ interval_list do
        $\epsilon \leftarrow$ COMPUTEEPS ($T, \boldsymbol{x}_0, \boldsymbol{p}_0, \sigma_x, \sigma_p, \boldsymbol{z}_{sl}, \boldsymbol{z}_{su}$)                    ▷ compute $\epsilon$ according to equation (13)

        $q_u, q_l \leftarrow$ GETEMPIRICALPERC ($n, \epsilon, c, \alpha/$|interval_list|)                    ▷ compute $q_u, q_l$ according to equation (10)

        $\hat{x} \leftarrow$ ADDGAUSSIANNOISE ($T_x(\boldsymbol{x}_0, \boldsymbol{z}_{sl}), \sigma_x, n$)
                    ▷ sample $n$ Gaussian noises $\delta_x \sim \mathcal{N}(0, \sigma_x^2 \boldsymbol{I}_d)$ and add to $T_x(\boldsymbol{x}_0, \boldsymbol{z}_{sl})$ to get $n$ noisy samples $\hat{x}$

        $\hat{p} \leftarrow$ ADDGAUSSIANNOISE ($T_p(\boldsymbol{p}_0, \boldsymbol{z}_{sl}), \sigma_p, n$)
                    ▷ sample $n$ Gaussian noises $\delta_p \sim \mathcal{N}(0, \sigma_p^2 \boldsymbol{I}_{3 \times N})$ and add to $(T_p(\boldsymbol{p}_0, \boldsymbol{z}_{sl})$ to get $n$ noisy samples $\hat{p}$

        $\hat{c} \leftarrow$ COMPUTECONFIDENCESCORES ($F, \hat{x}, \hat{p}, y$)
                    ▷ collect the confidence score of $y$ for each $(\boldsymbol{x}, \boldsymbol{p}) \in (\hat{x}, \hat{p})$ based on $F(\boldsymbol{x}, \boldsymbol{p}) = \max_{1 \leq k \leq \ell : y_k = y} c_k$

        $\hat{c} \leftarrow$ SORT ($\hat{c}$)
        $c_{median} \leftarrow \hat{c}_{\lfloor 0.5n \rfloor}$
        if $q_l = -1$ then
            $c_l \leftarrow -\infty$                    ▷ $-\infty$ means cannot certify
        else
            $c_l \leftarrow \hat{c}_{q_l}$
        if $q_u = \infty$ then
            $c_u \leftarrow \infty$                    ▷ $\infty$ means cannot certify
        else
            $c_u \leftarrow \hat{c}_{q_u}$
        $c_{median}\_list$ add $c_{median}$, $c_u\_list$ add $c_u$, $c_l\_list$ add $c_l$
    $b \leftarrow \mathbb{I}[\min c_l\_list \geq \gamma_0]$
    return $\min c_l\_list, \max c_u\_list, b$
function GETEMPIRICALPERC ($n, \epsilon, c, \alpha$):
    $\underline{p} \leftarrow \Phi\left(\Phi^{-1}(p) - \epsilon\right)$
    $q_l \leftarrow$ BINARYSEARCHLEFT($BinomialCDF(n, \underline{p}), 1 - \alpha$)                    ▷ do binary search and choose the left endpoint

    if BinomialCDF($q_l, \underline{p}$) $> 1 - \alpha$ then
        $q_l \leftarrow -1$
    $\overline{p} \leftarrow \Phi\left(\Phi^{-1}(p) + \epsilon\right)$
    $q_u \leftarrow$ BINARYSEARCHRIGHT($BinomialCDF(n, \overline{p}), \alpha$)                    ▷ do binary search and choose the right endpoint

    if BinomialCDF($q_u, \overline{p}$) $< \alpha$ then
        $q_u \leftarrow \infty$
    return $q_u, q_l$
```

Algorithm 2: Bounding Box Detection and Certification

```
Input: clean image input $\boldsymbol{x}_0 \in \mathcal{X} \subseteq \mathbb{R}^d$, clean point cloud input $\boldsymbol{p}_0 \in \mathcal{P} \subseteq \mathbb{R}^{3\times N}$, multi-sensor fusion pipeline
       $F : \mathcal{X} \times \mathcal{P} \to ([X] \times [Y] \times [Z] \times [W] \times [H] \times [L] \times [R] \times [C] \times [0, 1])_n$, image Gaussian noise $\delta_x \sim \mathcal{N}(0, \sigma_x^2 \boldsymbol{I}_d)$, point Gaussian noise $\delta_p \sim \mathcal{N}(0, \sigma_p^2 \boldsymbol{I}_{3\times N})$,
       transformation space $[\boldsymbol{z}_l, \boldsymbol{z}_u]$.
Output: smoothed prediction $bbox_{median}$, lower bound of IoU between predicted bounding boxes and ground truth bounding boxes $\underline{IoU}$
function Inference ($F, \boldsymbol{x}_0, \boldsymbol{p}_0, \sigma_x, \sigma_p, n$):
    $\hat{x} \leftarrow$ AddGaussianNoise ($\boldsymbol{x}_0, \sigma_x, n$)
    $\hat{p} \leftarrow$ AddGaussianNoise ($\boldsymbol{p}_0, \sigma_p, n$)
    $\hat{bbox} \leftarrow$ ComputeBBoxParams ($F(\hat{x}, \hat{p})$)
    $\hat{bbox} \leftarrow$ Sort ($\hat{bbox}$)                                    ▷ sort on each parameter

    $bbox_{median} \leftarrow \hat{bbox}_{\lfloor 0.5n \rfloor}$                          ▷ take the median of each parameter

    return $bbox_{median}$
function Certify ($F, \boldsymbol{x}_0, \boldsymbol{p}_0, \sigma_x, \sigma_p, n, [\boldsymbol{z}_{sl}, \boldsymbol{z}_{su}], c, \alpha$):
    $interval\_list \leftarrow$ SplitInterval ($[\boldsymbol{z}_{sl}, \boldsymbol{z}_{su}]$)
    $IoU\_list \leftarrow []$
    for $[z_{sl}, z_{su}] \in$ interval_list do
        $\epsilon \leftarrow$ ComputeEps ($T, \boldsymbol{x}_0, \boldsymbol{p}_0, \sigma_x, \sigma_p, \boldsymbol{z}_l, \boldsymbol{z}_u$)     ▷ compute $\epsilon$ according to equation (13)

        $q_u, q_l \leftarrow$ GetEmpiricalPerc ($n, \epsilon, c, \alpha/|\text{interval\_list}|$)    ▷ compute $q_u, q_l$ according to equation (10)

        $\hat{x} \leftarrow$ AddGaussianNoise ($T_x(\boldsymbol{x}_0, \boldsymbol{z}_{sl}), \sigma_x, n$)
            ▷ sample $n$ Gaussian noises $\delta_x \sim \mathcal{N}(0, \sigma_x^2 \boldsymbol{I}_d)$ and add to $T_x(\boldsymbol{x}_0, \boldsymbol{z}_{sl})$ to get $n$ noisy samples $\hat{x}$

        $\hat{p} \leftarrow$ AddGaussianNoise ($T_p(\boldsymbol{p}_0, \boldsymbol{z}_{sl}), \sigma_p, n$)
            ▷ sample $n$ Gaussian noises $\delta_p \sim \mathcal{N}(0, \sigma_p^2 \boldsymbol{I}_{3\times N})$ and add to $(T_p(\boldsymbol{p}_0, \boldsymbol{z}_{sl})$ to get $n$ noisy samples $\hat{p}$

