TPC: Transformation Specific Smoothing for Point Cloud Models

Wenda Chu[1], Linyi Li[2], Bo Li[2]

[1]Tsinghua University [2]University of Illinois Urbana-Champaign

Threat Model

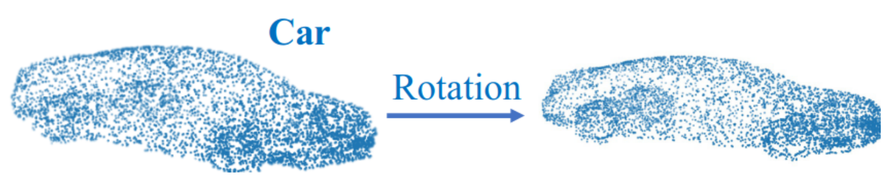
Semantic transformation attacks:

- Adversary can manipulate point clouds by semantic transformations.

Parameterized transformations:

$$\phi : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{X}$$

($\mathcal{X}$: Point cloud; $\mathcal{Z}$: Parameter space)



Goals

- Provide certified robustness conditions for point cloud classifiers against various semantic transformation attacks.
- Design concrete defense strategies and certification protocols for different transformation attacks based on randomized smoothing.

Certification goals:

Given a point cloud classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$

For a specific types of transformation, $\phi : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{X}$,

find a subset of parameters $\mathcal{Z}_{\text{robust}} \subseteq \mathcal{Z}$, such that,

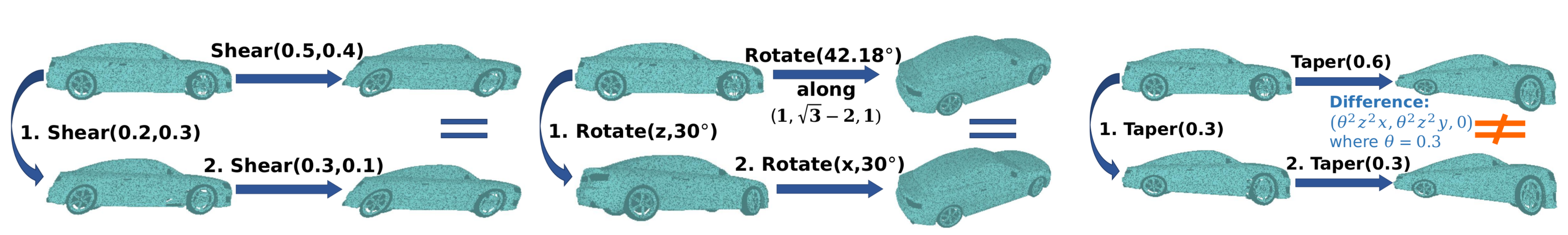
$$h(\phi(x, z)) = h(x), \forall x \in \mathcal{X}, z \in \mathcal{Z}_{\text{robust}}$$

Transformation Taxonomy

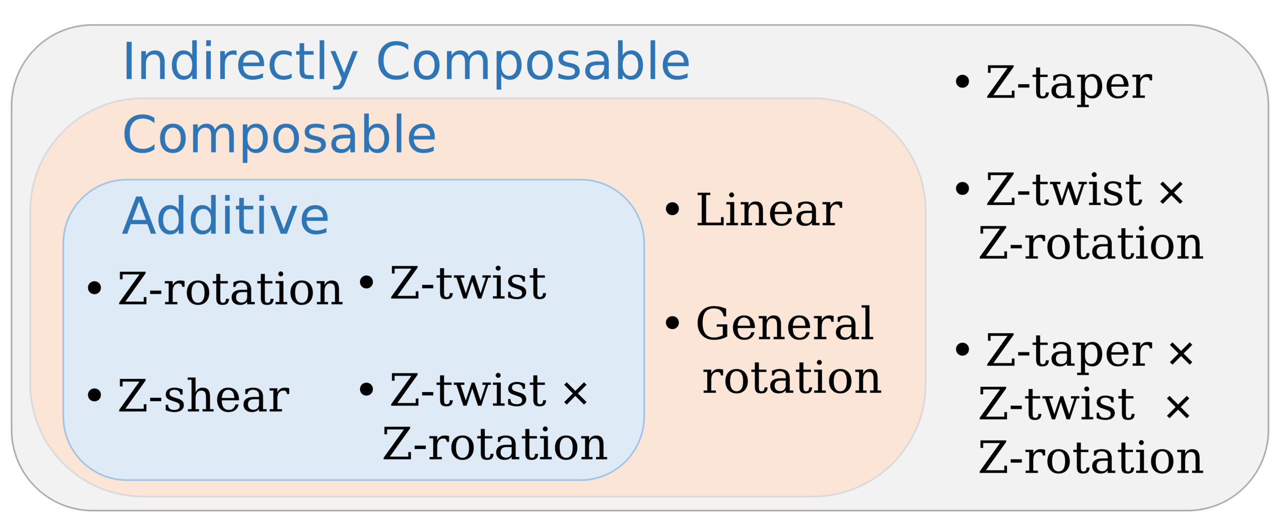
We categorize transformations into three classes based on composition property

