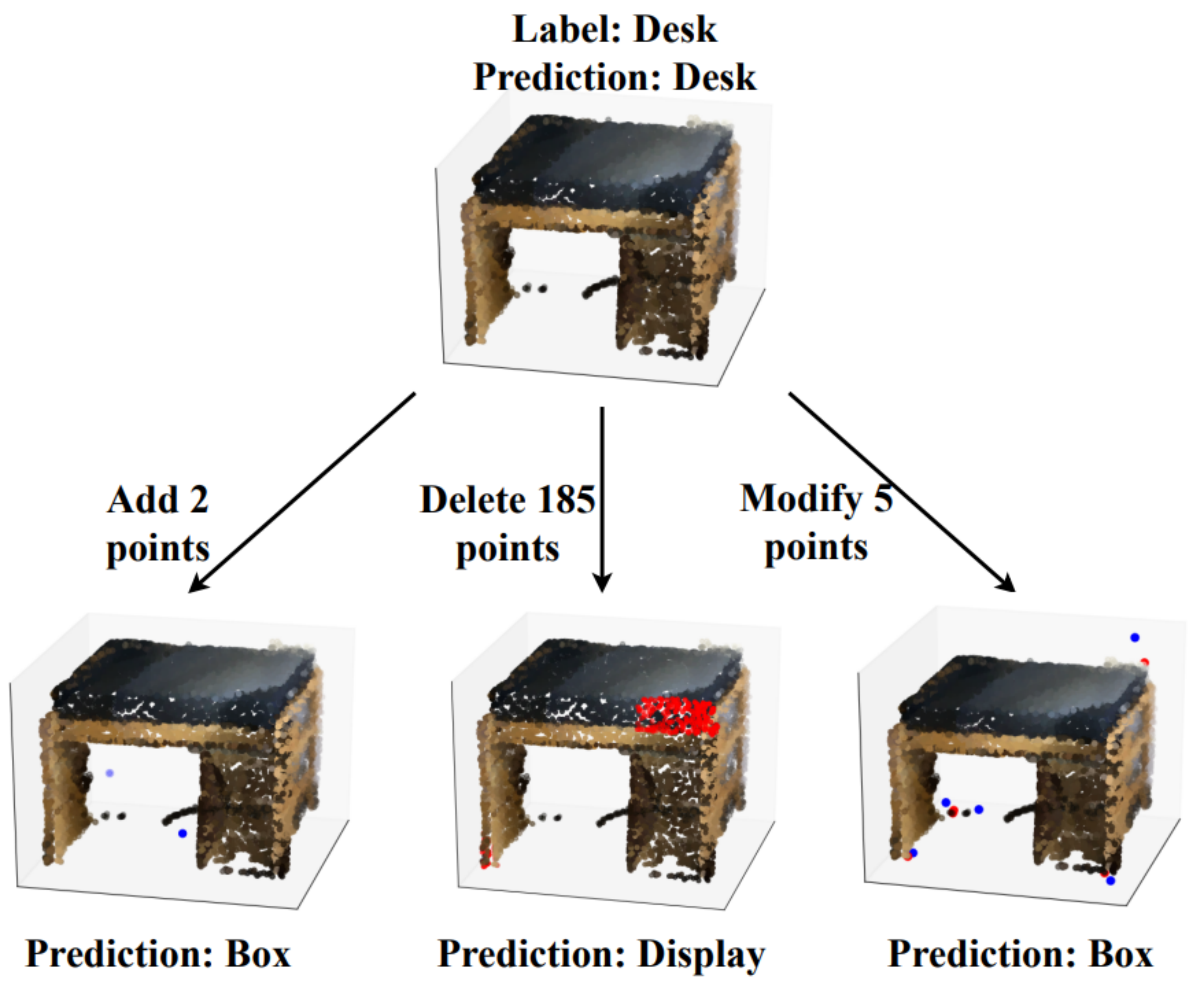
PaperTag: WED-AM-121

PointCert: Point Cloud Classification with Deterministic Certified Robustness Guarantees

Jinghuai Zhang[1] *Jinyuan Jia*[2] *Hongbin Liu*[1] *Neil Zhenqiang Gong*[1]
Duke University[1] UIUC[2]

Motivation

Point cloud classifiers are vulnerable to *adversarial point clouds*.



Overview

We propose PointCert, the first certified defense that has deterministic robustness guarantees against adversarial point clouds. Moreover, we propose methods to optimize the performance of PointCert in multiple application scenarios.