        $\hat{bbox} \leftarrow$ ComputeBBoxParams ($F(\hat{x}, \hat{p})$)
        $\hat{bbox} \leftarrow$ Sort ($\hat{bbox}$)                                ▷ sort on each parameter

        if $q_l = -1$ then
            $\underline{IoU} \leftarrow -\infty$                                   ▷ $-\infty$ means cannot certify

            return $\underline{IoU}$
        else
            $bbox_l \leftarrow \hat{bbox}_{q_l}$
        if $q_u = \infty$ then
            $\underline{IoU} \leftarrow \infty$                                    ▷ $\infty$ means cannot certify

            return $\underline{IoU}$
        else
            $bbox_u \leftarrow \hat{bbox}_{q_u}$
        $\underline{IoU} \leftarrow$ IoULowerBound ($\underline{bbox}, \overline{bbox}, bbox$)
        $IoU\_list$ add $\underline{IoU}$
    return $\min(IoU\_list)$
function GetEmpiricalPerc ($n, \epsilon, c, \alpha$):
    $\underline{p} \leftarrow \Phi\left(\Phi^{-1}(p) - \epsilon\right)$
    $q_l \leftarrow$ BinarySearchLeft($BinomialCDF(n, \underline{p}), 1 - \alpha$)        ▷ do binary search and choose the left endpoint

    if BinomialCDF($q_l, \underline{p}$) $> 1 - \alpha$ then
        $q_l \leftarrow -1$
    $\overline{p} \leftarrow \Phi\left(\Phi^{-1}(p) + \epsilon\right)$
    $q_u \leftarrow$ BinarySearchRight($BinomialCDF(n, \overline{p}), \alpha$)          ▷ do binary search and choose the right endpoint

    if BinomialCDF($q_u, \overline{p}$) $< \alpha$ then
        $q_u \leftarrow \infty$
    return $q_u, q_l$
```

Algorithm 3: IoU Lower Bound

```
Input: upper bound of bounding boxes' parameters $\overline{bbox}$, lower bound of bounding boxes'
       parameters $\underline{bbox}$, ground truth bounding boxes' parameters $bbox$
Output: lower bound of IoU between predicted bounding boxes and ground truth bounding boxes
        $\underline{IoU}$
function IoULowerBound($\overline{bbox}$, $\underline{bbox}$, $bbox$):
    $\underline{V_I} \leftarrow$ IntersectionLowerBound($\overline{bbox}$, $\underline{bbox}$, $bbox$)
    $\overline{V_U} \leftarrow$ UnionUpperBound($\overline{bbox}$, $\underline{bbox}$, $bbox$)
    return $\underline{V_I}/\overline{V_U}$

function IntersectionLowerBound($\overline{bbox}$, $\underline{bbox}$, $bbox$):
    if $y < (\overline{y} + \underline{y})/2$ then
        $y_l = \overline{y}$
    else
        $y_l = \underline{y}$

    if $y \leq y_l - \underline{h}$ or $y_l \leq y - h$ then
        return 0
    else
        $h_I = \min(y_l, y) - \max(y_l - \underline{h}, y - h)$

    $xz_l, xz_u \leftarrow$ CornerXZIntervals($\underline{l}, \underline{w}, \underline{x}, \underline{z}, \underline{r}, \overline{x}, \overline{z}, \overline{r}$)
    $\underline{S_{overlap}} \leftarrow$ OverlapAreaLower($xz_l, xz_u, bbox$)
    if $\underline{S_{overlap}} \leq 0$ then
        return 0
    else
        return $h_I \cdot \underline{S_{overlap}}$

function UnionUpperBound($\overline{bbox}$, $\underline{bbox}$, $bbox$):
    if $y < (\overline{y} + \underline{y})/2$ then
        $y_u = \underline{y}$
    else
        $y_u = \overline{y}$

    if $y \leq y_u - \overline{h}$ or $y_u \leq y - h$ then
        return 0
    else
        $h_I = \min(y_u, y) - \max(y_u - \overline{h}, y - h)$

    $xz_l, xz_u \leftarrow$ CornerXZIntervals($\overline{l}, \overline{w}, \underline{x}, \underline{z}, \underline{r}, \overline{x}, \overline{z}, \overline{r}$)
    $\underline{S_{overlap}} \leftarrow$ OverlapAreaLower($xz_l, xz_u, bbox$)
    if $\underline{S_{overlap}} \leq 0$ then
        return $h \cdot w \cdot l + \overline{h} \cdot \overline{w} \cdot \overline{l}$
    else
        return $h \cdot w \cdot l + \overline{h} \cdot \overline{w} \cdot \overline{l} - h_I \cdot \underline{S_{overlap}}$
```

D Additional Experimental Details

D.1 Dataset

As introduced in Section 4, we generate the certification data via spawning the ego vehicle at a few randomly chosen spawn points.

We use CARLA simulator to generate the benchmark, which is widely used in the literature to study the robustness of autonomous driving models (e.g.,(Xu et al. 2022)). It is studied that model performance in CARLA-generated scenarios aligns with that in the real world (Osinski et al. 2020). Hence, we believe the fidelity of CARLA-generated data is suitable for studying and improving the robustness of MSF systems.

For settings with 15 spawn points in Table 2, the randomly-chosen spawn point index in the CARLA Town01 map is 15, 30, 43, 46, 57, 86, 102, 11, 136, 14, 29, 6, 61, 81, and 88. For settings with 4 spawn points in Table 2, the randomly-chosen spawn point index is 15, 30, 43, and 46.

Certification setup. As reflected in Lemma 2, to certify the robustness, we need to partition the transformation's parameter space. For rotation certification, we split the rotation angle interval $[-30°, 30°]$ uniformly into 600 tiny intervals of 0.1 degree; for shifting certification, we split the distance interval $[10, 15]$ uniformly into 500 tiny intervals of 0.01 meter.

Empirical attack setup. In Section 4, we evaluate the empirical robustness, i.e., the robustness under attacks, to show vanilla models' vulnerability and estimate the certification tightness. For rotation, we conduct the attack by enumerating the lowest detection rate and IoU score among 6000 parameters uniformly sampled with a distance 0.01 degree (distance = $60°/6000 = 0.01°$); for shifting, we conduct the attack by enumerating the lowest detection rate and IoU score among 5000 parameters uniformly sampled with a distance of 0.001 meter (distance = $5/5000 = 0.001$).

D.2 Detailed Experimental Evaluation

In this section, we present the complete experimental results, which include rotation transformation (Table 3) and shifting transformation (Table 4) considering different thresholds for detection and IoU certification.

Certification against rotation transformation. As shown in Table 3 and discussed in Section 4.1, the order of robustness against rotation transformation in the detection metric is FocalsConv $>$ MVX-Net $>$ MonoCon $>$ CLOCs $>$ SECOND, but the most robust model in the IoU metric is CLOCs. Moreover, with experimental results in different thresholds (Table 3b, Figure 6a and Figure 3), the performance of models all drop when the threshold or attack radius increase, and all models' certified IoU drop to 0 when $\text{TH}_{\text{IoU}} \approx 0.6$, which reflects the problem of models' robustness against rotation transformation. We can also find that the performance difference between models increases when the threshold or attack radius increases.

Certification against shifting transformation. As shown in Table 4 and discussed in Section 4.2, the order of robustness against shifting transformation in the detection metric is MVX-Net $>$ CLOCs $>$ SECOND $\approx$ MonoCon $>$ FocalsConv, and the order in the IoU metric is CLOCs $>$ MonoCon $<$ MVX-Net $>$ SECOND $>$ FocalsConv. This shows the advantage of fusion models (e.g. CLOCs) on the one hand but also shows this makes the attack space larger on another hand (e.g. FocalsConv), which leads to a new question of robust fusion mechanism design.

Algorithm 4: Corners' x, z Intervals

