Toward Availability Attacks in 3D Point Clouds

Yifan Zhu [* 1 2]   **Yibo Miao** [* 1 2]   **Yinpeng Dong** [3 4]   **Xiao-Shan Gao** [1 2 5]

Abstract

Despite the great progress of 3D vision, data privacy and security issues in 3D deep learning are not explored systematically. In the domain of 2D images, many availability attacks have been proposed to prevent data from being illicitly learned by unauthorized deep models. However, unlike images represented on a fixed dimensional grid, point clouds are characterized as unordered and unstructured sets, posing a significant challenge in designing an effective availability attack for 3D deep learning. In this paper, we theoretically show that extending 2D availability attacks directly to 3D point clouds under distance regularization is susceptible to the degeneracy, rendering the generated poisons weaker or even ineffective. This is because in bi-level optimization, introducing regularization term can result in update directions out of control. To address this issue, we propose a novel Feature Collision Error-Minimization (FC-EM) method, which creates additional shortcuts in the feature space, inducing different update directions to prevent the degeneracy of bi-level optimization. Moreover, we provide a theoretical analysis that demonstrates the effectiveness of the FC-EM attack. Extensive experiments on typical point cloud datasets, 3D intracranial aneurysm medical dataset, and 3D face dataset verify the superiority and practicality of our approach. Code is available at https://github.com/hala64/fc-em.

1. Introduction

Vision for 3D has been developed rapidly and become more popular in real-world applications, such as autonomous driv-

*Equal contribution [1]Academy of Mathematics and Systems Science, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences [3]Tsinghua University [4]RealAI [5]Kaiyuan International Mathematical Sciences Institute. Correspondence to: Xiao-Shan Gao <xgao@mmrc.iss.ac.cn>, Yibo Miao <miaoyibo@amss.ac.cn>.

ing (Chen et al., 2017b; Yue et al., 2018), medical image processing (Taha & Hanbury, 2015), scene reconstruction (Malihi et al., 2016), and face recognition (Zhou & Xiao, 2018). Since PointNet (Qi et al., 2017a) was first proposed, many deep learning-based methods have been applied to the 3D domain and shown tremendous success in various tasks. Although great progress has been made, data privacy and security issues in 3D deep learning are not explored systematically. Users may be reluctant to contribute their privacy-sensitive data, such as online healthcare records, to training large-scale commercial 3D models (Xu et al., 2023; Zhou et al., 2023). In fact, according to a report by Das (2023), an insurance company illicitly acquired a substantial amount of biomedical data to develop a commercial chronic disease risk prediction model. Another investigation indicated a growing number of lawsuits between data owners and machine learning companies (Vincent, 2019; Burt, 2020; Conklin, 2020). Consequently, there is an increasing focus on safeguarding data from unauthorized use for training.

In the realm of 2D images, many availability attacks have been proposed to prevent data from being illicitly learned by unauthorized deep models (Feng et al., 2019; Huang et al., 2020; Fowl et al., 2021; Fu et al., 2022; Sandoval-Segura et al., 2022b). They add imperceptible perturbations to the training data so that the model cannot learn much information from the data and thereby significantly reducing the model's accuracy on test data. Within the literature, two representative 2D availability attack methods stand out. Huang et al. (2020) propose an error-minimization poisoning approach, known as Unlearnable Examples, leveraging iterative optimization of a bi-level min-min problem to create shortcuts by minimizing network loss. On the other hand, Fowl et al. (2021) propose an error-maximization poisoning approach, known as Adversarial Poisons, employing stronger adversarial attacks on the pre-trained source classifier to generate poisoning noises.

However, to date, there is a notable absence of research in the domain of data privacy and security concerning 3D data. The data structure of 3D point clouds is inherently different from 2D images, posing a significant challenge in designing an effective availability attack. Specifically, unlike images represented on a fixed dimensional grid, point clouds are characterized as unordered and unstructured sets. As shown in Figure 1, the poisons have noticeable outliers, lacking im-
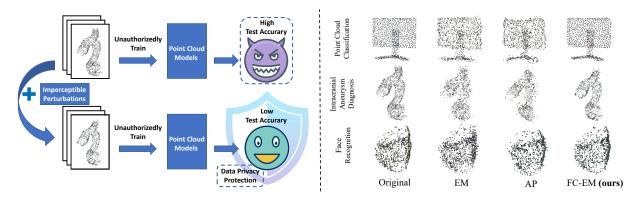
Figure 1. **Left:** An illustration of availability attacks. The poisoner adds imperceptible perturbations to the training data, aiming to reduce the model's generalization ability and prevent data from being illicitly learned by unauthorized deep models. **Right:** An illustration of poisons crafted by EM (Huang et al., 2020), AP (Fowl et al., 2021), and our FC-EM attack for point cloud classification, intracranial aneurysm diagnosis, and face recognition tasks. Notably, poisons generated by EM and AP have noticeable outliers, lacking imperceptibility. Poisons generated by our FC-EM are more natural and imperceptible, maintaining the semantic integrity.

perceptibility, thus making the ℓ_p norm constraints used in 2D availability attacks less suitable (Miao et al., 2022). On the other hand, the Chamfer distance (Fan et al., 2017) as an alternative metric is often used as a regularization term as it is difficult to directly constrain point clouds through projection onto a hypersphere due to the presence of the minimum operator. However, as theoretically analyzed in Section 4, extending 2D poison methods directly to 3D point clouds under distance regularization can be susceptible to the degeneracy of poisons, making the generated poisons weaker or even ineffective. This is because the complexity introduced by the regularization term in bi-level optimization leads to the optimization directions out of control.

To address this issue, we propose a novel method, **Feature Collision Error-Minimization (FC-EM)**. For the bi-level min-min problem of availability attacks, during the inner minimization, FC-EM employs our proposed feature collision loss rather than the original cross-entropy loss to optimize the poisoning noise, which creates additional shortcuts in the feature space to encourage the concentration of intra-class features and the dispersion of inter-class features. By doing so, different optimization directions can prevent the degeneracy of bi-level optimization, thereby breaking the equilibrium. We also provide a theoretical analysis that demonstrates the effectiveness of the FC-EM attack. In summary, we propose the first availability attack specifically designed for point cloud classifiers and hope it could serve as a baseline for further studies into the data privacy and security issues of 3D deep learning.

Extensive experiments on typical point cloud recognition models (Qi et al., 2017a;b; Wang et al., 2019) demonstrate that, in comparison with other baseline poisoning methods, our FC-EM method significantly reduces the model's generalization ability, while making the generated poisons more natural and imperceptible. Experiments on real-world datasets, including a 3D intracranial aneurysm medical

dataset (Yang et al., 2020) and a 3D face dataset (Gerig et al., 2018), further confirms the effectiveness and practicality of our approach in real-world scenarios.

2. Related Work

In this section, we introduce availability attacks and 3D point cloud attacks. More discussions on related work can be found in Appendix B.

Availability attacks. Privacy issues have been extensively studied in the field of privacy-preserving machine learning (Shokri & Shmatikov, 2015; Abadi et al., 2016; Shokri et al., 2017), including studies on availability attacks (Huang et al., 2020; Yu et al., 2022; Sandoval-Segura et al., 2022a). Availability attacks are a type of data poisoning, which allow the attacker to perturb the training dataset under a small norm restriction. These attacks aim to cause test-time errors while maintaining the semantic integrity and not affecting the normal usage by legitimate users. Huang et al. (2020) propose a bi-level error-minimizing approach to generate effective poisons, which is called unlearnable examples. Fowl et al. (2021) use stronger adversarial poisons to achieve availability attacks. Other availability attacks (Feng et al., 2019; Yuan & Wu, 2021; Yu et al., 2022; Sandoval-Segura et al., 2022b; Chen et al., 2023; He et al., 2023b) have also been proposed. However, the problem of data privacy and security of 3D deep learning is still unexplored. The data structure of 3D point clouds is inherently different from 2D images, posing a significant challenge in designing an effective availability attack method for 3D neural networks.

Attacks on 3D point cloud. Vulnerability of 3D point cloud classifiers has become a grave concern due to the popularity of 3D sensors in applications. Many works (Xiang et al., 2019; Cao et al., 2019; Miao et al., 2022) apply adversarial attacks to the 3D point cloud domain. Xiang et al. (2019) propose point generation attacks by adding a limited number

of synthetic points. Further research (Huang et al., 2022; Liu & Hu, 2022; Tsai et al., 2020; Zheng et al., 2019; Miao et al., 2022) has employed gradient-based point perturbation attacks. However, adversarial attacks are test-time evasion attacks. In contrast, our objective is to attack the model at train-time. The backdoor attack is another type of attack that poisons training data with a stealthy trigger pattern (Chen et al., 2017a). Some works (Li et al., 2021; Xiang et al., 2021; Gao et al., 2023) inject backdoors into 3D point clouds from a geometric perspective. However, the backdoor attack does not harm the model's performance on clean data. Thus, it is not a valid method for data protection. Different from these works, we introduce imperceptible perturbations into the training data, thereby reducing the model's generalization ability and achieving privacy protection.

3. Preliminaries

3.1. Problem Formulation

We formulate the problem of creating a clean-label poison in the context of 3D point cloud recognition. The poisoner has full access to the clean training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i = \{x_i^j\}_{j=1}^{n} \in \mathbb{R}^{n \times 3}$ represents the input point cloud and y_i is the ground-truth label, and is able to add perturbations δ to each sample, releasing the poisoned version $D_\delta = \{(x_i + \delta_i, y_i)\}_{i=1}^{N}$. The goal of the poisoner is to decrease the clean test accuracy of the point cloud model trained on D_δ, while ensuring that the perturbation is imperceptible and can escape any detection from the victim. To this end, the poisoner constrains perturbations δ by a certain budget ϵ, i.e., $\|\delta\|_p \leq \epsilon$, where $\|\cdot\|_p$ represents the ℓ_p norm. There are two representative availability attack methods.