Motivation

1. Existing empirical defenses cannot provide formal guarantees and are often broken by advanced and adaptive attacks [1].

2. Existing certified defenses [2, 3] produce incorrect robustness guarantees with some probability, i.e., their certified robustness guarantees are probabilistic.

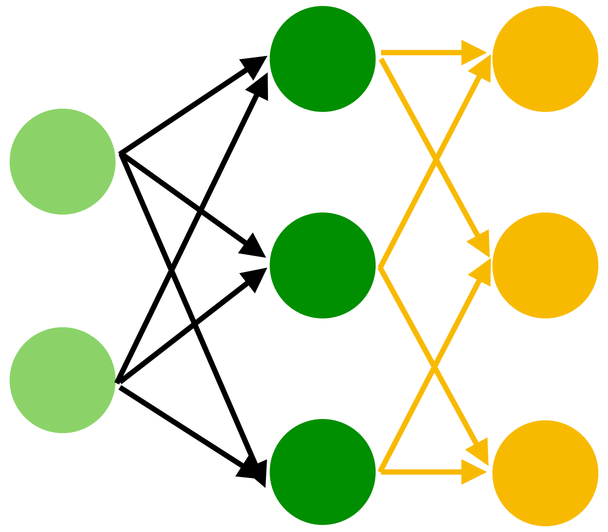
Our work

We propose PointCert, the first certified defense that has deterministic robustness guarantees against adversarial point clouds.

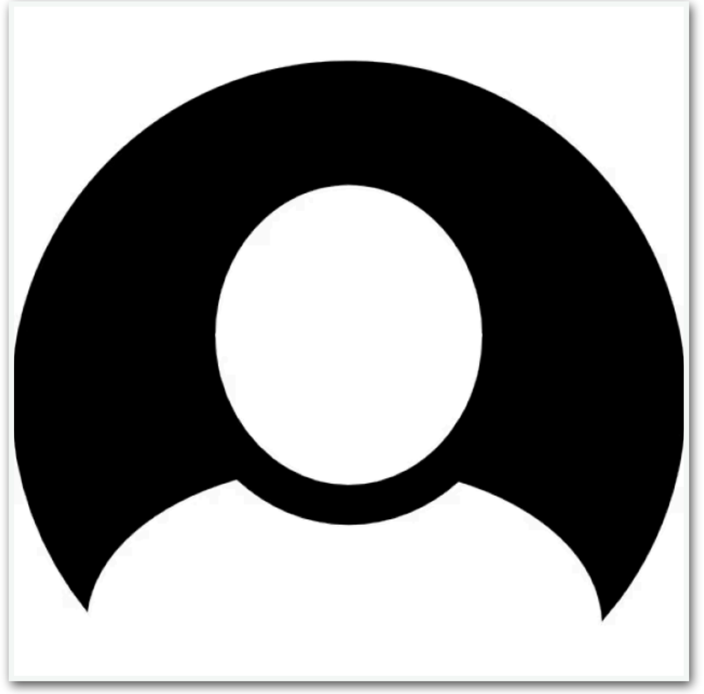
PointCert certifiably predicts the same label for a point cloud when the number of points arbitrarily added, deleted, modified by an attacker is less than the certified perturbation size.

Our work

Optimize PointCert in three scenarios, in which base point cloud classifier *f* is trained by the model provider differently and/or used by a customer differently.