```
Input: bounding box length $l$, width $w$, $x$ lower bound $\underline{x}$, $z$ lower bound $\underline{z}$, rotation angel lower bound $\underline{r}$, $x$ upper bound $\overline{x}$, $z$ upper bound $\overline{z}$, rotation upper bound $\overline{r}$
Output: bounding box corners' $xz$ coordinates lower bound $xz_l$, bounding box corners' xz coordinates upper bound $xz_u$
function CornerXZIntervals ($l, w, \underline{x}, \underline{z}, \underline{r}, \overline{x}, \overline{z}, \overline{r}$):
    $xzCoorsList \leftarrow []$
    for $x$, $z$ in $[[\underline{x}, \underline{z}]$ , $[\overline{x}, \overline{z}]]$ do
        for $r$ in $\underline{r}, \overline{r}$ do
            $xzCoors \leftarrow$ ComputeXZ $(x, z, r)$
            $xzCoorsList$ add $xzCoors$
    $xz_l, xz_u \leftarrow$ ComputeXZInterval $(xzCoorsList)$            ▷ compute $x, z$ intervals roughly

    ▷ consider extremum cases
    $\alpha \leftarrow \arctan(w/l)$
    $d \leftarrow \sqrt{(l/2)^2 + (w/2)^2}$
    if $\pi - \alpha \geq \underline{r}$ and $\pi - \alpha \leq \overline{r}$ then
        $xz_l[0] \leftarrow \underline{x} - d$
        $xz_u[4] \leftarrow \overline{x} + d$
    if $2\pi - \alpha \geq \underline{r}$ and $2\pi - \alpha \leq \overline{r}$ then
        $xz_u[0] \leftarrow \overline{x} + d$
        $xz_l[4] \leftarrow \underline{x} - d$
    if $\pi/2 - \alpha \geq \underline{r}$ and $\pi/2 - \alpha \leq \overline{r}$ then
        $xz_u[1] \leftarrow \overline{z} + d$
        $xz_l[5] \leftarrow \underline{z} - d$
    if $3\pi/2 - \alpha \geq \underline{r}$ and $3\pi/2 - \alpha \leq \overline{r}$ then
        $xz_l[1] \leftarrow \underline{z} - d$
        $xz_u[5] \leftarrow \overline{z} + d$
    if $\alpha \geq \underline{r}$ and $\alpha \leq \overline{r}$ then
        $xz_u[2] \leftarrow \overline{x} + d$
        $xz_l[6] \leftarrow \underline{x} - d$
    if $\pi + \alpha \geq \underline{r}$ and $\pi + \alpha \leq \overline{r}$ then
        $xz_l[2] \leftarrow \underline{x} - d$
        $xz_u[6] \leftarrow \overline{x} + d$
    if $\pi/2 + \alpha \geq \underline{r}$ and $\pi/2 + \alpha \leq \overline{r}$ then
        $xz_u[3] \leftarrow \overline{z} + d$
        $xz_l[7] \leftarrow \underline{z} - d$
    if $3\pi/2 + \alpha \geq \underline{r}$ and $3\pi/2 + \alpha \leq \overline{r}$ then
        $xz_l[3] \leftarrow \underline{z} - d$
        $xz_u[7] \leftarrow \overline{z} + d$
return $xz_l, xz_u$
```

Table 2: Certification data. Each row stands for one setting, and the columns “vehicle color”, “building”, “pedestrian”, and “amount” represent the color of the car, whether buildings exist, whether a pedestrian exists, and the number of corresponding data respectively.

vehicle color	building	pedestrian	amount
blue	yes	no	15
red	yes	no	4
red	yes	yes	4
black	yes	no	4
black	yes	yes	4
blue	no	no	15
red	no	no	4
red	no	yes	4
black	no	no	4
black	no	yes	4

D.3 Effect of Sample Strategy

To study the effect of sampling strategies, we compare two different sampling strategies and show the results in Table 5. The first strategy is fixing the size of intervals (e.g. 0.1° rotation intervals in out case), which is named “Certification (sparse)” in Table 5, and the second strategy is fixing small interval number (e.g. 600 small intervals in each big rotation interval), which is called “Certification (dense)” in Table 5.

From certified detection rate Table 5a and certified IoU Table 5b, we can find that the “Certification (sparse)” is already tight enough when the sample number in each small interval stays the same since the certified detection rate and IoU in “Certification (sparse)” is almost the same as those in “Certification (dense)”.

D.4 Effect of Smoothing Parameter σ

To study the effect of smoothing parameter σ, we test our approach with random Gaussian noise whose $\sigma = 0.5$ in our rotation setting to compare the previous results on rotation transformation with noises whose $\sigma = 0.25$ (Table 3).

As shown in Figure 6c, Figure 8 and Table 6, the models’ performance might degrade in small attack radius with larger smoothing σ, but smoothing with larger σ can also improve the robustness of models against large attack radius, especially for the model which is more stable than others (CLOCs in our case).

D.5 Examples

In this subsection, we present some failure cases and possible reasons.

As shown in Figure 9a, where MonoCon fails to detect the front vehicle when the rotation angle r is somewhere larger than 20° but is able to detect it when $r \leq 20^\circ$, where CLOCs can detect cars with all rotation angles between -30° and 30°. The possible reason for this situation is that there are some objects (e.g. the puddle on the sidewalk and trees far away in this case) with similar color to the vehicle, which impacts the detection ability of camera-based detection modules. This problem can be mitigated by the LiDAR-based detection modules.