Error-minimization (EM) (Huang et al., 2020) involves iteratively optimizing a bi-level min-min problem to minimize the loss of the network, leading to the generation of poisoning noises, referred to as unlearnable examples:

$$\min_{\theta} \min_{\delta} \mathbb{E}_{(x,y)\in D} \big[\mathcal{L}_{\text{cls}}(x + \delta, y; \theta)\big], s.t. \|\delta\|_p \leq \epsilon, \quad (1)$$

where θ represents the model parameters, and $\mathcal{L}_{\text{cls}}$ denotes the loss function. The outer minimization can imitate the training process, while the inner minimization can induce δ to have the property of minimizing the supervised loss. Due to this property, deep models will pay more attention to the easy-to-learn δ and ignore x.

Error-maximization (Adversarial Poisons, AP) (Fowl et al., 2021) generates poisoning noises by conducting stronger adversarial attacks on a pre-trained source classifier, referred to as adversarial poisons:

$$\max_{\delta} \mathbb{E}_{(x,y)\in D} \big[\mathcal{L}_{\text{cls}}(x + \delta, y; \theta^*)\big], s.t. \|\delta\|_p \leq \epsilon,$$
$$\theta^* \in \arg\min_{\theta} \mathbb{E}_{(x,y)\in D} \big[\mathcal{L}_{\text{cls}}(x, y; \theta)\big], \quad (2)$$

where θ^* represents the pre-trained model's parameters. AP poisons have non-robust features in the incorrect class, causing DNNs generalize poorly on clean examples.

3.2. 3D Point Cloud Availability Attack

However, in contrast to images which are fixed dimension grids, point clouds are represented as unordered and unstructured sets. As illustrated in Figure 1, the poisons constrained by ℓ_p norms have noticeable point outliers, which are visually perceptible, making ℓ_p norm less suitable (Miao et al., 2022). Instead, within 3D point clouds, *Chamfer distance* (Fan et al., 2017) often serves as an alternative to measure the distance of two point clouds, which is defined as:

$$\mathcal{D}_c(x, x') = \frac{1}{n}\sum_{x_j \in x} \min_{x_j' \in x'} \|x_j - x_j'\|_2^2 + \frac{1}{n'}\sum_{x_j' \in x'} \min_{x_j \in x} \|x_j' - x_j\|_2^2,$$

where n and n' are the number of points for x and x' respectively. The Chamfer distance finds the nearest original point for each adversarial point and averages all the distances among the nearest point pairs.

Owing to the presence of the minimum operator in the Chamfer distance, it is difficult to directly constrain the point cloud through projection onto a hypersphere, akin to the ℓ_p norm. In the context of 3D point cloud adversarial attacks (Xiang et al., 2019; Tsai et al., 2020; Wen et al., 2020), Chamfer distance is commonly employed as a regularization term for perceptual distance. Thus, EM can be augmented by incorporating a distance regularization term, thus straightforwardly extending to:

$$\min_{\theta} \min_{\delta} \mathbb{E}_{(x,y)\in D} \big[\mathcal{L}_{\text{cls}}(x + \delta, y; \theta) + \beta \cdot \mathcal{L}_{\text{dis}}(x + \delta, x)\big], \quad (3)$$

where β is a balancing hyperparameter between these two losses. We denote it as *Regularized Error-Minimization* (REG-EM). Similarly, AP also can be extended:

$$\max_{\delta} \mathbb{E}_{(x,y)\in D} \big[\mathcal{L}_{\text{cls}}(x + \delta, y; \theta^*) - \beta \cdot \mathcal{L}_{\text{dis}}(x + \delta, x)\big],$$
$$s.t. \quad \theta^* \in \arg\min_{\theta} \mathbb{E}_{(x,y)\in D} \big[\mathcal{L}_{\text{cls}}(x, y; \theta)\big]. \quad (4)$$

We denote it as *Regularized Adversarial Poisons* (REG-AP). However, in the next section, we will show that Eq. (3) and Eq. (4) exhibit significant degeneracy issues, rendering the poisoning ineffective.

4. Poisoning Degeneracy

While Eq. (3) and Eq. (4) straightforwardly extend the 2D poisoning methods to 3D point clouds under a distance regularization, this section demonstrates that such extension is susceptible to poisoning degeneracy, which makes the generated poisons weaker or even ineffective.

Attack degeneracy on 3D unlearnable examples. In bi-level optimization, the occurrence of degeneracy phenomena, such as catastrophic forgetting and mode collapse observed in GAN (Thanh-Tung & Tran, 2020), is likely if the optimization directions are not well-controlled. Unfortunately, the bi-level optimization process REG-EM is not exempted from this degeneracy. Theorem 4.1 demonstrates that, the final convergence point of Eq. (3) compels the poison $\delta \rightarrow 0$, rendering the poisoning ineffective.

Theorem 4.1. *(Proof in Appendix A) Assume that $\mathcal{L}_{\mathrm{cls}}$ and $\mathcal{L}_{\mathrm{dis}}$ are continuous, and the network's hypothesis space $\mathcal{H}_{\mathcal{F}}$ is compact. Let $D_\delta = \{(x_i + \delta_i, y_i)\}_{i=1}^N$ be the poisoned dataset of D. For simplicity, we denote $\underset{(x,y) \in D}{\mathbb{E}} [\mathcal{L}_{\mathrm{cls}}(x, y; \theta)]$ as $\mathcal{L}_{\mathrm{cls}}(D; \theta)$. Then, there exists an optimal point (δ^*, θ^*) for the bi-level optimization (3), and the optimal perturbation δ^* satisfies $\mathcal{L}_{\mathrm{dis}}(D, \delta^*) \leq \frac{1}{\beta} [\min_\theta \mathcal{L}_{\mathrm{cls}}(D; \theta) - \min_{\delta,\theta} \mathcal{L}_{\mathrm{cls}}(D_\delta; \theta)]$.*

Remark 4.2. Theorem 4.1 indicates that when the gap of $\min_\theta \mathcal{L}_{\mathrm{cls}}(D; \theta) - \min_{\delta,\theta} \mathcal{L}_{\mathrm{cls}}(D_\delta; \theta)$ is extremely small, the optimal perturbation δ^* is subject to a very limited distance constraint, significantly reducing the poisoning power.

Corollary 4.3. *(Proof in Appendix A) Let $\mathcal{L}_{\mathrm{cls}}(x, y; \theta) = \max(\max_{t \neq y} f_\theta(x)_t - f_\theta(x)_y, 0)$, where $f_\theta(\cdot)$ is the output of logits layer, and $\mathcal{L}_{\mathrm{dis}}$ be a metric function. If $\mathcal{H}_{\mathcal{F}}$ includes the function capable of correctly classifying D, then the equilibrium of the bi-level optimization (3) will degenerate to $\delta^* = 0$.*

Corollary 4.3 demonstrates that the optimal perturbation δ^* of Eq. (3) degenerates to 0 when $\mathcal{L}_{\mathrm{cls}}$ is constructed as suggested by Carlini & Wagner (2017). In practice, when $\mathcal{L}_{\mathrm{cls}}$ is cross-entropy, $\min_\theta \mathcal{L}_{\mathrm{cls}}(D; \theta)$ is nearly zero. Consequently, the gap in Theorem 4.1 is very small, compelling the degeneracy of poison δ. For REG-EM, i.e., Eq. (3), we vary the regularization strength β when generating poisons. The results in Figure 2(a) suggest that, despite changes in β, there is no significant impact on the Chamfer distance and test accuracy. This implies that the poison has already degenerated when distance regularization is applied.

Stronger attack becomes weaker for 3D adversarial poisons. AP aims to employ stronger adversarial examples for poison generation (Fowl et al., 2021). However, when extending AP to REG-AP, the addition of a distance regularization term weakens δ rather than strengthens it. This occurs because, during the iterative process of REG-AP, once the adversarial attack succeeds, the optimization focus in Eq. (4) shifts towards minimizing the regularization term, attempting to reduce δ. This leads to weaker adversarial attacks compared to AP, which, under the constraint of ℓ_p norm, continues to seek stronger adversarial samples even after a successful attack. Further detailed discussions will be deferred to Appendix E. This will significantly diminish the
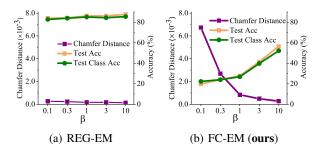

(a) REG-EM   (b) FC-EM (**ours**)

Figure 2. The effects of different distance regularization strength β under REG-EM and our FC-EM. Poisons crafted by REG-EM exhibit consistent Chamfer distance and test accuracy across different values of β. This implies that the poison has already degenerated when distance regularization is applied. In contrast, poisons crafted by FC-EM are more vulnerable to the balancing hyperparameter β, demonstrating resilience against poison degeneracy.

potency of the poisons, as demonstrated in the experiments in Section 6. Therefore, directly extending the AP process to 3D point clouds proves to be ineffective.

5. Method

As analyzed in Section 4, the inclusion of distance regularization in 3D point clouds would lead to degeneracy, significantly diminishing the efficacy. Addressing the challenges arising from the complexity introduced by the distance regularization term in bi-level optimization becomes crucial. In this section, we propose a novel method, **Feature Collision Error-Minimization (FC-EM)**, to mitigate this degeneracy and theoretically analyze its poisoning performance.

5.1. Feature Collision Error-Minimization

To generate potent poisons, a necessary condition is to disrupt the equilibrium provided in Theorem 4.1 and Corollary 4.3. To address this issue, our key insight is to employ a new loss for poison generation which is distinct from $\mathcal{L}_{\mathrm{cls}}$ in model optimization. By doing so, different update directions can prevent the degeneracy of bi-level optimization, thereby breaking the constraint provided in Theorem 4.1.