Base Point Cloud Classifier f

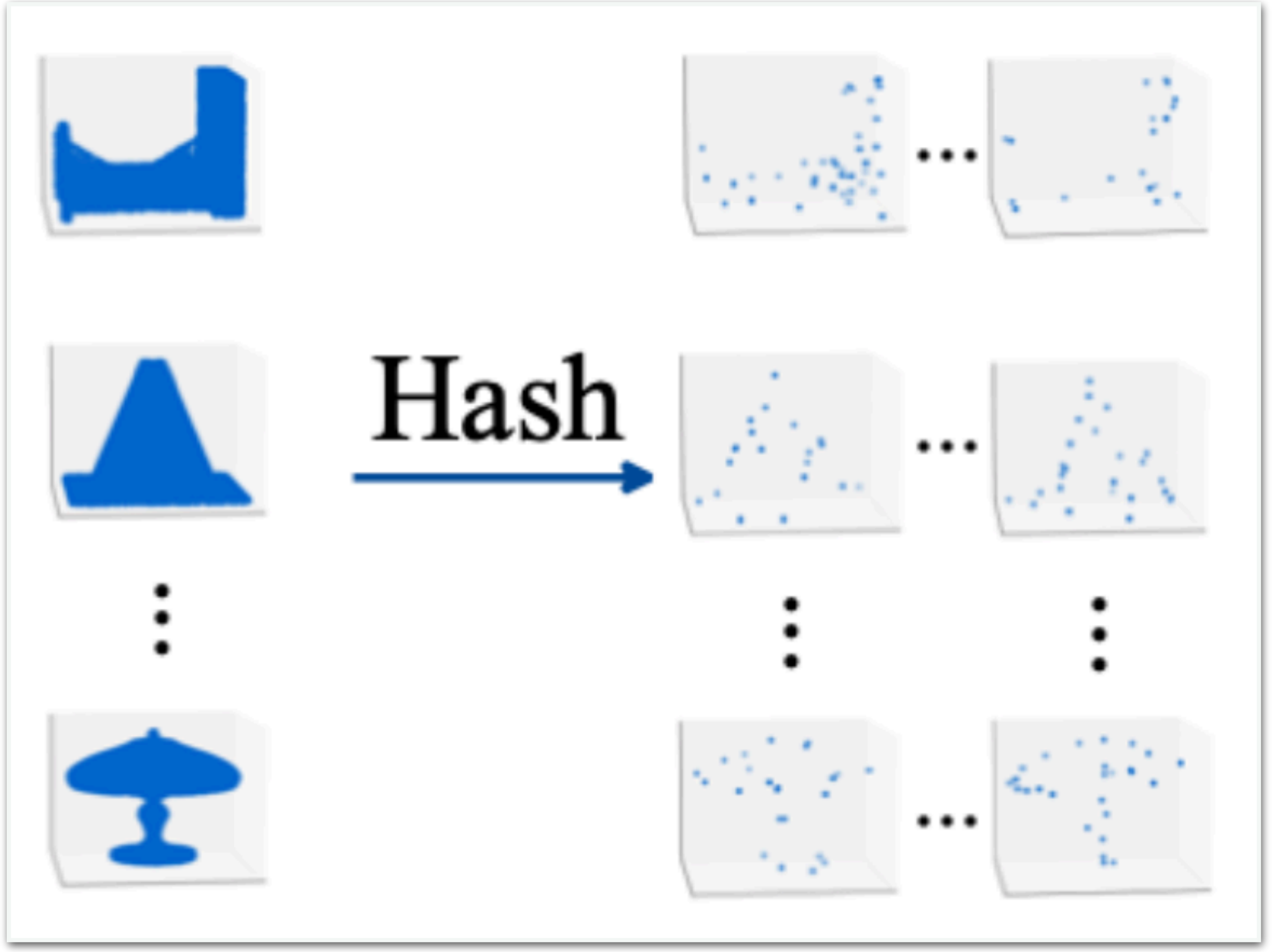


Model provider

Customer

Key idea

Step 1. Dividing a point cloud into m disjoint subpoint clouds using cryptographic hash function (e.g., MD5).