Figure 9c shows another failure case of camera-based models. Although there is no object with a similar color to the vehicle that we want to detect. The car farther is relatively small in the image, which is harder for camera-based models to detect. The application of point cloud data is very helpful in this case because of the perception ability of objects at long distances.

Figure 9b and Figure 9d shows two failure cases of SECOND, which is representative of LiDAR-based models where fusion models can detect relatively well. The possible reason for this case is that there are some objects very close to the vehicle (e.g.benches and grass in these cases), which affects the detection ability of point cloud modules, which can be mitigated by the combination with camera-based modules.

E Side Discussions

E.1 Potential Limitations

- The first potential limitation of our framework is inference overhead incured by inference-time smoothing. Such smoothing usually requires sampling around 100 samples to make the prediction. This limitation is also shared by other randomized smoothing approaches and its mitigation is another important research topic (Horváth et al. 2022).
- The second potential limitation is achieving robustness comes at the expense of normal accuracy degradation. Such a robustness-accuracy tradeoff is a lasting topic and can be partly mitigated by principled training (Zhang et al. 2019) or gated selection (Mueller, Balunovic, and Vechev 2020).

Table 3: Overview of rotation transformation experiment results (smoothing $\sigma = 0.25$). Each row represents the corresponding model and attack radius. “Benign”, “Adv (Vanilla)”, “Adv (Smoothed)”, and “Certification” stands for benign performance, vanilla models’ performance under attacks, smoothed models’ performance under attacks, and certified lower bound of smoothed model performance under attacks. Each column represents the results under different thresholds.

(a) Detection rate under rotation transformation

Model	Attack Radius	Benign			Adv (Vanilla)			Adv (Smoothed)			Certification		
		Det@20	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**
MonoCon (Liu, Xue, and Wu 2022)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	70.97%	70.97%	58.06%	98.39%	98.39%	80.65%	95.16%	95.16%	75.81%
	$\lvert r\rvert \leq 15^\circ$				70.97%	70.97%	58.06%	98.39%	98.39%	80.65%	95.16%	95.16%	75.81%
	$\lvert r\rvert \leq 20^\circ$				70.97%	70.97%	58.06%	98.39%	98.39%	80.65%	95.16%	95.16%	75.81%
	$\lvert r\rvert \leq 25^\circ$				70.97%	70.97%	45.16%	98.39%	98.39%	80.65%	95.16%	95.16%	75.81%
	$\lvert r\rvert \leq 30^\circ$				70.97%	70.97%	32.26%	96.77%	96.77%	80.65%	91.94%	91.94%	75.81%
SECOND (Yan, Mao, and Li 2018)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				100.00%	100.00%	0.00%	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				100.00%	100.00%	0.00%	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				100.00%	12.90%	0.00%	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				67.74%	3.23%	0.00%	100.00%	100.00%	0.00%	100.00%	62.90%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	88.71%	100.00%	100.00%	88.71%
	$\lvert r\rvert \leq 15^\circ$				100.00%	100.00%	100.00%	98.39%	79.03%	66.13%	98.39%	77.42%	66.13%
	$\lvert r\rvert \leq 20^\circ$				100.00%	100.00%	100.00%	98.39%	69.35%	50.00%	98.39%	67.74%	50.00%
	$\lvert r\rvert \leq 25^\circ$				100.00%	91.94%	20.97%	98.39%	69.35%	50.00%	98.39%	67.74%	50.00%
	$\lvert r\rvert \leq 30^\circ$				100.00%	74.19%	3.23%	98.39%	69.35%	50.00%	98.39%	67.74%	50.00%
FocalsConv (Chen et al. 2022)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 15^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 20^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 25^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 30^\circ$				100.00%	100.00%	98.39%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
MVX-Net (Sindagi, Zhou, and Tuzel 2019)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	51.61%
	$\lvert r\rvert \leq 15^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	51.61%
	$\lvert r\rvert \leq 20^\circ$				100.00%	100.00%	90.32%	100.00%	100.00%	90.32%	100.00%	100.00%	51.61%
	$\lvert r\rvert \leq 25^\circ$				100.00%	100.00%	3.23%	100.00%	100.00%	3.23%	100.00%	100.00%	50.00%
	$\lvert r\rvert \leq 30^\circ$				100.00%	100.00%	3.23%	100.00%	100.00%	3.23%	100.00%	100.00%	50.00%

(b) IoU with ground truth under rotation transformation

Model	Attack Radius	Benign			Adv (Vanilla)			Adv (Smoothed)			Certification		
		AP@30	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**