Specifically, due to the alternate iteration of bi-level optimization (3), we propose the use of a new loss $\mathcal{L}_{\mathrm{new}}$ during the optimization of δ. For clarity, let $\mathcal{L}_1(\theta, \delta) = \mathbb{E}_{(x,y) \in D} \mathcal{L}_{\mathrm{cls}}(x + \delta, y; \theta) + \beta \cdot \mathcal{L}_{\mathrm{dis}}(x + \delta, x)$ and $\mathcal{L}_2(\theta, \delta) = \mathbb{E}_{(x,y) \in D} \mathcal{L}_{\mathrm{new}}(x + \delta, y; \theta) + \beta \cdot \mathcal{L}_{\mathrm{dis}}(x + \delta, x)$. θ is updated by minimizing $\mathcal{L}_1$, while δ is updated by minimizing $\mathcal{L}_2$. The subsequent theorem guides us in constructing a new loss to mitigate degeneracy. The results indicate that when there exists a substantial gap between $\mathcal{L}_2(\theta^*, 0)$ and $\mathcal{L}_2(\theta^*, \delta^*)$, the ℓ_2 norm of δ^* is also significant, preventing degeneracy.

Theorem 5.1. *(Proof in Appendix A) Let $D = \{(x_i, y_i)\}_{i=1}^N$, δ_i be the perturbations corresponding to x_i. Denote the*

Algorithm 1 Feature Collision Error-Minimization Poisoning Attack (FC-EM)

Input: A 3D point cloud training dataset $D = \{(x_i, y_i)\}_{i=1}^N$. Total epoch T. Batch size N_B. Distance loss $\mathcal{L}_{\text{dis}}$ and regularization strength β. Feature collision loss $\mathcal{L}_{\text{fc}}$ and temperature t. Classifier parameters α_θ and T_θ. Attack parameters α_δ, T_δ and T_a.

Output: Poisoned dataset $D_\delta = \{(x_i + \delta_i, y_i)\}_{i=1}^N$

Initialize: $\delta_i \leftarrow 0, i = 1, 2, \cdots, N$

for $t = 1, \cdots, T$ **do**

  for $t_\theta = 1, \cdots, T_\theta$ **do**

    Sample a mini batch $B = \{(x_{b_j}, y_{b_j})\}_{j=1}^{N_B}$.

    $\theta \leftarrow \theta - \alpha_\theta \cdot \nabla_\theta \mathbb{E}_{(x_{b_j}, y_{b_j}) \in B}\big[\mathcal{L}_{\text{cls}}(x_{b_j} + \delta_{b_j}, y_{b_j}; \theta) + \beta \cdot \mathcal{L}_{\text{dis}}(x_{b_j} + \delta_{b_j}, x_{b_j})\big]$

  for $t_\theta = 1, \cdots, T_\delta$ **do**

    Sample a mini batch $B = \{(x_{b_j}, y_{b_j})\}_{j=1}^{N_B}$.

    for $t_a = 1, \cdots, T_a$ **do**

      Compute class-wise feature collision loss $\mathcal{L}_{\text{fc}}$.

      $\delta_{b_j} \leftarrow \delta_{b_j} - \alpha_\delta \cdot \nabla_{\delta_{b_j}} \mathbb{E}_{(x_{b_j}, y_{b_j}) \in B}\big[\mathcal{L}_{\text{fc}}(x_{b_j} + \delta_{b_j}, y_{b_j}; \theta, t) + \beta \cdot \mathcal{L}_{\text{dis}}(x_{b_j} + \delta_{b_j}, x_{b_j})\big]$

---

equilibrium for bi-level optimization $\min_{\theta,\delta}\{\mathcal{L}_1, \mathcal{L}_2\}$ *as* (θ^*, δ^*). *Assume that* $\mathcal{L}_{\text{new}}$ *is L-lipschitz with respect to* δ *under* ℓ_2 *norm,* $\mathcal{L}_{\text{dis}}$ *is Chamfer distance, and for each point* x_i^j, *the* ℓ_2 *norm of perturbation* δ_i^j *is bound by* $\frac{1}{2}\min_{j,k}\|x_i^j - x_i^k\|_2$. *Then it holds* $\|\delta_i^*\|_2 \geq \frac{1}{4\beta}(L - \sqrt{L^2 - 8\beta\Delta})$, *where* $\Delta = \mathcal{L}_2(\theta^*, 0) - \mathcal{L}_2(\theta^*, \delta^*)$.

Remark 5.2. In light of Theorem 5.1, when the gap Δ is substantial, the optimal perturbation δ_i under the ℓ_2 norm also increases. To maintain a substantial gap Δ and disrupt the equilibrium in bi-level optimization, the new loss $\mathcal{L}_{\text{new}}$ should not converge to the optimal value θ^* under $\mathcal{L}_{\text{cls}}$.

Theorem 5.1 guides us to construct a new loss with distinct optimization directions. Effective availability attacks often require creating shortcuts (Yu et al., 2022; Sandoval-Segura et al., 2022a) that are so simplistic, being linearly separable, that deep models tend to rely on them for predictions, neglecting genuine features. We consider generating poisons containing shortcuts in the feature space rather than using $\mathcal{L}_{\text{cls}}$ in the logit space to induce different update directions. Motivated by Chen et al. (2020), we propose a novel class-wise feature similarity loss $\mathcal{L}_{\text{fc}}$ with its j-th dimension:

$$\mathcal{L}_{\text{fc}}(x + \delta, y; \theta, t)_j = -\log$$
$$\frac{\sum_{k \in B} \mathbb{I}(y_{b_k} = y_{b_j}) \exp(s(g(x_{b_j} + \delta_{b_j}), g(x_{b_k} + \delta_{b_k}))/t)}{\sum_{k \in B} \exp(s(g(x_{b_j} + \delta_{b_j}), g(x_{b_k} + \delta_{b_k}))/t)}, \quad (5)$$

where $B = \{x_{b_j}, y_{b_j}\}_{j=1}^{N_B}$ is the mini-batch, g represents the feature extractor of the network f_θ, $s(\cdot, \cdot) = \frac{\langle\cdot,\cdot\rangle}{\|\cdot\|_2\|\cdot\|_2}$ denotes the cosine similarity between two vectors, and t is the temperature of the loss. $\mathcal{L}_{\text{fc}}$ represents inter-class feature similarity as a surrogate loss that reflects the clustering process, generating shortcuts for 3D point clouds. Specifically,
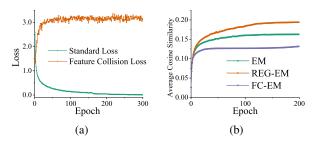

Figure 3. **(a):** Epoch-loss curves of cross-entropy loss and feature collision loss under standard training. The feature collision loss fails to converge and even increases. This implies that they optimize towards the different direction. **(b):** Average cosine similarity between the rows of last layer weight matrix. FC-EM yields smaller cosine similarities compared to EM and REG-EM.

at each step i, we iteratively update δ and θ:

$$\Delta_\theta^i = \arg\min_\theta \mathbb{E}_{(x,y) \in D}\big[\mathcal{L}_{\text{cls}}(x + \delta^{i-1}, y; \theta) + \beta \cdot \mathcal{L}_{\text{dis}}(x + \delta^{i-1}, x)\big],$$

$$\Delta_\delta^i = \arg\min_\delta \mathbb{E}_{(x,y) \in D}\big[\mathcal{L}_{\text{fc}}(x + \delta, y; \theta^i, t) + \beta \cdot \mathcal{L}_{\text{dis}}(x + \delta, x)\big],$$

$$\theta^i = \theta^{i-1} + \Delta_\theta^i, \delta^i = \delta^{i-1} + \Delta_\delta^i. \quad (6)$$

It is noteworthy that $\mathcal{L}_{\text{fc}}$ modifies logits by replacing softmax and negative log-likelihood operations with cosine similarity. In the numerator of $\mathcal{L}_{\text{fc}}$, we incorporate all samples from the same class, not just the logits corresponding to true labels. Therefore, the objective of minimizing Eq. (5) is to increase feature similarity within the same class while decreasing feature similarity across different classes, which we referred to as *feature collision* by similarity. We create shortcuts in the feature space, causing the network to focus more on the intra/inter-class features of δ, ignoring those of x. The detailed procedure of FC-EM is provided in Algorithm 1.

As shown in Figure 3(a), when vanilla training a network using the cross entropy loss $\mathcal{L}_{\text{cls}}$, the feature collision loss $\mathcal{L}_{\text{fc}}$ fails to converge and even increases. In other words, while minimizing $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{fc}}$ both aim to create shortcuts, they do not optimize towards the same direction. This discrepancy leads to a larger $\mathcal{L}_2(\theta^*, 0)$, as derived in Theorem 5.1, resulting in a larger Δ and subsequently causing a larger δ^* to break the degeneracy under distance regularization. In comparison to REG-EM depicted in Figure 2(a), Figure 2(b) reveals that FC-EM poisons are more vulnerable to the distance regularization strength β, demonstrating resilience against poison degeneracy.