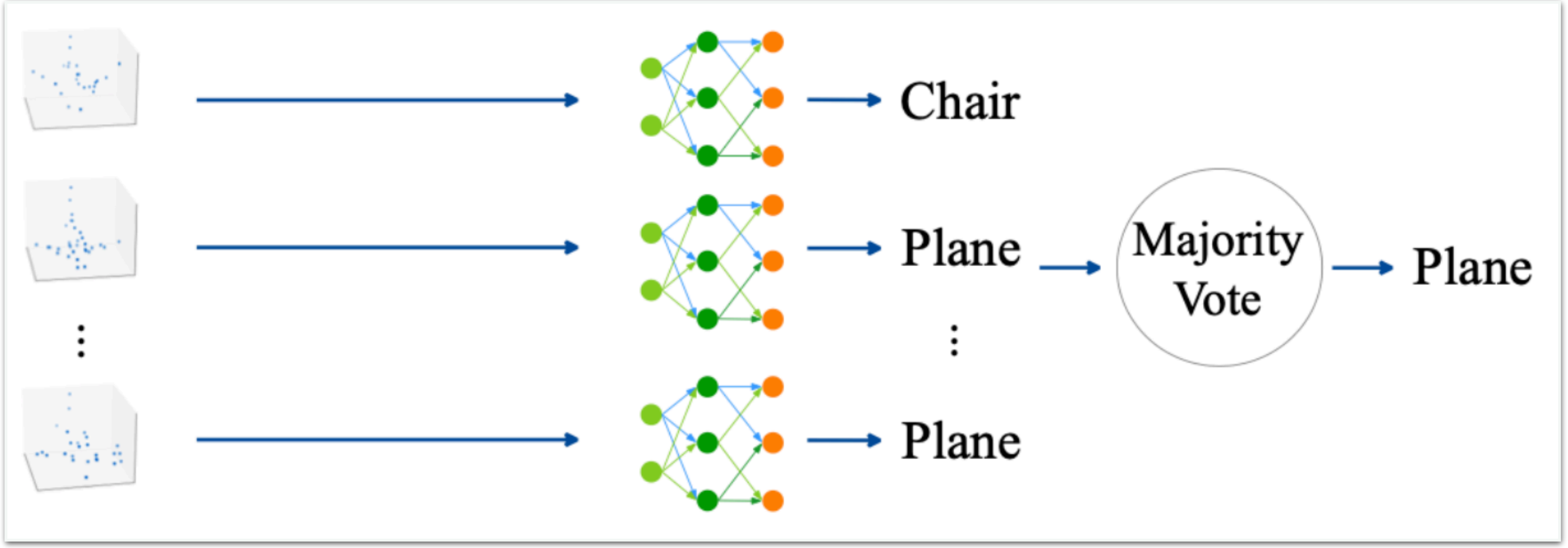


Key idea

Step 2. Building an ensemble point cloud classifier *h*.

h predicts label *y* for a point cloud *P* *if:* $M_y(P) \geq \max_{l \neq y}(M_l(P) + \mathbb{I}(y > l))$

where $M_l(P)$ indicates label frequency for label *l*.



Theoretical Analysis

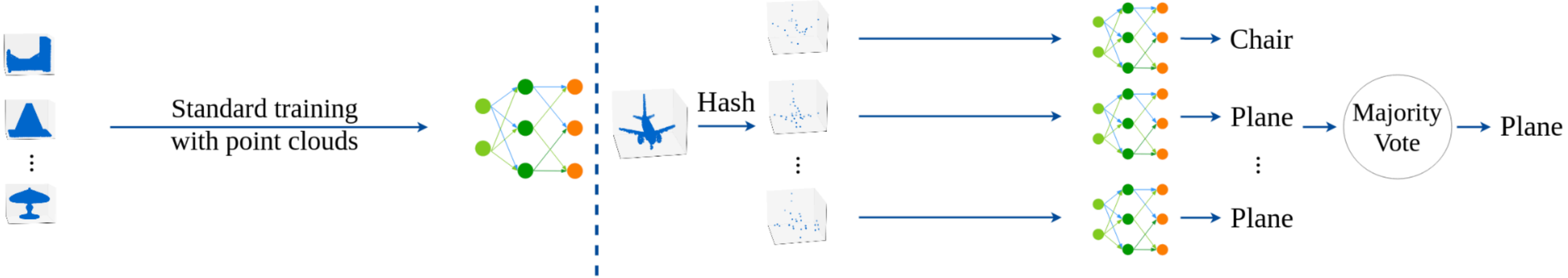
Derive the largest certified perturbation size *t(P)* such that our PointCert is guaranteed to predicts the same label *y* for *P* and its adversarially perturbed version:

$$t(P) = \left\lfloor \frac{M_y(P) - \max_{l \neq y}(M_l(P) + \mathbb{I}(y > l))}{2 \cdot \tau} \right\rfloor$$

τ is 1 for point addition and deletion attacks, while it is 2 for point modification and perturbation attacks.

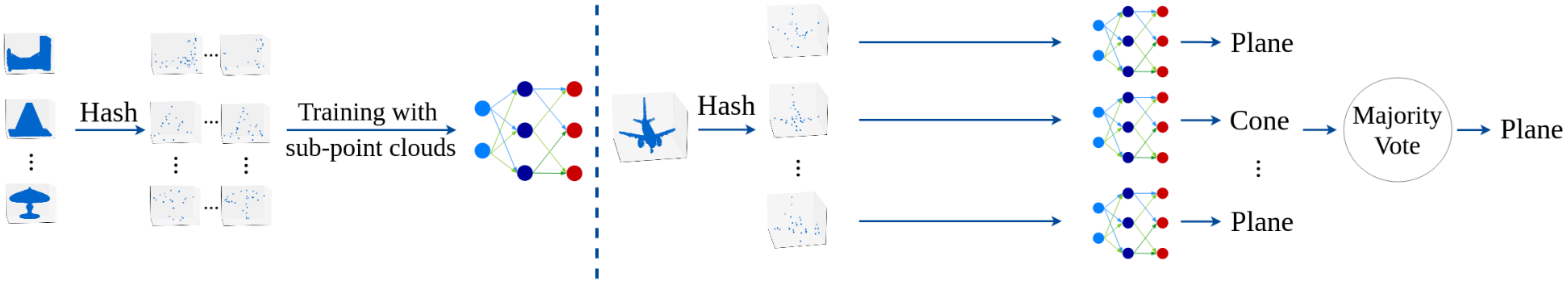
Application Scenario I

Naive application of PointCert.