1. Additive: $\phi(\phi(x, \alpha), \beta) = \phi(x, \alpha + \beta)$
2. Composable: $\phi(\phi(x, \alpha), \beta) = \phi(x, \gamma_\alpha(\beta))$
3. Indirectly composable: there exists a composable transformation $\psi : \mathcal{X} \times \mathcal{Z}_\psi \rightarrow \mathcal{X}$ such that $\forall x \in \mathcal{X}$ there exists a function $\delta_x : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathcal{Z}_\psi$ with

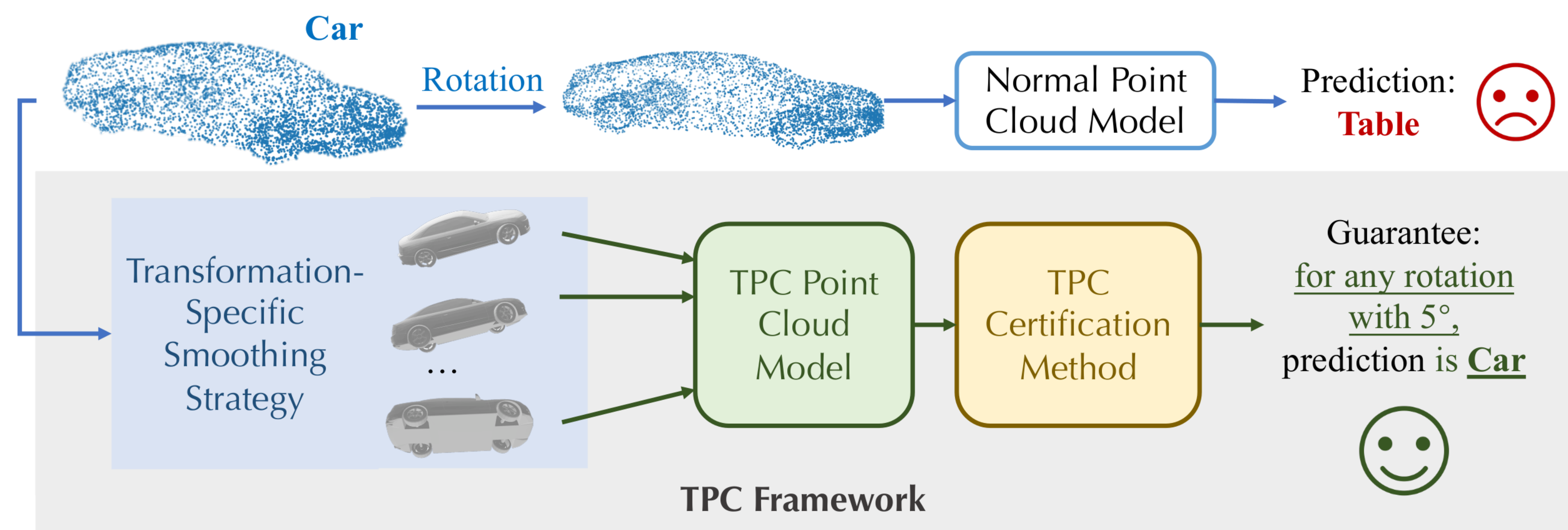
$$\phi(x, \alpha) = \psi(\phi(x, \beta), \delta_x(\alpha, \beta)), \forall \alpha, \beta \in \mathcal{Z}_\phi$$



(a) Additive (b) Composable (c) Indirectly composable



TPC Framework Overview



Transformation specific smoothed classifier

Definition 1 (Transformation Specific Smoothed Classifier) Suppose we have a base classifier $h : \mathcal{X} \rightarrow \mathcal{Y}$. For a given semantic transformation $\phi : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{X}$ and a random variable ϵ in the parameter space, the transformation specific smoothed classifier for this transformation is defined as

$$g(x; \epsilon) = \arg\max_{y \in \mathcal{Y}} q(y|x, \epsilon) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_\epsilon(p(y|\phi(x, \epsilon)))$$

Concrete certification protocols

Intuition 1: Additive transformations can be certified following the same protocol as that of additive noises. Suppose the class probability of the smoothed classifier satisfies $q(y_A|x, \epsilon) \geq p_A > p_B \geq \max q(y|x, \epsilon)$. The classifier is guaranteed to be robust if $\|\alpha\|_2 \leq \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right), \alpha \in \mathcal{Z}$.

Intuition 2: The above does not hold for composable but not additive transformations. We take linear transformations as an example: $\phi(x, \alpha) = (I + \alpha)x, \quad \alpha \in \mathbb{R}^{3\times 3}$

We guarantee the robustness of the smoothed classifier when

$$\|\alpha\|_F \leq R, \quad R = \frac{\sigma\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(1 - \tilde{p}_A)\right)}{2 + \sigma\left(\Phi^{-1}(\tilde{p}_A) - \Phi^{-1}(1 - \tilde{p}_A)\right)}.$$

Intuition 3: For indirectly composable transformations, we draw multiple samples in the parameter space and certify the neighbor areas of these samples by adding additive noise. These small certified areas are combined to cover a larger parameter space.