5.2. Theoretical Analysis

To assess the suitability of our proposed FC-EM attack for 3D point clouds, we provide a theoretical analysis that demonstrates the effectiveness of FC-EM attack. Yu et al. (2022); Zhu et al. (2023) find both experimentally and theoretically that the effectiveness of availability attacks is
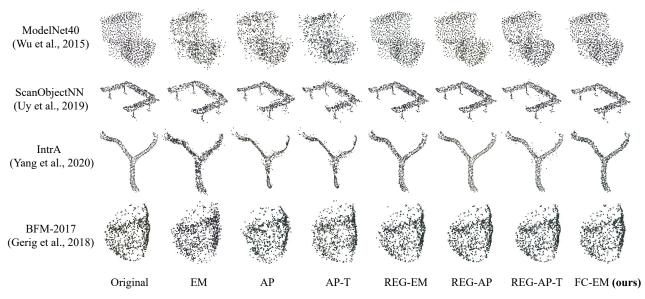
Figure 4. Qualitative visualization results of baseline methods and our FC-EM. Poisons generated by EM, AP, AP-T and REG-AP-T exhibit conspicuous outliers, thus lacking imperceptibility. Although REG-EM and REG-AP successfully achieve imperceptibility, they fail to reduce model's accuracy on test data. In contrast, our FC-EM approach not only demonstrates enhanced naturalness and imperceptibility but also effectively reduce model's generalization ability.

closely tied to linear separability. The reason why current availability attacks work may stem from that the imperceptible poisons create shortcuts. They are almost linearly separable, that deep models would overwhelmingly rely on spurious features for predictions. Below, we demonstrate that FC-EM attack exhibits stronger linear separability compared to traditional EM and REG-EM attacks. Firstly, we establish a correlation between the linear separability of the poisoned dataset and the cosine similarity between the last layer weights.

Theorem 5.3. *(Proof in Appendix A) Let α be the loss minimum of dataset D under all linear classifier, the weight matrix of a linear classifier be denoted as W, where ℓ_2 norm of each row is normalized to $\sqrt{d}$. For any two rows of W, their cosine similarity does not exceed γ. The loss function is $\mathcal{L}_{cls}(x, y; \theta) = \max(\max_{t \neq y} f_\theta(x)_t - f_\theta(x)_y, 0)$. Then, there exist poisons $\{\delta_i\}_{i=1}^N$ satisfying $\mathbb{E}_{(x_i,y_i) \in D} \mathcal{D}_c(x_i, x_i + \delta_i) \leq O(\frac{\alpha}{(1-\gamma)^2})$, such that the poisoned dataset is linearly separable.*

Remark 5.4. Theorem 5.3 implies that as the cosine similarity γ decreases, the upper bound on the Chamfer distance of existing effective poisons becomes smaller. Consequently, to achieve linear separability in poisoned datasets, models with lower cosine similarity require less Chamfer distance as a poisoning budget, making poisons more effective.

FC-EM optimizes the feature collision loss by enhancing intra-class cosine similarity and reducing inter-class cosine similarity simultaneously. The weight matrix W of the final layer linear classifier has each row representing a distinct pattern match for a specified class, tending to align with the corresponding features. Consequently, due to the consistency in cosine similarity between inter-class features and different rows in W, FC-EM tends to diminish the cosine similarity of rows. In fact, as illustrated in Figure 3(b), FC-EM indeed yields smaller cosine similarities compared to EM and REG-EM. According to Theorem 5.3, this implies that if the Chamfer distance is constrained, FC-EM exhibits stronger linear separability than EM and REG-EM. Therefore, it possesses stronger poison, resulting in lower clean test accuracy of the victim model.

6. Experiments

6.1. Experimental Setup

Dataset. In our experiments, we employ several datasets for evaluation, including the generated point clouds dataset, ModelNet40 (Wu et al., 2015), the scanned point clouds dataset, ScanObjectNN (Uy et al., 2019), the 3D intracranial aneurysm medical dataset, IntrA (Yang et al., 2020), and the BFM-2017 generated face dataset (Gerig et al., 2018). Details of these datasets are provided in Appendix C.

Victim Models. We select three commonly used point cloud classification networks as our victim models, including PointNet (Qi et al., 2017a), PointNet++ (Qi et al., 2017b) and DGCNN (Wang et al., 2019).

Evaluation metrics. To quantitatively assess the effectiveness of various attacks, we train victim models on the poisoned training dataset, and evaluate the accuracy (Acc) on the clean test set. Besides, to measure the imperceptibility,

Table 1. Quantitative results of baselines and our FC-EM on ModelNet40. Our FC-EM achieves the lowest test accuracy (Acc) and exhibits suitable imperceptibility (measured by Chamfer distance ($\mathcal{D}_c$) and Hausdorff distance ($\mathcal{D}_h$)), causing the strongest availability attack.

Poison	PointNet			PointNet++			DGCNN		
	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$
Clean	90.70%	-	-	93.06%	-	-	92.64%	-	-
EM	82.98%	29.7	14.1	82.41%	27.5	11.8	80.72%	27.2	11.6
AP	69.37%	27.3	17.0	89.26%	25.6	11.1	78.04%	29.2	10.9
AP-T	61.95%	26.7	16.8	56.33%	22.8	9.7	57.45%	19.6	8.7
REG-EM	86.22%	1.7	7.8	87.52%	13.2	8.2	85.21%	6.5	5.4
REG-AP	75.32%	5.5	11.5	88.11%	5.4	3.4	88.49%	7.8	5.2
REG-AP-T	72.61%	5.1	11.5	77.88%	3.0	5.3	75.97%	3.3	5.1
FC-EM (**ours**)	**27.67%**	8.4	13.5	**31.12%**	8.9	9.0	**32.46%**	8.1	8.0

Table 2. Quantitative results of baselines and our FC-EM on ScanObjectNN. Our FC-EM obtains the lowest test accuracy and gains good imperceptibility (measured by Chamfer distance ($\mathcal{D}_c$) and Hausdorff distance ($\mathcal{D}_h$)), outperforming all other existing availability attacks.

Poison	PointNet			PointNet++			DGCNN		
	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$
Clean	74.87%	-	-	85.54%	-	-	81.07%	-	-
EM	63.68%	32.5	12.7	66.78%	22.6	7.9	46.82%	22.3	8.0
AP	49.91%	14.4	14.5	54.04%	11.2	7.0	32.01%	13.7	8.5
AP-T	49.74%	13.5	13.9	34.77%	10.4	9.3	26.68%	13.1	8.2
REG-EM	69.71%	2.3	1.8	72.46%	4.0	2.8	65.23%	4.0	2.8
REG-AP	67.13%	2.0	3.9	73.49%	4.7	3.9	56.63%	9.5	7.7
REG-AP-T	50.60%	3.2	8.1	52.32%	2.3	4.0	43.72%	5.9	7.9
FC-EM (**ours**)	**33.21%**	4.6	10.4	**27.71%**	8.7	9.2	**23.41%**	4.5	6.8

we use the Chamfer distance ($\mathcal{D}_c$) and Hausdorff distance (Rockafellar & Wets, 2009) ($\mathcal{D}_h$) as evaluation metrics.

Implementation details. We use PyTorch learning rate scheduler, ReduceLROnPlateau for both poisoning and evaluation process, with initial learning rate 10^{-3}. Except for above baseline methods, we also evaluate the targeted version of AP and REG-AP, namely AP-T and REG-AP-T. The poisoning epoch for EM, REG-EM and FC-EM is set to 200, and for AP(-T), REG-AP(-T) is set to 100. The evaluation epoch is set to 200. The coefficient of regularization term β is set to 1.0 originally. Across all datasets, we uniformly sample 1,024 points for attacks and classification. For further details, including more ablation studies and experiments, please refer to Appendices C and D.

6.2. Main Results

We compare the proposed FC-EM with with various baselines, including the distance regularization methods, REG-EM and REG-AP(-T), and the ℓ_∞-norm restraint methods, EM and AP(-T). Table 1 shows the results on ModelNet40, our FC-EM achieves the lowest test accuracy (Acc) compared with all other baselines and exhibits suitable imperceptibility (measured by Chamfer distance $\mathcal{D}_c$) and Hausdorff distance $\mathcal{D}_h$). Consequently, it exhibits the worst generalization and causes strongest poison. Visualization on ModelNet40 under different methods is depicted in the first row of Figure 4. Our FC-EM demonstrates great imperceptibility. This is because FC-EM creates poisons under distance

regularization rather than simple ℓ_p norm restriction, leading poisons to focus more on the structural aspects of point clouds. In contrast, the ℓ_∞-norm restraint methods, EM and AP(-T) exhibit many outliers and irregular deformations, simultaneously inducing high $\mathcal{D}_c$ and $\mathcal{D}_h$. On the other hand, although REG-EM and REG-AP successfully achieve imperceptibility, they fail to reduce model's accuracy on test data, remaining Acc even larger than 70%. In summary, our FC-EM achieves better availability attack on both performance and imperceptibility.

Evaluation results and corresponding visualizations on ScanObjectNN are provided in Table 2 and the second row of Figure 4, respectively. Results on ScanObjectNN reveal similar trends with those on ModelNet40. FC-EM obtains lowest test accuracy and gains good imperceptibility. Thus it outperforms all other baselines. More discussions are deferred to Appendix D.1.

6.3. Transferability between Models

We further verify the transferability of our FC-EM. Results presented in Table 4 shows the decent transferability of FC-EM, as all the test accuracy remains consistently below 40%. Our FC-EM is stable, displaying small sensitivity in both the source and victim models. This implies that FC-EM poison appears to manifest as intrinsic availability defects within the dataset itself rather than arising from shortcomings in specific models.

Table 3. Quantitative results of baselines and our FC-EM on 3D intracranial aneurysm medical dataset IntrA. Our FC-EM attains the lowest F1-score while keeping imperceptibility (measured by $\mathcal{D}_c$ and $\mathcal{D}_h$), obtaining strongest availability attack.

	PointNet			PointNet++			DGCNN		
Poison	F1-score	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$	F1-score	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$	F1-score	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$
Clean	0.667	-	-	0.821	-	-	0.741	-	-
EM	0.458	36.4	16.4	0.391	24.4	8.7	0.417	29.5	11.4
AP	0.360	17.9	16.2	0.288	15.4	6.7	0.175	15.0	9.6
AP-T	0.324	17.9	16.1	0.288	15.4	6.7	0.179	15.1	9.7
REG-EM	0.471	2.1	2.1	0.545	4.1	3.3	0.485	4.0	2.9
REG-AP	0.585	0.9	1.2	0.571	7.6	3.4	0.485	10.1	5.4
REG-AP-T	0.410	1.6	4.5	0.747	0.4	1.5	0.207	0.7	2.9
FC-EM (**ours**)	**0.285**	2.0	2.0	**0.000**	0.9	2.0	**0.000**	0.9	2.0

Table 4. Quantitative results of model's transferability of our FC-EM on ModelNet40. Our FC-EM exhibits good transferability, with all the test accuracy remaining consistently below 40%.