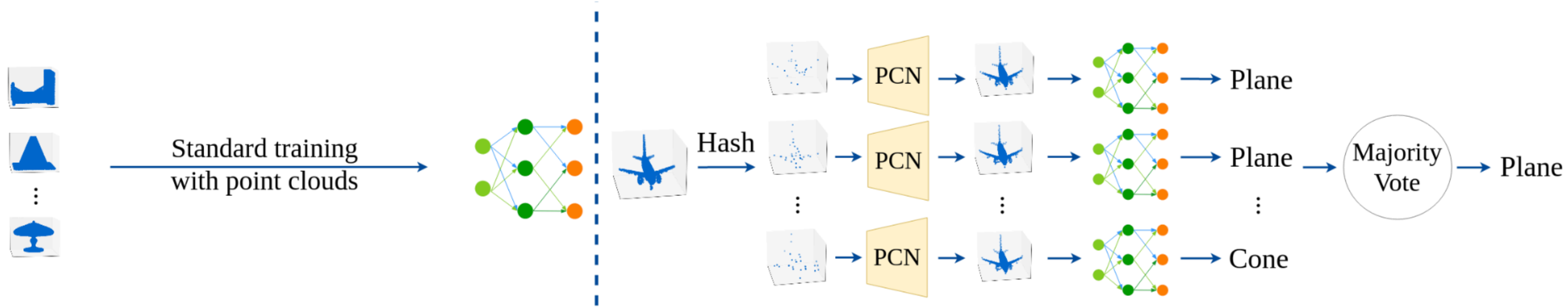
Application Scenario II

The model provider trains *f* on sub-point clouds to optimize the performance of PointCert.

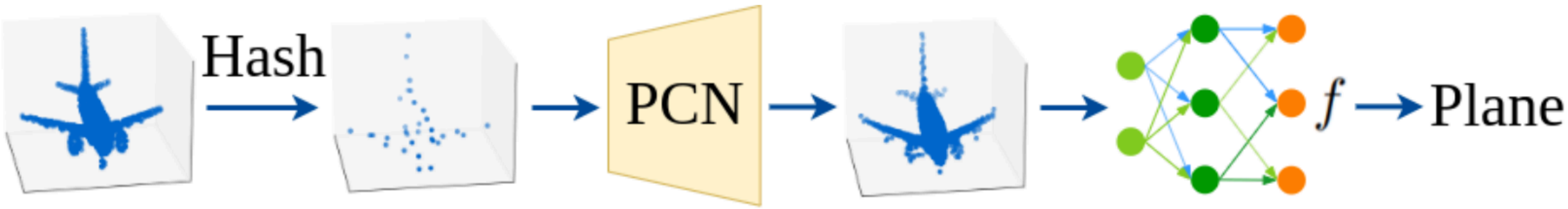


Application Scenario III

The customer trains a Point Completion Network [4] using unlabeled/ partially labeled data and adds it before f to improve the accuracy for subpoint clouds.



Application Scenario III



Completion loss $\longrightarrow$ $L_p(\mathcal{D}_u, \mathcal{C}) + \lambda \cdot L_c(\mathcal{D}_l, \mathcal{C}, f)$ $\longleftarrow$ **CE loss**