MonoCon (Liu, Xue, and Wu 2022)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	56.45%	56.45%	0.00%	82.26%	82.26%	0.00%	75.81%	0.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				54.84%	54.84%	0.00%	82.26%	82.26%	0.00%	74.19%	0.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				54.84%	53.23%	0.00%	82.26%	74.19%	0.00%	6.45%	0.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				51.61%	16.13%	0.00%	80.65%	16.13%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				46.77%	0.00%	0.00%	79.03%	3.23%	0.00%	0.00%	0.00%	0.00%
SECOND (Yan, Mao, and Li 2018)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	96.77%	96.77%	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				96.77%	96.77%	0.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				96.77%	96.77%	0.00%	100.00%	100.00%	54.84%	100.00%	100.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				83.87%	83.87%	0.00%	100.00%	96.77%	0.00%	100.00%	0.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				51.61%	51.61%	0.00%	54.84%	54.84%	0.00%	11.29%	0.00%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	90.32%	90.32%	90.32%	100.00%	100.00%	98.39%	100.00%	100.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				90.32%	90.32%	90.32%	98.39%	98.39%	85.48%	98.39%	87.10%	0.00%
	$\lvert r\rvert \leq 20^\circ$				88.71%	88.71%	77.42%	98.39%	98.39%	67.74%	98.39%	69.35%	0.00%
	$\lvert r\rvert \leq 25^\circ$				87.10%	87.10%	0.00%	98.39%	98.39%	67.74%	98.39%	67.74%	0.00%
	$\lvert r\rvert \leq 30^\circ$				80.65%	80.65%	0.00%	98.39%	98.39%	67.74%	98.39%	53.23%	0.00%
FocalsConv (Chen et al. 2022)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	100.00%	0.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				96.77%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				96.77%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				96.77%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				96.77%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
MVX-Net (Sindagi, Zhou, and Tuzel 2019)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	96.77%	96.77%	0.00%	95.16%	72.58%	0.00%	72.58%	0.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				96.77%	96.77%	0.00%	95.16%	72.58%	0.00%	72.58%	0.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				96.77%	96.77%	0.00%	95.16%	72.58%	0.00%	72.58%	0.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				96.77%	75.81%	0.00%	95.16%	72.58%	0.00%	66.13%	0.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				96.77%	75.81%	0.00%	95.16%	72.58%	0.00%	0.00%	0.00%	0.00%

- The third potential limitation is evaluation without real-world AV systems. On the one hand, we remark that such evaluation will require a much longer evaluation time period, and related works (Hu et al. 2023; Prakash, Chitta, and Geiger 2021) usually skip it for faster technique evolvement. On the other hand, there is no technical challenge in deploying and evaluating our approach in real-world AV systems, and it is our ongoing effort to conduct evaluations on real-world AV systems.

E.2 Connection with Other Trustworthy Research Directions

- **Conformal Prediction** The conformal prediction is another important tool for qualifying and improving the trustworthiness in autonomous driving. Since our framework certifies **instance-level** prediction correctness for semantically transformed input, and conformal prediction guarantees **distribu-tional-level** error rate for exchangeable data distribution, our framework cannot directly compare with conformal prediction. However, our framework can be combined with conformal prediction to improve trustworthiness under these two notions at the same time. We leave as future work to study the interactions and connections between our framework and conformal prediction.
- **Corruption Parameterization** Though our method does not support common corruption yet, we believe that this is not a technical limitation. Indeed, as long as the transformation can be parameterized, we can apply our method to certify. As a result, to apply our proposed method to common corruptions, we only need to parameterize common corruptions, e.g., by learning the perturbations sets (Wong and Kolter 2021).

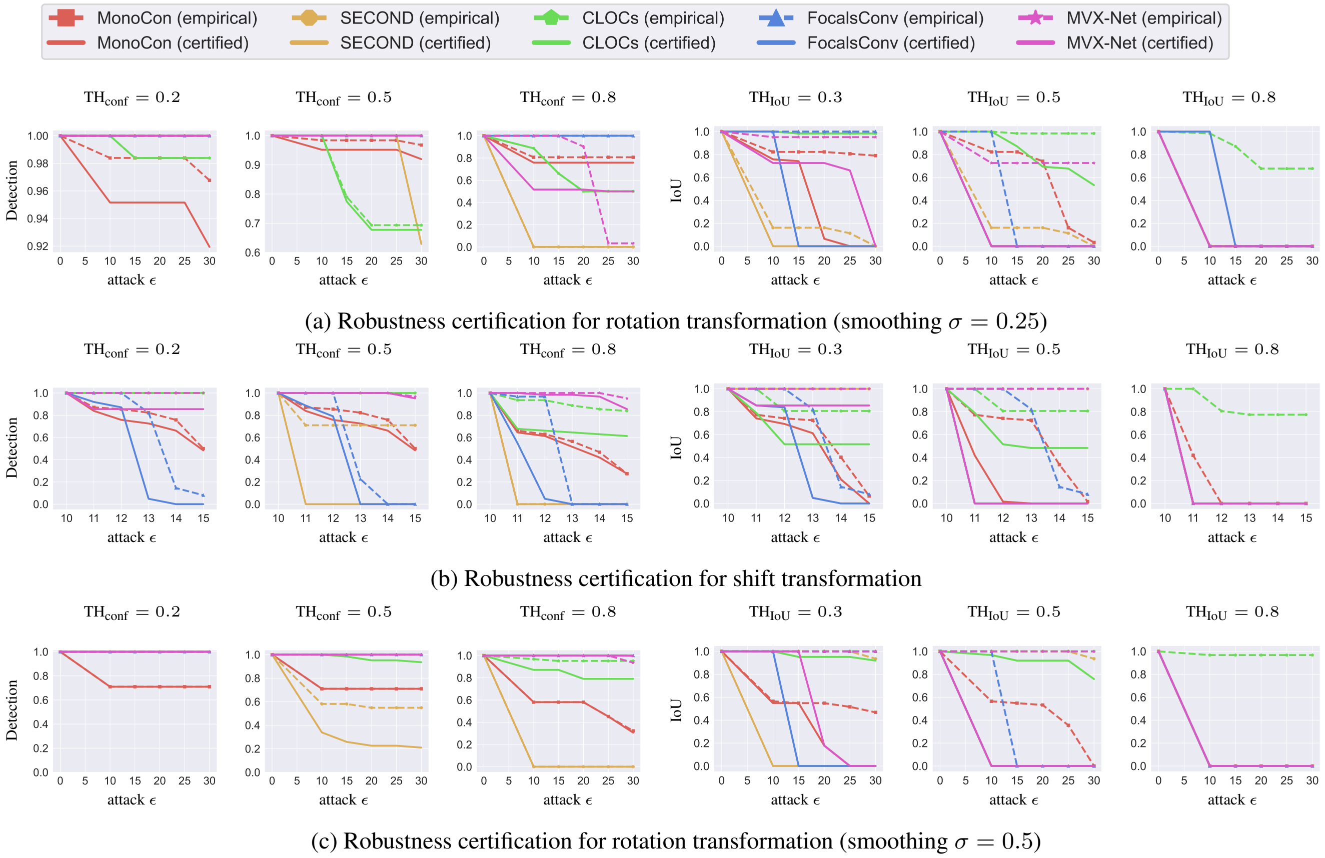
Table 4: Overview of shifting transformation experiment results. Each row represents the corresponding model and attack radius. “Benign”, “Adv (Vanilla)”, “Adv (Smoothed)”, and “Certification” stands for benign performance, vanilla models’ performance under attacks, smoothed models’ performance under attacks, and certified lower bound of smoothed model performance under attacks. Each column represents the results under different thresholds.

(a) Detection rate under shifting transformation

Model	Attack Radius	Benign			Adv (Vanilla)			Adv (Smoothed)			Certification		
		Det@20	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**
MonoCon (Liu, Xue, and Wu 2022)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	87.10%	87.10%	66.13%	87.10%	87.10%	66.13%	83.87%	83.87%	64.52%