Eva./Gen. Model	PointNet	PointNet++	DGCNN
PointNet	27.67%	38.01%	30.83%
PointNet++	35.05%	31.12%	33.95%
DGCNN	33.31%	37.88%	32.50%

Table 5. Quantitative results of the test accuracy (Acc) under different defense methods. Our FC-EM outperforms all other availability attacks in all defense methods.

Acc(%)	ST	PGD	TRADES	Mixup	Rsmix	IF-D
Clean	90.70	82.62	78.77	89.95	89.26	87.12
EM	82.98	65.48	70.22	84.00	82.90	80.79
AP	69.37	78.97	76.94	76.05	71.27	75.81
AP-T	61.95	75.97	72.97	67.02	63.70	76.05
REG-EM	86.22	82.09	78.44	89.14	87.76	86.06
REG-AP	75.32	74.35	77.39	82.58	80.88	84.93
REG-AP-T	72.61	77.63	74.67	73.95	70.83	82.70
FC-EM (**ours**)	**27.67**	**49.59**	**52.80**	**65.32**	**51.86**	**75.73**

6.4. Performance under Defense

We evaluate various availability attacks on a range of defense methods, including AT-based approaches, PGD (Madry et al., 2018) and TRADES (Zhang et al., 2019), data augmentation, Mixup (Zhang et al., 2018) and Rsmix (Lee et al., 2021), and a defense tailored for 3D point clouds, IF-Defense (Wu et al., 2020). The results provided in Table 5 demonstrates that FC-EM outperforms all other availability attacks in all defense methods.

6.5. Results on Real-World Datasets

Results on 3D medical dataset. We conduct evaluation on the medical dataset IntrA (Yang et al., 2020), a 3D intracranial aneurysm dataset framed as binary classification problem for distinguishing aneurysms from healthy vessel segments. The evaluation is based on the F1-score. The results provided in Table 3 reveal that FC-EM attains the lowest F1-score while keeping distance small, and even reaches

Table 6. Quantitative results of baselines and our FC-EM on the BFM-2017 generated 3D face dataset. Our FC-EM attains the lowest test accuracy, inducing strongest availability attack.

Poison	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$
Clean	98.1%	-	-
EM	53.5%	23.8	11.4
AP	76.1%	10.1	11.9
AP-T	46.9%	7.3	10.8
REG-EM	50.9%	4.5	5.6
REG-AP	79.3%	2.0	3.7
REG-AP-T	91.1%	0.8	1.2
FC-EM (**ours**)	**11.0%**	8.5	11.3

to zero F1-score when utilizing PointNet++ and DGCNN. In such scenarios, victim models' predictions collapse entirely, consistently yielding negative outputs regardless of the input, rendering the F1-score to be zero. IntrA is visualized in the third row of Figure 4. It illustrates that both regularization methods and FC-EM maintain the structural integrity of vessels. In contrast, ℓ_∞ methods largely disrupt the structure, introducing numerous outliers. Consequently, FC-EM not only preserves the imperceptibility, but also obtains strongest availability attack.

Results on 3D face dataset. We employ Basel Face Model 2017 (Gerig et al., 2018) as the source face models, generating 3D point clouds from facial scans. Details regarding the generation process are outlined in Appendix C. The poison results on PointNet and the visualizations are presented in Table 6 and the fourth row of Figure 4, respectively. Notably, FC-EM and regularization methods display superior face contour. In contrast, ℓ_∞ methods severely disrupt the contour and have more outliers. Moreover, compared to regularization methods, FC-EM induces stronger attack, prompting a substantial 40% decrease in the test accuracy, succinctly showcasing its superiority. More ablation studies and experiments can be found in Appendix D.

7. Conclusion

In this paper, we propose a Feature Collision Error-Minimization (FC-EM) method for effective 3D availability

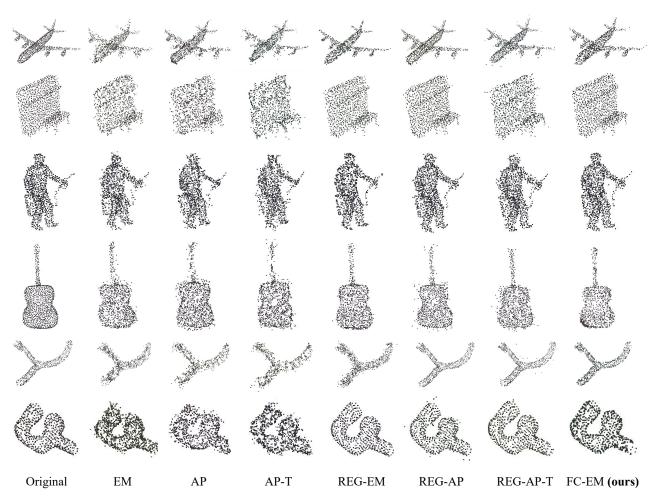| Original | EM | AP | AP-T | REG-EM | REG-AP | REG-AP-T | FC-EM **(ours)** |

Figure 5. More qualitative visualization results of baseline methods and our FC-EM. Our FC-EM approach demonstrates enhanced naturalness and imperceptibility.

attacks to safeguard data privacy. Theoretically, directly extending 2D availability attacks to 3D under distance regularization terms can lead to degeneracy, resulting in weaker poisons, due to optimization directions out of control in the bi-level optimization. FC-EM establishes additional shortcuts in the feature space, inducing diverse update directions to prevent degeneracy. We conduct a theoretical analysis of the effectiveness of FC-EM attack, demonstrating that FC-EM exhibits stronger linear separability, consequently possessing stronger poison. Extensive experiments consistently validate the superiority and practicality of FC-EM.

Limitations and future work. While availability attacks are intended to prevent the unauthorized use of data, malicious entities can exploit availability attacks to compromise 3D deep models. Therefore, defensive strategy against 3D availability attacks need to be developed to safeguard models from corruption. Although our FC-EM method demonstrates strong attack ability and decent imperceptibility, it still exists some sparsity disparity, with some clustering phenomena as observed in Figure 4. Consequently, addressing

the challenge of availability attacks on 3D point clouds with enhanced imperceptibility remains a promising direction for future research.

Impact Statement

Data privacy and security issues in 3D deep learning is a severe problem towards safe and reliable 3D perception. Our work proposes an effective method to solve this issue, aiming to thwart unauthorized deep models from illegitimately learning data, which does not raise negative social impact. We aspire for our work to serve as a baseline for advancing research in the protection of privacy data within the 3D domain.

Acknowledgement

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Burt, C. Facial biometrics training dataset leads to bipa lawsuits against amazon, alphabet and microsoft, 2020. URL https://reurl.cc/dV4rD8.

Cao, Y., Xiao, C., Yang, D., Fang, J., Yang, R., Liu, M., and Li, B. Adversarial objects against lidar-based autonomous driving systems. *arXiv preprint arXiv:1907.05418*, 2019.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security & Privacy*, pp. 39–57. IEEE, 2017.

Chen, S., Yuan, G., Cheng, X., Gong, Y., Qin, M., Wang, Y., and Huang, X. Self-ensemble protection: Training checkpoints are good data protectors. In *The Eleventh International Conference on Learning Representations*, 2023.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017a.

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

Conklin, A. Facebook agrees to pay record $650m to settle facial recognition lawsuit, 2020. URL https://reurl.cc/Gd2kev.

Das, S. Private uk health data donated for medical research shared with insurance companies, 2023. URL https://reurl.cc/WRODdy.

Dolatabadi, H. M., Erfani, S., and Leckie, C. The devil's advocate: Shattering the illusion of unexploitable data using diffusion models. *arXiv preprint arXiv:2303.08500*, 2023.

Fan, H., Su, H., and Guibas, L. J. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 605–613, 2017.

Feng, J., Cai, Q.-Z., and Zhou, Z.-H. Learning to confuse: generating training time adversarial data with autoencoder. *Advances in Neural Information Processing Systems*, 32, 2019.

Fowl, L., Goldblum, M., Chiang, P.-y., Geiping, J., Czaja, W., and Goldstein, T. Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*, 34:30339–30351, 2021.

Fu, S., He, F., Liu, Y., Shen, L., and Tao, D. Robust unlearnable examples: Protecting data against adversarial learning. In *International Conference on Learning Representations*, 2022.

Gao, K., Bai, J., Wu, B., Ya, M., and Xia, S.-T. Imperceptible and robust backdoor attack in 3d point cloud. *IEEE Transactions on Information Forensics and Security*, 19:1267–1282, 2023.

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., and Vetter, T. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 75–82. IEEE, 2018.

He, B., Liu, J., Li, Y., Liang, S., Li, J., Jia, X., and Cao, X. Generating transferable 3d adversarial point cloud via random perturbation factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 764–772, 2023a.

He, H., Zha, K., and Katabi, D. Indiscriminate poisoning attacks on unsupervised contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022.

He, P., Xu, H., Ren, J., Cui, Y., Liu, H., Aggarwal, C. C., and Tang, J. Sharpness-aware data poisoning attack. *arXiv preprint arXiv:2305.14851*, 2023b.

Huang, H., Ma, X., Erfani, S. M., Bailey, J., and Wang, Y. Unlearnable examples: Making personal data unexploitable. In *International Conference on Learning Representations*, 2020.

Huang, Q., Dong, X., Chen, D., Zhou, H., Zhang, W., and Yu, N. Shape-invariant 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15335–15344, 2022.