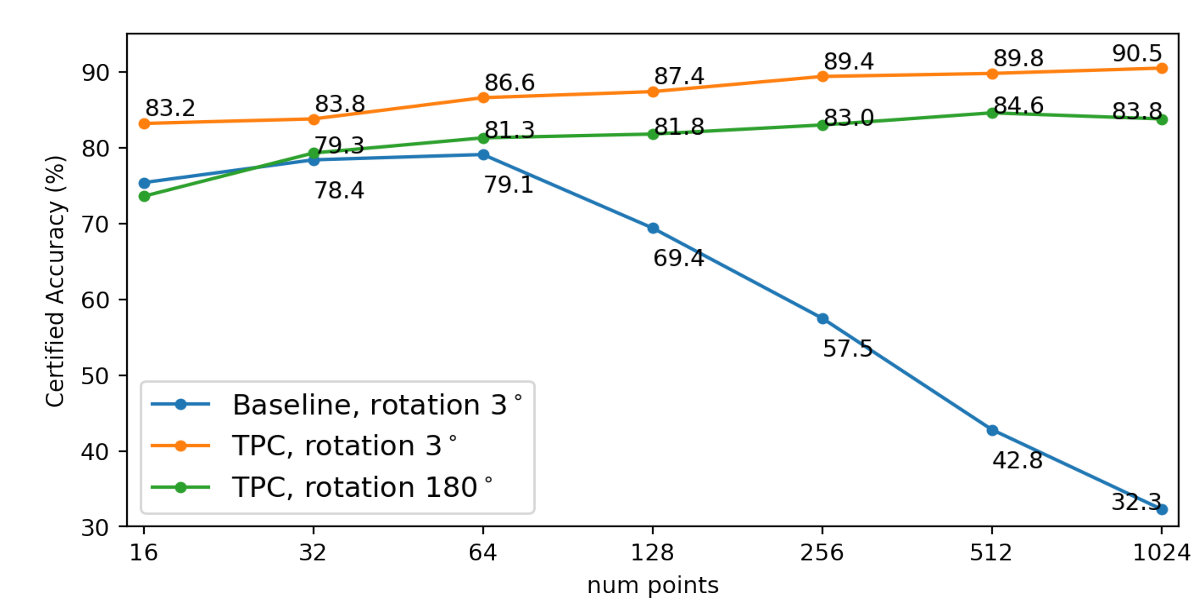
Certification Pipeline:

- Bound the interpolation error $\mathcal{M}_\mathcal{Z} := \max_{\alpha \in \mathcal{Z}} \min_{1 \leq i \leq N} \mathcal{M}(\alpha, \alpha_i)$ $= \max_{\alpha \in \mathcal{Z}} \min_{1 \leq i \leq N} \|\phi(x, \alpha) - \phi(x, \alpha_i)\|_2$
- Guaranteed to be robust if $\mathcal{M}_\mathcal{Z} \leq \frac{\sigma}{2} \min_{1 \leq i \leq N} \left(\Phi^{-1}(p_A^{(i)}) - \Phi^{-1}(p_B^{(i)})\right)$

Numerical Results

- We evaluate our TPC method on ModelNet40 dataset, using a model with PointNet architecture.
- Metric: certified accuracy, defined by the fraction of point clouds that are classified both *correctly* and *consistently* within certain transformation space.
- Our TPC method scales up well to large point clouds. The certified accuracy increases as the number of points increases.

Transformation	Attack radius	Certified Accuracy (%) TPC	Certified Accuracy (%) DeepG3D
ZYX-rotation	2°	**81.4**	61.6
	5°	**69.2**	49.6
General rotation	5°	**78.5**	-
	10°	**69.2**	-
	15°	**55.5**	-
Z-rotation	20°	**84.2**	81.8
	60°	**83.8**	81.0
	180°	**81.3**	-
Z-shear	0.03	**83.4**	59.8
	0.1	**82.2**	-
	0.2	**77.7**	-
Z-twist	20°	**83.8**	20.3
	60°	**80.1**	-
	180°	**64.3**	-
Z-taper	0.1	**78.1**	69.0
	0.2	**76.5**	23.9
	0.5	**66.0**	-
Linear	0.1	**74.0**	-
	0.2	**59.9**	-
Z-twist ∘ Z-rotation	20°, 1°	**78.9**	13.8
	20°, 5°	**78.5**	-
	50°, 5°	**76.9**	-
Z-taper ∘ Z-rotation	0.1, 1°	**76.1**	58.2
	0.2, 1°	**72.9**	17.5
Z-twist ∘ Z-taper ∘ Z-rotation	10°, 0.1, 1°	**68.8**	17.5
	20°, 0.2, 1°	**63.1**	4.6



Paper Code Author Lab