	$10 \leq z \leq 12$				85.48%	85.48%	62.90%	85.48%	85.48%	62.90%	75.81%	75.81%	61.29%
	$10 \leq z \leq 13$				82.26%	82.26%	56.45%	82.26%	82.26%	56.45%	72.58%	72.58%	51.61%
	$10 \leq z \leq 14$				75.81%	75.81%	46.77%	75.81%	75.81%	46.77%	66.13%	66.13%	41.94%
	$10 \leq z \leq 15$				50.00%	50.00%	27.42%	50.00%	50.00%	27.42%	48.39%	48.39%	27.42%
SECOND (Yan, Mao, and Li 2018)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	100.00%	70.97%	0.00%	100.00%	70.97%	0.00%	100.00%	0.00%	0.00%
	$10 \leq z \leq 12$				100.00%	70.97%	0.00%	100.00%	70.97%	0.00%	100.00%	0.00%	0.00%
	$10 \leq z \leq 13$				100.00%	12.90%	0.00%	100.00%	70.97%	0.00%	100.00%	0.00%	0.00%
	$10 \leq z \leq 14$				100.00%	12.90%	0.00%	100.00%	70.97%	0.00%	100.00%	0.00%	0.00%
	$10 \leq z \leq 15$				100.00%	12.90%	0.00%	100.00%	70.97%	0.00%	100.00%	0.00%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	100.00%	100.00%	93.54%	100.00%	100.00%	93.54%	100.00%	100.00%	67.74%
	$10 \leq z \leq 12$				100.00%	100.00%	93.54%	100.00%	100.00%	93.54%	100.00%	100.00%	66.13%
	$10 \leq z \leq 13$				100.00%	100.00%	85.45%	100.00%	100.00%	88.71%	100.00%	100.00%	64.52%
	$10 \leq z \leq 14$				100.00%	100.00%	64.52%	100.00%	100.00%	85.48%	100.00%	100.00%	62.90%
	$10 \leq z \leq 15$				100.00%	100.00%	64.52%	100.00%	100.00%	83.87%	100.00%	100.00%	61.29%
FocalsConv (Chen et al. 2022)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	100.00%	100.00%	96.77%	100.00%	100.00%	96.77%	91.94%	88.71%	54.84%
	$10 \leq z \leq 12$				100.00%	100.00%	96.77%	100.00%	100.00%	96.77%	87.10%	79.03%	4.84%
	$10 \leq z \leq 13$				82.26%	22.58%	0.00%	82.26%	22.58%	0.00%	4.84%	0.00%	0.00%
	$10 \leq z \leq 14$				14.52%	0.00%	0.00%	14.52%	0.00%	0.00%	0.00%	0.00%	0.00%
	$10 \leq z \leq 15$				8.06%	0.00%	0.00%	8.06%	0.00%	0.00%	0.00%	0.00%	0.00%
MVX-Net (Sindagi, Zhou, and Tuzel 2019)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	100.00%	100.00%	88.71%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$10 \leq z \leq 12$				100.00%	100.00%	88.71%	100.00%	100.00%	100.00%	100.00%	100.00%	98.39%
	$10 \leq z \leq 13$				100.00%	100.00%	88.71%	100.00%	100.00%	100.00%	100.00%	100.00%	98.39%
	$10 \leq z \leq 14$				100.00%	100.00%	88.71%	100.00%	100.00%	100.00%	100.00%	100.00%	96.77%
	$10 \leq z \leq 15$				100.00%	100.00%	20.97%	100.00%	96.77%	95.16%	100.00%	96.77%	85.48%

(b) IoU with ground truth under shifting transformation

Model	Attack Radius	Benign			Adv (Vanilla)			Adv (Smoothed)			Certification		
		AP@30	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**
MonoCon (Liu, Xue, and Wu 2022)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	77.42%	77.42%	41.94%	77.42%	77.42%	41.94%	74.19%	41.94%	0.00%
	$10 \leq z \leq 12$				74.19%	74.19%	0.00%	74.19%	74.19%	0.00%	69.35%	1.61%	0.00%
	$10 \leq z \leq 13$				72.58%	72.58%	0.00%	72.58%	72.58%	0.00%	61.29%	0.00%	0.00%
	$10 \leq z \leq 14$				40.32%	33.87%	0.00%	40.32%	33.87%	0.00%	20.97%	0.00%	0.00%
	$10 \leq z \leq 15$				6.45%	1.61%	0.00%	6.45%	1.61%	0.00%	0.00%	0.00%	0.00%
SECOND (Yan, Mao, and Li 2018)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	93.55%	93.55%	93.55%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$10 \leq z \leq 12$				93.55%	93.55%	93.55%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$10 \leq z \leq 13$				87.10%	87.10%	87.10%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$10 \leq z \leq 14$				87.10%	87.10%	87.10%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$10 \leq z \leq 15$				87.10%	87.10%	87.10%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	93.55%	93.55%	93.55%	100.00%	100.00%	100.00%	79.03%	79.03%	0.00%
	$10 \leq z \leq 12$				80.65%	80.65%	80.65%	80.65%	80.65%	80.65%	51.61%	51.61%	0.00%
	$10 \leq z \leq 13$				80.65%	80.65%	77.42%	80.65%	80.65%	77.42%	51.61%	48.39%	0.00%
	$10 \leq z \leq 14$				80.65%	80.65%	77.42%	80.65%	80.65%	77.42%	51.61%	48.39%	0.00%
	$10 \leq z \leq 15$				80.65%	80.65%	77.42%	80.65%	80.65%	77.42%	51.61%	48.39%	0.00%
FocalsConv (Chen et al. 2022)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	97.77%	0.00%	0.00%	100.00%	100.00%	0.00%	85.48%	0.00%	0.00%
	$10 \leq z \leq 12$				0.00%	0.00%	0.00%	100.00%	100.00%	0.00%	83.87%	0.00%	0.00%
	$10 \leq z \leq 13$				0.00%	0.00%	0.00%	82.26%	82.26%	0.00%	4.84%	0.00%	0.00%
	$10 \leq z \leq 14$				0.00%	0.00%	0.00%	14.52%	14.52%	0.00%	0.00%	0.00%	0.00%
	$10 \leq z \leq 15$				0.00%	0.00%	0.00%	8.06%	8.06%	0.00%	0.00%	0.00%	0.00%
MVX-Net (Sindagi, Zhou, and Tuzel 2019)	$10 \leq z \leq 11$	100.00%	100.00%	100.00%	96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	85.48%	0.00%	0.00%
	$10 \leq z \leq 12$				96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	85.48%	0.00%	0.00%
	$10 \leq z \leq 13$				96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	85.48%	0.00%	0.00%
	$10 \leq z \leq 14$				96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	85.48%	0.00%	0.00%