Experimental results

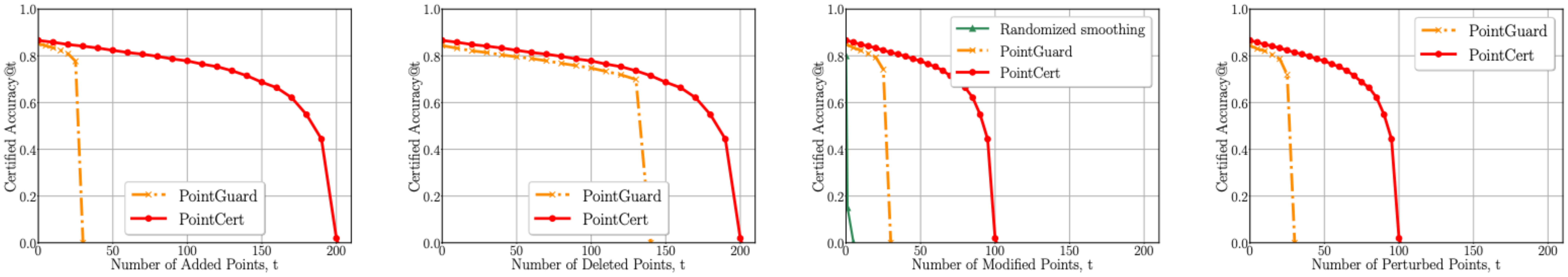
Dataset: ModelNet40[5] and ScanObjectNN[6]. We split the training point clouds into two balanced halves. One is used for the model provider to train base point cloud classifiers f, and the other is used for a customer to train a PCN in Scenario III.

Compared methods: Randomized smoothing [2], PointGuard[3].

Certified Accuracy@t: The fraction of testing point clouds whose certified perturbation sizes are at least t and whose labels are correctly predicted.

Certified accuracy

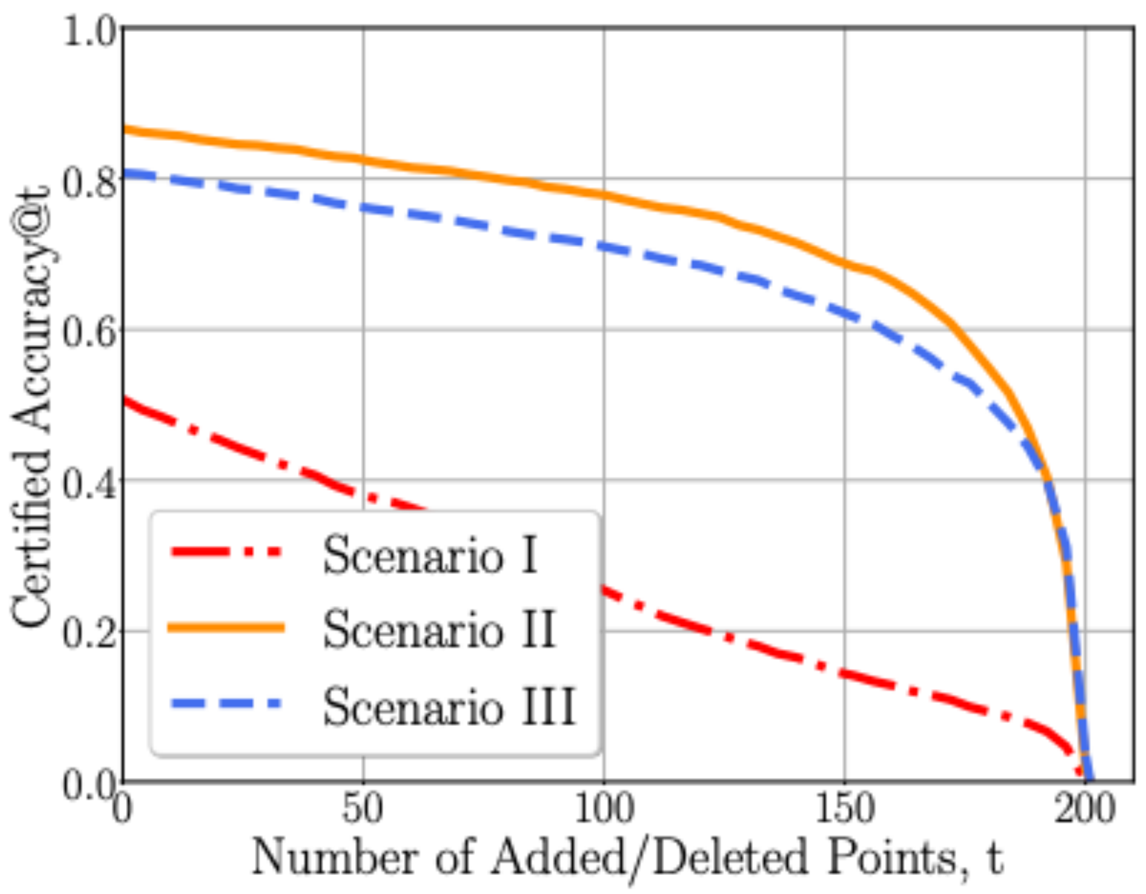
Make each method have similar certified accuracy under no attacks.