Jiang, W., Diao, Y., Wang, H., Sun, J., Wang, M., and Hong, R. Unlearnable examples give a false sense of security: Piercing through unexploitable data with learnable examples. *arXiv preprint arXiv:2305.09241*, 2023.

Lee, D., Lee, J., Lee, J., Lee, H., Lee, M., Woo, S., and Lee, S. Regularization strategy for point cloud via rigidly

mixed sample. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15900–15909, 2021.

Li, X., Chen, Z., Zhao, Y., Tong, Z., Zhao, Y., Lim, A., and Zhou, J. T. Pointba: Towards backdoor attacks in 3d point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16492–16501, 2021.

Li, X., Liu, M., and Gao, S. Make text unlearnable: Exploiting effective patterns to protect personal data. In *The Third Workshop on Trustworthy Natural Language Processing*, pp. 249, 2023.

Liu, D. and Hu, W. Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Liu, Z., Zhao, Z., and Larson, M. Image shortcut squeezing: Countering perturbative availability poisons with compression. In *International conference on machine learning*, 2023.

Lou, T., Jia, X., Gu, J., Liu, L., Liang, S., He, B., and Cao, X. Hide in thicket: Generating imperceptible and rational adversarial perturbations on 3d point clouds. *arXiv preprint arXiv:2403.05247*, 2024.

Ma, X., Qin, C., You, H., Ran, H., and Fu, Y. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022.

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

Malihi, S., Valadan Zoej, M., Hahn, M., Mokhtarzade, M., and Arefi, H. 3d building reconstruction using dense photogrammetric point cloud. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41:71–74, 2016.

Miao, Y., Dong, Y., Zhu, J., and Gao, X.-S. Isometric 3d adversarial examples in the physical world. In *Advances in Neural Information Processing Systems*, 2022.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Qin, T., Gao, X., Zhao, J., Ye, K., and Xu, C.-Z. Learning the unlearnable: Adversarial augmentations suppress unlearnable example attacks. *arXiv preprint arXiv:2303.15127*, 2023.

Ren, J., Xu, H., Wan, Y., Ma, X., Sun, L., and Tang, J. Transferable unlearnable examples. In *The Eleventh International Conference on Learning Representations*, 2022.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Sadasivan, V. S., Soltanolkotabi, M., and Feizi, S. Cuda: Convolution-based unlearnable datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3862–3871, 2023.

Sandoval-Segura, P., Singla, V., Fowl, L., Geiping, J., Goldblum, M., Jacobs, D., and Goldstein, T. Poisons that are learned faster are more effective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 198–205, 2022a.

Sandoval-Segura, P., Singla, V., Geiping, J., Goldblum, M., Goldstein, T., and Jacobs, D. Autoregressive perturbations for data poisoning. *Advances in Neural Information Processing Systems*, 35:27374–27386, 2022b.

Sandoval-Segura, P., Singla, V., Geiping, J., Goldblum, M., and Goldstein, T. What can we learn from unlearnable datasets? *arXiv preprint arXiv:2305.19254*, 2023.

Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Taha, A. A. and Hanbury, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.

Tao, L., Feng, L., Yi, J., Huang, S.-J., and Chen, S. Better safe than sorry: Preventing delusive adversaries with adversarial training. *Advances in Neural Information Processing Systems*, 34:16209–16225, 2021.

Tao, L., Feng, L., Wei, H., Yi, J., Huang, S.-J., and Chen, S. Can adversarial training be manipulated by non-robust features? *Advances in Neural Information Processing Systems*, 35:26504–26518, 2022.

Thanh-Tung, H. and Tran, T. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pp. 1–10. IEEE, 2020.

Tsai, T., Yang, K., Ho, T.-Y., and Jin, Y. Robust adversarial objects against deep learning models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 954–962, 2020.

Uy, M. A., Pham, Q.-H., Hua, B.-S., Nguyen, T., and Yeung, S.-K. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1588–1597, 2019.

Vincent, J. Google accused of inappropriate access to medical data in potential class-action lawsuit, 2019. URL https://reurl.cc/bzK69v.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38 (5):1–12, 2019.

Wang, Y., Zhu, Y., and Gao, X.-S. Efficient availability attacks against supervised and contrastive learning simultaneously. *arXiv preprint arXiv:2402.04010*, 2024.

Wen, R., Zhao, Z., Liu, Z., Backes, M., Wang, T., and Zhang, Y. Is adversarial training really a silver bullet for mitigating data poisoning? In *International Conference on Learning Representations*, 2023.

Wen, Y., Lin, J., Chen, K., Chen, C. P., and Jia, K. Geometry-aware generation of adversarial point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Wu, S., Chen, S., Xie, C., and Huang, X. One-pixel shortcut: On the learning preference of deep neural networks. In *The Eleventh International Conference on Learning Representations*, 2022.

Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 803–814, 2023.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Wu, Z., Duan, Y., Wang, H., Fan, Q., and Guibas, L. J. If-defense: 3d adversarial point cloud defense via implicit function based restoration. *arXiv preprint arXiv:2010.05272*, 2020.

Xiang, C., Qi, C. R., and Li, B. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9136–9144, 2019.

Xiang, Z., Miller, D. J., Chen, S., Li, X., and Kesidis, G. A backdoor attack against 3d point cloud classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7597–7607, 2021.

Xu, M., Ding, R., Zhao, H., and Qi, X. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3173–3182, 2021a.

Xu, M., Zhang, J., Zhou, Z., Xu, M., Qi, X., and Qiao, Y. Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3056–3064, 2021b.

Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., and Lin, D. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.

Yang, X., Xia, D., Kin, T., and Igarashi, T. Intra: 3d intracranial aneurysm dataset for deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2656–2666, 2020.

Yu, D., Zhang, H., Chen, W., Yin, J., and Liu, T.-Y. Availability attacks create shortcuts. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2367–2376, 2022.

Yuan, C.-H. and Wu, S.-H. Neural tangent generalization attacks. In *International Conference on Machine Learning*, pp. 12230–12240. PMLR, 2021.

Yue, X., Wu, B., Seshia, S. A., Keutzer, K., and Sangiovanni-Vincentelli, A. L. A lidar point cloud generator: from a virtual world to autonomous driving. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pp. 458–464, 2018.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

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

Zhang, J., Chen, L., Ouyang, B., Liu, B., Zhu, J., Chen, Y., Meng, Y., and Wu, D. Pointcutmix: Regularization

strategy for point cloud classification. *Neurocomputing*, 505:58–67, 2022a.

Zhang, J., Chen, L., Liu, B., Ouyang, B., Xie, Q., Zhu, J., Li, W., and Meng, Y. 3d adversarial attacks beyond point cloud. *Information Sciences*, 633:491–503, 2023a.

Zhang, J., Ma, X., Yi, Q., Sang, J., Jiang, Y.-G., Wang, Y., and Xu, C. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3984–3993, 2023b.

Zhang, Z., Da, F., and Yu, Y. Learning directly from synthetic point clouds for "in-the-wild" 3d face recognition. *Pattern Recognition*, 123:108394, 2022b.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., and Koltun, V. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

Zheng, T., Chen, C., Yuan, J., Li, B., and Ren, K. Pointcloud saliency maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1598–1606, 2019.

Zhou, J., Wang, J., Ma, B., Liu, Y.-S., Huang, T., and Wang, X. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023.

Zhou, S. and Xiao, S. 3d face recognition: a survey. *Human-centric Computing and Information Sciences*, 8(1):1–27, 2018.

Zhu, Y., Yu, L., and Gao, X.-S. Detection and defense of unlearnable examples. *arXiv preprint arXiv:2312.08898*, 2023.

A. Theorems and Proofs

Theorem A.1 (Restate of Theorem 5.1). *Let $D = \{(x_i, y_i)\}_{i=1}^N$, δ_i be the perturbations corresponding to x_i. Denote the equilibrium for bi-level optimization $\min_{\theta,\delta}\{\mathcal{L}_1, \mathcal{L}_2\}$ as (θ^*, δ^*). Assume that $\mathcal{L}_{new}$ is L-lipschitz with respect to δ under ℓ_2 norm, $\mathcal{L}_{dis}$ is Chamfer distance, and for each point x_i^j, the ℓ_2 norm of perturbation δ_i^j is bound by $\frac{1}{2}\min_{j,k}\|x_i^j - x_i^k\|_2$. Then it holds $\|\delta_i^*\|_2 \geq \frac{1}{4\beta}(L - \sqrt{L^2 - 8\beta\Delta})$, where $\Delta = \mathcal{L}_2(\theta^*, 0) - \mathcal{L}_2(\theta^*, \delta^*)$.*

Proof. For any δ, due to the L-lipschitz property, it has

$$\mathcal{L}_2(\theta^*, \delta) - \mathcal{L}_2(\theta^*, 0) = \frac{1}{N}\sum_{j=1}^N [\mathcal{L}_{new}(x_i + \delta_i, y_i, \theta^*) - \mathcal{L}_{new}(x_i + 0, y_i, \theta^*)] + \frac{\beta}{N}\sum_{j=1}^N \mathcal{L}_{dis}(x_i + \delta_i, x_i)$$

$$\geq -\frac{L}{N}\sum_{j=1}^N \|\delta_i\|_2 + \frac{\beta}{N}\sum_{j=1}^N \mathcal{L}_{dis}(x_i + \delta_i, x_i).$$

Because for each point cloud x_i^j of x_i, $\|\delta_i^j\|_2 \leq \frac{1}{2}\min_{j,k}\|x_i^j - x_i^k\|_2$, it holds that

$$\mathcal{L}_{dis}(x_i + \delta_i, x_i) = 2\|\delta_i\|_2^2.$$

Then there must exists $i \in [N]$, such that

$$\mathcal{L}_2(\theta^*, \delta) - \mathcal{L}_2(\theta^*, \delta^*) \geq -L\|\delta_i\|_2 + \Delta + 2\beta\|\delta_i\|_2^2.$$

Let $\delta = \delta^*$, one must have

$$L\|\delta_i^*\|_2 - 2\beta\|\delta_i^*\|_2^2 \geq \Delta.$$

To make the above inequality satisfy, the necessary condition for δ_i^* is

$$\|\delta_i^*\|_2 \geq \frac{L - \sqrt{L^2 - 8\beta\Delta}}{4\beta}.$$

$\square$

Theorem A.2 (Restate of Theorem 5.3). *Let α be the loss minimum of dataset D under all linear classifier, the weight matrix of a linear classifier be denoted as W, where ℓ_2 norm of each row is normalized to $\sqrt{d}$. For any two rows of W, their cosine similarity does not exceed γ. The loss function is $\mathcal{L}_{cls}(x, y; \theta) = \max(\max_{t \neq y} f_\theta(x)_t - f_\theta(x)_y, 0)$. Then, there exist poisons $\{\delta_i\}_{i=1}^N$ satisfying $\mathcal{D}_c(D, D_\delta) \leq O(\frac{\alpha}{(1-\gamma)^2})$, such that the poisoned dataset is linearly separable.*

Proof. Assume that $\delta_i = \beta_i W_{y_i}$. We only need to prove that there exists weight matrix $W = [W_1^T, \cdots, W_C^T]$, such that $W_{y_i}(x_i + \delta_i) > W_{y'}(x_i + \delta_i)$ holds for $y' \neq y_i$ for all $i \in [N]$.