	$10 \leq z \leq 15$				96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	85.48%	0.00%	0.00%



(a) Robustness certification for rotation transformation (smoothing $\sigma = 0.25$)

(b) Robustness certification for shift transformation

(c) Robustness certification for rotation transformation (smoothing $\sigma = 0.5$)

Figure 6: Robustness certification for rotation and shifting transformation, including detection rate bound and IoU bound. Solid lines represent the certified bounds, and dashed lines show the empirical performance under PGD.

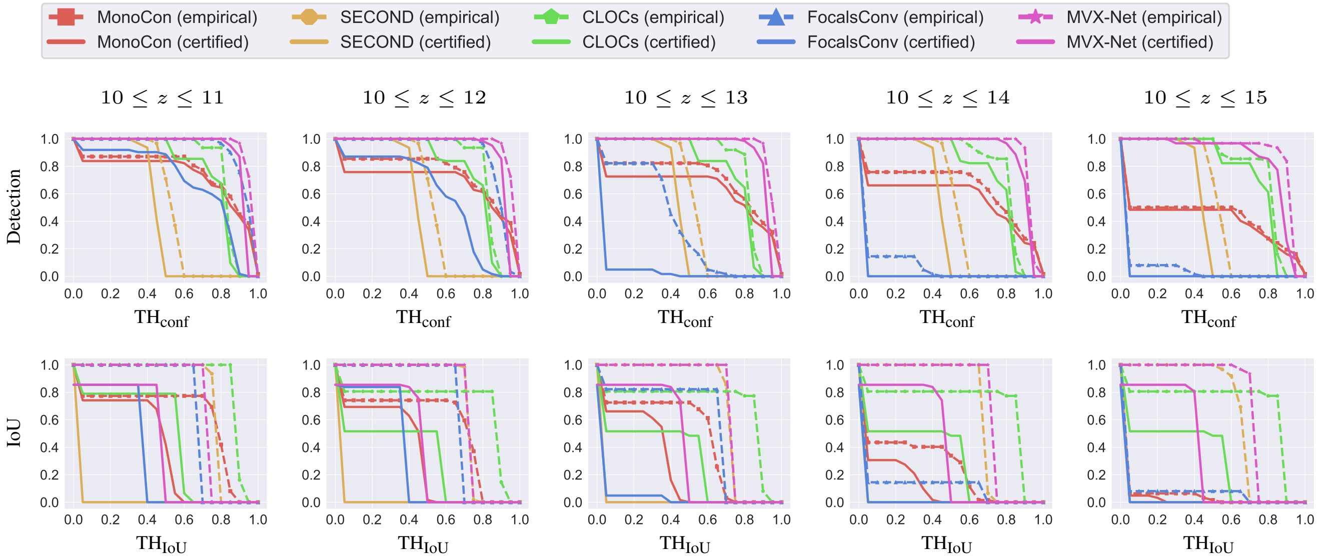


Figure 7: Robustness certification for shifting transformation, including detection rate bound and IoU bound. Solid lines represent the certified bounds, and dashed lines show the empirical performance under PGD.

Table 5: Overview of rotation transition experiment results with different sampling strategies under the condition "color 1 with buildings without pedestrian". Each row represents the corresponding model and attack radius. Columns "Certification (sparse)", and "Certification (dense)" represent the certified lower bound of performance under attacks with intervals of $0.1°$ and 600 intervals respectively.

(a) Certified rotation detection rate.

Model	Attack Radius	Certification (sparse)			Certification (dense)		
		Det@20	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**
MonoCon (Liu, Xue, and Wu 2022)	$r \pm 10^{\circ}$	93.33%	93.33%	86.67%	93.33%	93.33%	86.67%
	$r \pm 20^{\circ}$	93.33%	93.33%	86.67%	93.33%	93.33%	86.67%
	$r \pm 30^{\circ}$	86.67%	86.67%	86.67%	86.67%	86.67%	86.67%
SECOND (Yan, Mao, and Li 2018)	$r \pm 10^{\circ}$	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%
	$r \pm 20^{\circ}$	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%
	$r \pm 30^{\circ}$	100.00%	66.67%	0.00%	100.00%	66.67%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	$r \pm 10^{\circ}$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$r \pm 20^{\circ}$	100.00%	80.00%	73.33%	100.00%	80.00%	73.33%
	$r \pm 30^{\circ}$	100.00%	80.00%	73.33%	100.00%	80.00%	73.33%
FocalsConv (Chen et al. 2022)	$r \pm 10^{\circ}$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$r \pm 20^{\circ}$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$r \pm 30^{\circ}$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

(b) Certified rotation IoU.

Model	Attack Radius	Certification (sparse)			Certification (dense)		
		AP@30	**AP@50**	**AP@70**	**AP@30**	**AP@50**	**AP@70**
MonoCon (Liu, Xue, and Wu 2022)	$r \pm 10^{\circ}$	86.67%	0.00%	0.00%	86.67%	0.00%	0.00%
	$r \pm 20^{\circ}$	13.33%	0.00%	0.00%	13.33%	0.00%	0.00%
	$r \pm 30^{\circ}$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
SECOND (Yan, Mao, and Li 2018)	$r \pm 10^{\circ}$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$r \pm 20^{\circ}$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$r \pm 30^{\circ}$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	$r \pm 10^{\circ}$	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%
	$r \pm 20^{\circ}$	100.00%	93.33%	0.00%	100.00%	93.33%	0.00%
	$r \pm 30^{\circ}$	100.00%	53.33%	0.00%	100.00%	53.33%	0.00%
FocalsConv (Chen et al. 2022)	$r \pm 10^{\circ}$	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%
	$r \pm 20^{\circ}$	0.00%	0.00%	0.00%	13.33%	0.00%	0.00%
	$r \pm 30^{\circ}$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

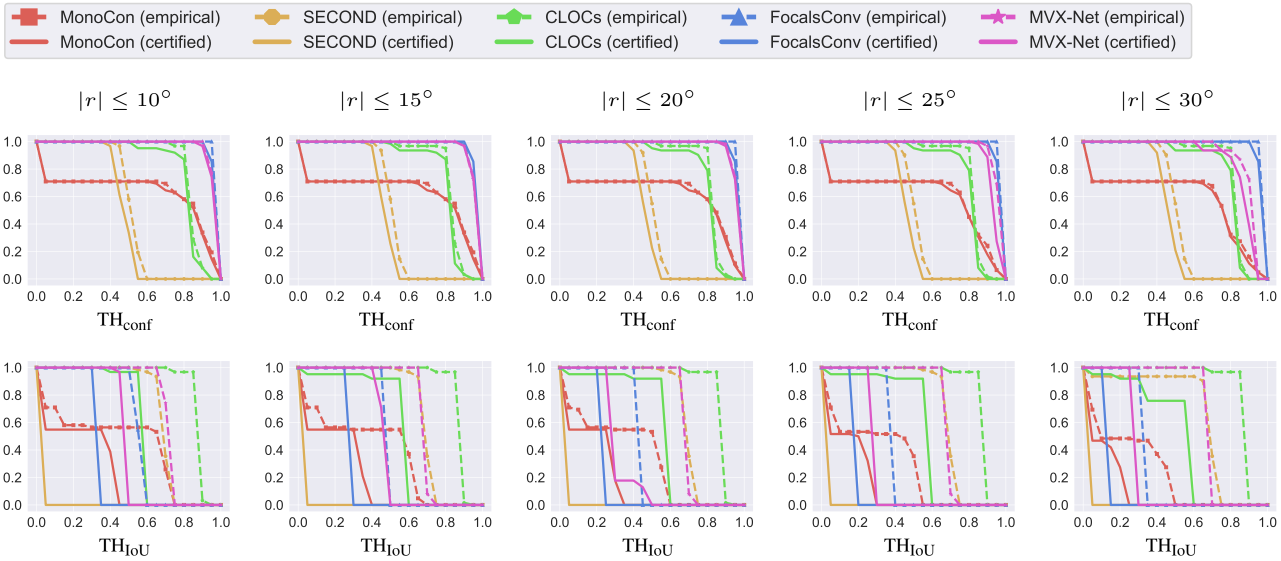


Figure 8: Robustness certification for rotation transformation (smoothing $\sigma = 0.5$), including detection rate bound and IoU bound. Solid lines represent the certified bounds, and dashed lines show the empirical performance under PGD. x-axis is the threshold for confidence score (TH_{conf}) and IoU score (TH_{IoU}), and y-axis is the ratio of detection whose confidence / IoU score is larger than the confidence / IoU threshold.

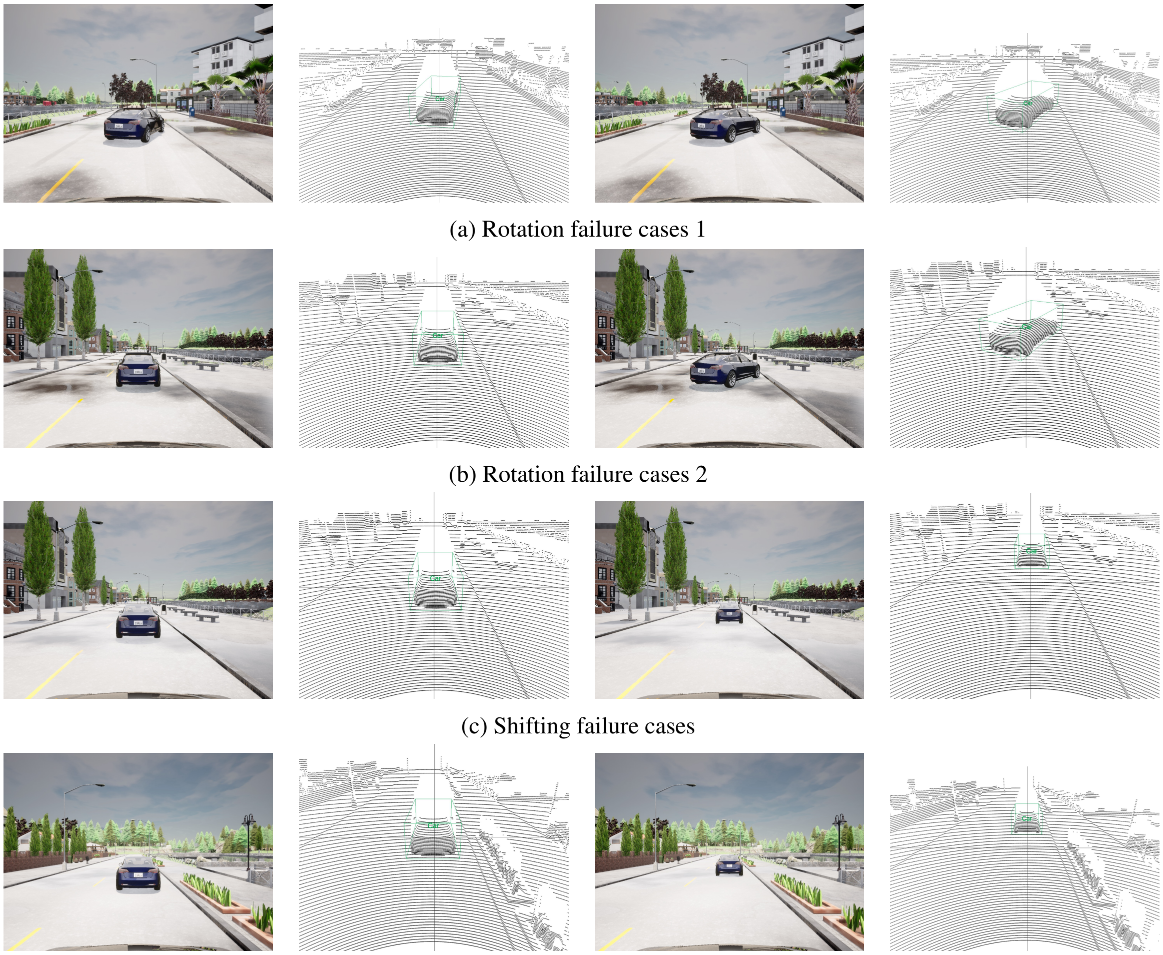
Table 6: Overview of rotation transformation experiment results (smoothing $\sigma = 0.5$). Each row represents the corresponding model and attack radius. “Benign”, “Adv (Vanilla)”, “Adv (Smoothed)”, and “Certification” stands for benign performance, vanilla models’ performance under attacks, smoothed models’ performance under attacks, and certified lower bound of smoothed model performance under attacks. Each column represents the results under different thresholds.

(a) Detection rate under rotation transformation

Model	Attack Radius		Benign			Adv (Vanilla)			Adv (Smoothed)			Certification	
		Det@20	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**	**Det@20**	**Det@50**	**Det@80**
MonoCon (Liu, Xue, and Wu 2022)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	70.97%	70.97%	58.06%	70.97%	70.97%	58.06%	70.97%	70.97%	58.06%