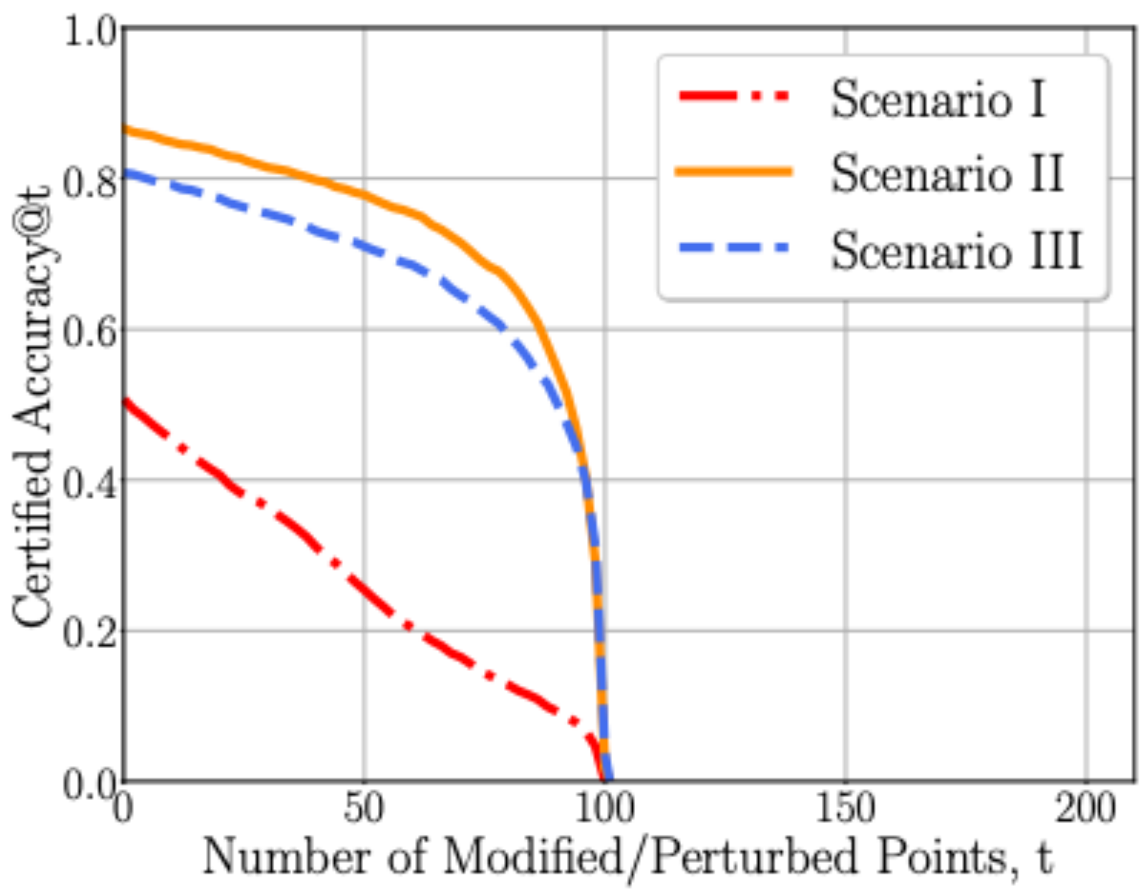
(a) Point addition attacks (b) Point deletion attacks (c) Point modification attacks (d) Point perturbation attacks

Comparing the certified accuracy of different defenses. (Scenario II)