This can be concluded from

$$\beta_i(\|W_{y_i}\|_2^2 - \|W_{y_i}\|_2\|W_{y'}\|_2 s(W_{y_i}, W_{y'})) > \max_{y' \neq y_i}(W_{y'}x_i - W_{y_i}x_i), \forall i,$$

which can be derived from

$$\beta_i > \frac{\mathcal{L}_{cls}(x_i, y_i, W)}{\|W_{y_i}\|_2(\|W_{y_i}\|_2 - \gamma\|W_{y'}\|_2)}, \forall i.$$

Therefore, as long as δ_i satisfies

$$\|\delta_i\|_2 > \frac{\mathcal{L}_{cls}(x_i, y_i, W)}{d(1 - \gamma)}, \forall i,$$

the above inequality satisfies.

The chamfer distance of clean data and poisoned data satisfies that

$$\mathcal{D}_c(D, \delta) \leq \frac{2}{N}\sum_{i=1}^N \|\delta_i\|_2^2,$$

14

meanwhile, we can find $\|\delta_i\|_2 \leq \frac{2Loss(x_i, y_i, W)}{d(1-\gamma)}$ satisfies abve inequalities. Due to normalization,

$$\mathcal{L}_{\text{cls}}(x_i, y_i, W) \leq \max_{y' \neq y_i}(W_{y'}x_i - W_{y_i}x_i) \leq 2\sqrt{d}\|x_i\|_2 \leq 2d,$$

as x_i lies in $[-1, 1]$.

Therefore,

$$\frac{2}{N}\sum_{i=1}^{N}\|\delta_i\|_2^2 \leq \frac{8}{(1-\gamma)^2 d^2}\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{cls}}^2(x_i, y_i, W) \leq \frac{32}{(1-\gamma)^2}\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{cls}}(x_i, y_i, W).$$

The above inequality holds for all W, hence we only need $\mathcal{D}_c(D, D_\delta) \leq O(\frac{\alpha}{(1-\gamma)^2})$. $\square$

Theorem A.3 (Restate of Theorem 4.1). *Assume that $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{dis}}$ are continuous, and the network's hypothesis space $\mathcal{H}_{\mathcal{F}}$ is compact. Let $D_\delta = \{(x_i + \delta_i, y_i)\}_{i=1}^{N}$ be the poisoned dataset of D. Then, there exists an optimal point (δ^*, θ^*) for the bi-level optimization* (3)*, and the optimal perturbation δ^* satisfies $\mathcal{L}_{\text{dis}}(D, \delta^*) \leq \frac{1}{\beta}\left[\min_{\theta} \mathcal{L}_{\text{cls}}(D; \theta) - \min_{\delta,\theta} \mathcal{L}_{\text{cls}}(D_\delta; \theta)\right]$.*

Proof. Since $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{dis}}$ are continuous, their sum and expectation are also continuous. Therefore, since $\mathcal{H}_{\mathcal{F}}$ is compact, the optimal minimum point δ^* exists. In addition, as $\delta \in \mathbb{R}^{n \times 3}$ is the Euclid space, we may further assume that $\delta \in [-2, 2]^{n \times 3}$ as we can assume 3D point clouds always lie in $[-1, 1]^{n \times 3}$ after normalization. Therefore, the perturbation space is also compact, resulting in the optimal minimum point δ^* exists.

Denote

$$\alpha_1 = \min_{\theta} \mathcal{L}_{\text{cls}}(D; \theta) = \min_{\theta} \mathbb{E}_{(x,y) \in D}\mathcal{L}_{\text{cls}}(x, y; \theta),$$

$$\alpha_2 = \min_{\delta,\theta} \mathcal{L}_{\text{cls}}(D_\delta; \theta) = \min_{\theta} \min_{\delta} \mathbb{E}_{(x,y) \in D}\mathcal{L}_{\text{cls}}(x + \delta(x), y; \theta).$$

There exists a θ_0 such that $\mathcal{L}_{\text{cls}}(D; \theta_0) = \alpha_1$. Therefore, the optimal point (δ^*, θ^*) satisfies

$$\mathbb{E}_{(x,y) \in D}\left[\mathcal{L}_{\text{cls}}(x + \delta^*, y; \theta^*) + \beta \cdot \mathcal{L}_{\text{dis}}(x + \delta^*, x)\right] \leq \mathbb{E}_{(x,y) \in D}\left[\mathcal{L}_{\text{cls}}(x + 0, y; \theta_0) + \beta \cdot \mathcal{L}_{\text{dis}}(x + 0, x)\right] = \alpha_1.$$

Since $\mathcal{L}_{\text{cls}}$ is always not less than α_2, δ^* is in $\{\delta \,|\, \mathcal{L}_{\text{dis}}(D, \delta) \leq \frac{\alpha_1 - \alpha_2}{\beta}\}$. $\square$

Corollary A.4 (Restate of Corollary 4.3). *Let $\mathcal{L}_{\text{cls}}(x, y; \theta) = \max(\max_{t \neq y} f_\theta(x)_t - f_\theta(x)_y, 0)$, where $f_\theta(\cdot)$ is the output of logits layer, and $\mathcal{L}_{\text{dis}}$ be a metric function. If $\mathcal{H}_{\mathcal{F}}$ includes the function capable of correctly classifying D, then the equilibrium of the bi-level optimization* (3) *will degenerate to $\delta^* = 0$.*

Proof. Since $\mathcal{H}_{\mathcal{F}}$ contains the true function, it has

$$\min_{\theta} \mathcal{L}_{\text{cls}}(D; \theta) = 0, \min_{\delta,\theta} \mathcal{L}_{\text{cls}}(D_\delta; \theta) = 0.$$

Therefore, by Theorem 4.1, as $\mathcal{L}_{\text{dis}}$ is a metric function, δ^* lies in the perturbation space

$$\{\delta \,|\, \mathcal{L}_{\text{dis}}(D, \delta) \leq 0\} = \{0\}.$$

$\square$

B. More Discussions on Related Work

Availability attacks. Privacy issues have been extensively studied in the field of privacy-preserving machine learning (Shokri & Shmatikov, 2015; Abadi et al., 2016; Shokri et al., 2017), including studies on availability attacks (Feng et al., 2019; Huang et al., 2020; Sandoval-Segura et al., 2022a). Availability attacks are a type of data poisoning, which allow the attacker to perturb the training dataset under a small norm restriction. These attacks aim to cause test-time errors while maintaining the semantic integrity of the dataset and not affecting the normal usage of the data by legitimate users. Feng et al. (2019) used auto-encoder to generate adversarial poisons. Huang et al. (2020) proposed a bi-level error-minimizing approach to generate effective availability attacks for privacy preserving, which is called unlearnable examples. Fowl et al. (2021) used stronger adversarial poisons to achieve availability attacks under error-maximizing approach. Yuan & Wu

(2021) used neural tangent kernels to generate error-minimizing poison noises. Yu et al. (2022) viewed availability attacks as a type of shortcut learning, and proposed the model-free synthetic noises to achieve comparable attack performance. Sandoval-Segura et al. (2022b) also proposed a model-free availability attack based on autoregressive perturbations. Wu et al. (2022) expanded availability attacks to ℓ_0 norm, generating poisons on one pixel, known as one-pixel shortcuts. Sadasivan et al. (2023) generated effective unlearnable datasets based on class-wise convolution filters. Chen et al. (2023) proposed self-ensemble protection to improve adversarial poisons by using several checkpoints. In recent years, availability attacks have been generalized to adversarial training (Fu et al., 2022; Wen et al., 2023), robustness (Tao et al., 2022), self-supervised learning (He et al., 2022; Ren et al., 2022; Wang et al., 2024), unsupervised learning (Zhang et al., 2023b), natural language processing (Li et al., 2023). There are also many work to defend availability attacks recently. Tao et al. (2021) proposed an upper bound of clean test risk under defense of adversarial training. Qin et al. (2023); Liu et al. (2023) used stronger data augmentations to defend availability attacks more efficiently. Sandoval-Segura et al. (2023) proposed an orthogonal projection method to defend availability attacks. Jiang et al. (2023); Dolatabadi et al. (2023) defended availability attacks by using a pre-trained diffusion model to purify injected poison noises.