	$\lvert r\rvert \leq 15^\circ$				70.97%	70.97%	58.06%	70.97%	70.97%	58.06%	70.97%	70.97%	58.06%
	$\lvert r\rvert \leq 20^\circ$				70.97%	70.97%	58.06%	70.97%	70.97%	58.06%	70.97%	70.97%	58.06%
	$\lvert r\rvert \leq 25^\circ$				70.97%	70.97%	45.16%	70.97%	70.97%	45.16%	70.97%	70.97%	45.16%
	$\lvert r\rvert \leq 30^\circ$				70.97%	70.97%	32.26%	70.97%	70.97%	32.26%	70.97%	70.97%	30.65%
SECOND (Yan, Mao, and Li 2018)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%	100.00%	58.06%	0.00%	100.00%	33.87%	0.00%
	$\lvert r\rvert \leq 15^\circ$				100.00%	100.00%	0.00%	100.00%	58.06%	0.00%	100.00%	25.81%	0.00%
	$\lvert r\rvert \leq 20^\circ$				100.00%	100.00%	0.00%	100.00%	54.84%	0.00%	100.00%	22.58%	0.00%
	$\lvert r\rvert \leq 25^\circ$				100.00%	12.90%	0.00%	100.00%	54.84%	0.00%	100.00%	22.58%	0.00%
	$\lvert r\rvert \leq 30^\circ$				67.74%	3.23%	0.00%	100.00%	54.84%	0.00%	100.00%	20.97%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	96.77%	100.00%	100.00%	87.10%
	$\lvert r\rvert \leq 15^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	95.16%	100.00%	98.39%	87.10%
	$\lvert r\rvert \leq 20^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	95.16%	100.00%	95.16%	79.03%
	$\lvert r\rvert \leq 25^\circ$				100.00%	91.94%	20.97%	100.00%	100.00%	95.16%	100.00%	95.16%	79.03%
	$\lvert r\rvert \leq 30^\circ$				100.00%	74.19%	3.23%	100.00%	100.00%	95.16%	100.00%	93.55%	79.03%
FocalsConv (Chen et al. 2022)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 15^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 20^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 25^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 30^\circ$				100.00%	100.00%	98.39%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
MVX-Net (Sindagi, Zhou, and Tuzel 2019)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 15^\circ$				100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 20^\circ$				100.00%	100.00%	90.32%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 25^\circ$				100.00%	100.00%	3.23%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	$\lvert r\rvert \leq 30^\circ$				100.00%	100.00%	3.23%	100.00%	100.00%	93.55%	100.00%	100.00%	99.71%

(b) IoU with ground truth under rotation transformation

Model	Attack Radius		Benign			Adv (Vanilla)			Adv (Smoothed)			Certification	
		AP@30	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**	**AP@30**	**AP@50**	**AP@80**
MonoCon (Liu, Xue, and Wu 2022)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	56.45%	56.45%	0.00%	56.45%	56.45%	0.00%	54.84%	0.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				54.84%	54.84%	0.00%	54.84%	54.84%	0.00%	54.84%	0.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				54.84%	53.23%	0.00%	54.84%	53.23%	0.00%	17.74%	0.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				51.61%	35.48%	0.00%	51.61%	35.48%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				46.77%	0.00%	0.00%	46.77%	0.00%	0.00%	0.00%	0.00%	0.00%
SECOND (Yan, Mao, and Li 2018)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				83.87%	83.87%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				51.61%	51.61%	0.00%	93.55%	93.55%	0.00%	0.00%	0.00%	0.00%
CLOCs (Pang, Morris, and Radha 2020)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	90.32%	90.32%	90.32%	100.00%	100.00%	96.77%	100.00%	96.77%	0.00%
	$\lvert r\rvert \leq 15^\circ$				90.32%	90.32%	90.32%	100.00%	100.00%	96.77%	95.16%	91.93%	0.00%
	$\lvert r\rvert \leq 20^\circ$				88.71%	88.71%	77.42%	100.00%	100.00%	96.77%	95.16%	91.93%	0.00%
	$\lvert r\rvert \leq 25^\circ$				87.10%	87.10%	0.00%	100.00%	100.00%	96.77%	95.16%	91.93%	0.00%
	$\lvert r\rvert \leq 30^\circ$				88.71%	88.71%	3.26%	98.39%	98.39%	67.74%	98.39%	53.23%	0.00%
FocalsConv (Chen et al. 2022)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	100.00%	0.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				96.77%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				96.77%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				96.77%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				96.77%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
MVX-Net (Sindagi, Zhou, and Tuzel 2019)	$\lvert r\rvert \leq 10^\circ$	100.00%	100.00%	100.00%	96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	100.00%	0.00%	0.00%
	$\lvert r\rvert \leq 15^\circ$				96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	100.00%	0.00%	0.00%
	$\lvert r\rvert \leq 20^\circ$				96.77%	96.77%	0.00%	100.00%	100.00%	0.00%	17.74%	0.00%	0.00%
	$\lvert r\rvert \leq 25^\circ$				96.77%	75.81%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	$\lvert r\rvert \leq 30^\circ$				96.77%	75.81%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%



(a) Rotation failure cases 1

(b) Rotation failure cases 2

(c) Shifting failure cases

(d) Shifting failure cases

Figure 9: Rotation and shifting failure cases.