Different Scenarios



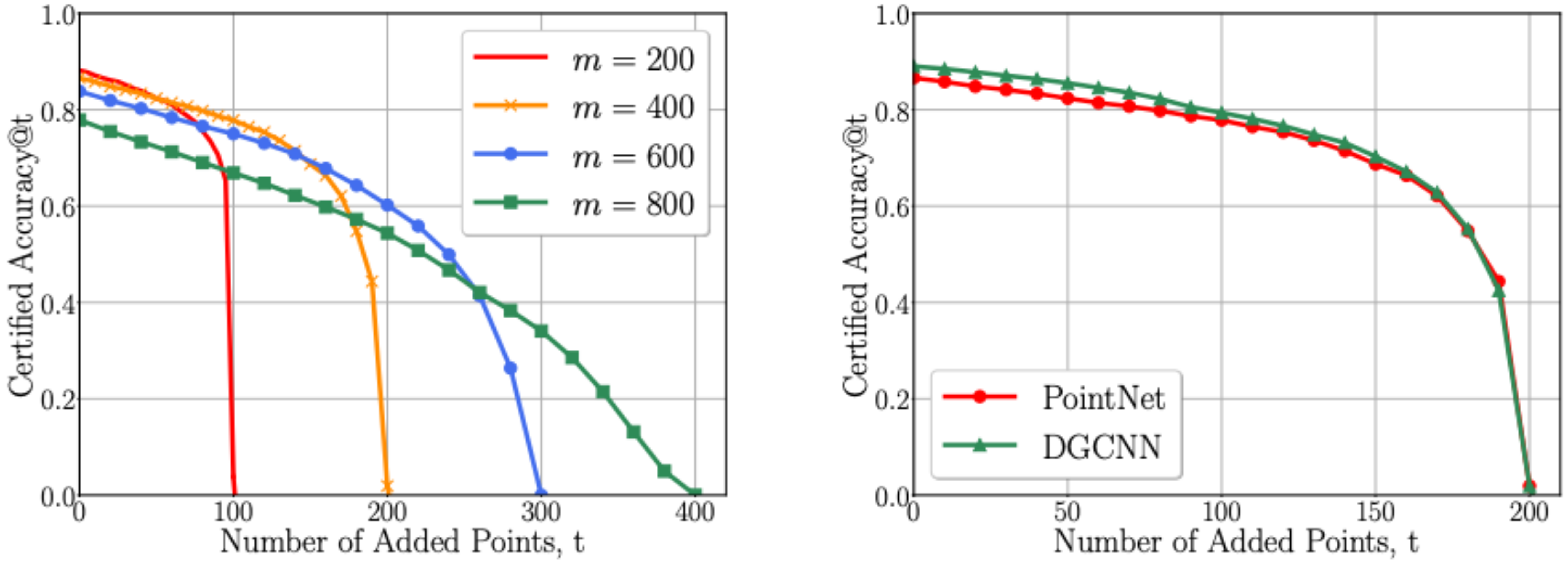
(a) Point addition/deletion attacks



(b) Point modification/perturbation attacks

Comparing the certified accuracy in three application scenarios under different attacks.

Ablation Study

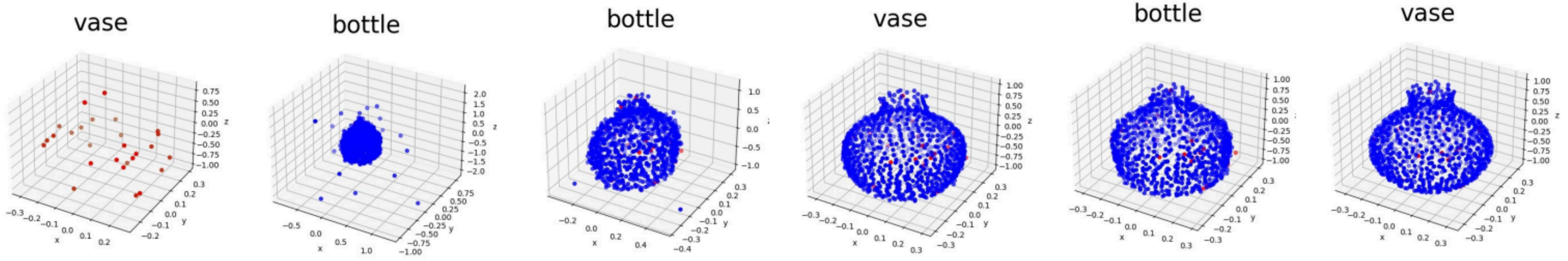


(a) (b)

Impact of (a) the number of sub-point clouds *m*. (b) different *f*. (Scenario II)

Ablation Study

$$L_p(\mathcal{D}_u, \mathcal{C}) + \lambda \cdot L_c(\mathcal{D}_l, \mathcal{C}, f)$$



(a) input PC (b) λ=0.05 (c) λ=0.005 (d) λ=0.0005 (e) λ=0.00005 (f) gt

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Reference

[1] Jiachen Sun, Karl Koenig, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. On adversarial robustness of 3d point cloud classification under adaptive attacks. *arXiv preprint arXiv:2011.11922*, 2020.
[2] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.
[3] Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. Pointguard: Provably robust 3d point cloud classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
[4] Wentao Yuan, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert. Pcn: Point completion network. In International Conference on 3D Vision (3DV), 2018.
[5] ModelNet40. https://modelnet.cs.princeton. edu/, 2015.
[6] ScanObjectNN. https://github.com/hkust-vgd/ scanobjectnn, 2019.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Thanks for listening!

Code available at *https://github.com/jzhang538/PointCert.*