Attacks on 3D point clouds.   Vulnerability of 3D point cloud classifiers has become a grave concern due to the popularity of 3D sensors in applications. Many works (Xiang et al., 2019; Cao et al., 2019; Miao et al., 2022) apply adversarial attacks to the 3D point cloud domain. Xiang et al. (2019) proposed point generation attacks by adding a limited number of synthetic points to the point cloud. Further research (Huang et al., 2022; Liu & Hu, 2022; Tsai et al., 2020; Zheng et al., 2019; Zhang et al., 2022a; Miao et al., 2022; Zhang et al., 2023a; He et al., 2023a; Lou et al., 2024) has employed gradient-based point perturbation attacks. However, adversarial attacks against 3D point cloud classifiers are all test-time evasion attacks. In contrast, our objective is to attack against the model at train-time. The backdoor attack is another type of attack that poisons training data with a stealthy trigger pattern (Chen et al., 2017a). (Li et al., 2021) generated triggers on 3D point clouds by conducting orientation and interaction implanting for both poison-label and clean-label approach. (Xiang et al., 2021) injected backdoors for 3D point clouds by optimizing local geometry and spatial location. (Gao et al., 2023) proposed imperceptible and robust backdoor attacks utilizing weighted local transformation. However, the backdoor attack does not harm the model's performance on clean data.

C. Experiment Details

C.1. Datasets

ModelNet40.   ModelNet40 (Wu et al., 2015) is a collection of CAD models with 40 object categories, which is divided into 9843 training data and 2468 test data[1].

ScanObjectNN.   We use a real-world point cloud object dataset namely ScanObjectNN (Uy et al., 2019), which contains 15 classes, and is divided into 2309 training data and 581 test data[2].

IntrA.   We use an open-access 3D intracranial aneurysm dataset IntrA (Yang et al., 2020), which is a binary classification dataset divided into 1619 training data and 406 test data[3].

BSM2017.   We use Basel Face Model 2017 (Gerig et al., 2018) as our source face models to generate 3D point clouds of face scans[4]. We follow the setting in (Zhang et al., 2022b) to generate 100 classes of face scans, each of them contains 50 point clouds. Furthermore, we divide the whole 5000 face data into training and test part, containing 4000 and 1000 data respectively. After that, we randomly choose 1024 points for each point clouds before our training and poisoning approach.

C.2. Networks

PointNet.   PointNet (Qi et al., 2017a) is a popular 3D point cloud classification network, and we use the architecture before the global feature as our feature extractor, the dimension of feature space is 1024.

---

[1]https://shapenet.cs.stanford.edu/media/modelnet40_ply_hdf5_2048.zip
[2]https://hkust-vgd.ust.hk/scanobjectnn/
[3]https://drive.google.com/drive/folders/1yjLdofRRqyklgwFOC0K4r7ee1LPKstPh/IntrA.zip
[4]https://faces.dmi.unibas.ch/bfm/bfm2017.html

PointNet++. We use the multi-scale grouping (MSG) network, a variant of PointNet, PointNet++ (Qi et al., 2017b), which feature dimension is 1024.

DGCNN. DGCNN (Wang et al., 2019) uses EdgeConv blocks to obtain effective network for classification and segmentation of point cloud. The feature dimension is 2048 as we use both max pooling and average pooling when obtaining features.

C.3. Details Setting of Algorithm

Table 7. Detailed generation process of baselines and our FC-EM attack.

Poison	FC-EM	REG-EM	REG-AP(-T)	EM	AP(-T)
Poison Batch Size	128	32	32	32	32
Epochs	200	200	100	200	100
Learning Rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
LR Schedule	Plateau	Plateau	Plateau	Plateau	Plateau
Dist Regularization	Chamfer	Chamfer	Chamfer	-	-
Reg Strength	1.0	1.0	1.0	-	-
ℓ_∞ Restraint	-	-	-	0.08	0.08
Temperature	0.1	-	-	-	-
PGD Steps	10	10	250	10	250
PGD Step Size	0.015	0.015	0.001	0.015	0.001

Adaptive regularization stregnth β. As shown in Table 10, choosing the suitable β is a tricky problem and could significantly influence the poison power and imperceptibility. Therefore, to mitigate this question, we may change β adaptively at each epoch for different inputs. In detail, we set an initial $\beta = 1.0$, then we multiply/divide a scale factor s at each epoch based on the $\mathcal{L}_{fc}$ of each input. If $\mathcal{L}_{fc}$ is large, we think the poison needs to focus more on similarity loss, otherwise, we think the poison should consider more on distance loss. We set $s = 1.1$, the threshold of $\mathcal{L}_{fc}$ be 20% largest inputs of the similarity loss, if $\mathcal{L}_{fc}$ is top 20%, we multiply β with s, otherwise if $\mathcal{L}_{fc}$ is last 80%, we divide β with s. The maximum scale factor is 10, i.e, β will change dynamicly in $[0.1, 10]$.

D. More Experiments

D.1. More Discussions on ScanObjectNN and OmniObject3D

Table 2 showcases the results of experiments conducted on ScanObjectNN, revealing similar trends to those observed in ModelNet40. Our FC-EM attack outperforms ℓ_∞ based availability attacks, yielding the lowest test accuracy and better Chamfer and Hausdorff distances, which implies better imperceptibility of the poisoned point clouds.

The second row of Figure 4 provides visualizations of clean and various poisoned point clouds on ScanObjectNN. Notably, FC-EM poisoned point clouds maintain the structure with fewer outliers, while EM, AP, AP-T, and REG-AP-T exhibit more outliers and significant deformations and spurs, which are easily perceived by humans visually. Although REG-EM and REG-AP show decent imperceptiblity on point clouds, their poisoning efficacy, as indicated in Table 2, is compromised by relatively high test accuracy.

It is worth noting that some ℓ_∞ methods like AP-T and REG-AP-T exhibit quite strong poisoning abilities on ScanObjectNN, unlike their weaker counterparts on ModelNet40. This discrepancy may stem from ScanObjectNN being a scanned dataset rather than a generated one, making it inherently more challenging to learn, evident in its lower clean accuracy. Therefore, ScanObjectNN proves to be more susceptible to poisoning, allowing even relatively weaker poison methods to achieve commendable attack performance.

We also conduct additional experiments on the point cloud version of OmniObject3D dataset (Wu et al., 2023), which has point clouds comprising 1024 points[5]. We compare FC-EM with other availability attacks on this dataset using the

---
[5]https://openxlab.org.cn/datasets/OpenXDLab/OmniObject3D-New/tree/main/raw/point_clouds/hdf5_/1024

Table 8. Quantitative results of baselines and our FC-EM on the OmniObject3D dataset. Our FC-EM attains the lowest test accuracy, inducing strongest effectiveness.

Poison	Acc	$\mathcal{D}_c(\times 10^{-7})$	$\mathcal{D}_h(\times 10^{-6})$
Clean	72.6%	-	-
EM	45.4%	39.2	9.2
AP	40.2%	35.8	9.1
AP-T	45.6%	33.1	9.1
REG-EM	66.4%	2.9	1.5
REG-AP	50.2%	9.6	3.3
REG-AP-T	53.5%	9.5	3.2
FC-EM (**ours**)	**20.1%**	11.3	6.9

PointNet architecture. as shown in Table 8, FC-EM yields lower test accuracy than the baseline methods under comparable Chamfer/Hausdorff distances, further validating the effectiveness of our proposed method.

D.2. Ablation Studies

Batch size. In the realm of 3D classification tasks, a common practice is to utilize a smaller batch size, like 32, during training (Qi et al., 2017a). However, given that FC-EM incorporates class-wise similarity for poison generation, a larger batch size may be more suitable to ensure that each batch contains both positive and negative samples. Like in 2D image tasks, people typically use 128 batch size for supervised training but a larger one for contrastive training (Chen et al., 2020). Motivated by this we set the batch size to be four times larger than the standard training, defaulting to 128. Nonetheless, we still assess poison performance across different batch sizes, as shown in Table 9. The results indicate that FC-EM exhibits a relatively low sensitivity to variations in batch size, although 128 appears to be the most optimal choice of batch size.

Table 9. Quantitative results on the test accuracy (Acc) of our FC-EM on different batch size. It demonstrates that FC-EM exhibits a relatively low sensitivity to variations in batch size.

Batch size	32	64	128	256	512
FC-EM	29.01%	28.57%	27.67%	31.73%	36.91%

Different regularization strength. To investigate the concrete impact of different distance regularization strength β, we systematically evaluate FC-EM across various β for comprehensive analysis, where the results are provided in Table 10. It is evident that with a smaller β, the distance regularization weakens, the poison performance enhances (lower test accuracy). However, this improvement comes at the cost of larger Chamfer and Hausdorff distance, indicating poorer imperceptibility. An pertinent question is the selection of an appropriate β maintaining decent imperceptibility and achieving lower test accuracy. In practice, we found experimentally that on ModelNet40 when the Chamfer distance is below 10^{-3}, the poisons appear nearly imperceptible. Conversely, a Chamfer distance exceeding 10^{-3} leads to severe deformations of the point clouds.

Table 10. Quantitative results on the test accuracy (Acc) and distances (Chamfer distance ($\mathcal{D}_c$) and Hausdorff distance ($\mathcal{D}_h$)) of our FC-EM under different β. With β increases, the test accuracy gets higher, while the Chamfer distance and the Hausdorff distance get lower.

β	0.1	0.2	0.4	0.6	0.8	1.0	1.5	2.0	3.0	5.0	10.0	20.0
Test Acc	19.73%	22.12%	25.36%	28.65%	27.19%	27.67%	28.32%	32.21%	41.45%	46.43%	56.52%	59.28%
$\mathcal{D}_c(\times 10^{-4})$	67.3	37.8	19.7	12.7	10.2	8.4	7.4	6.7	4.9	4.3	2.7	2.0
$\mathcal{D}_h(\times 10^{-3})$	119.9	75.7	41.9	21.6	19.7	13.5	12.4	11.7	8.1	7.6	5.3	4.5

FC-EM under ℓ_∞ restriction. Although FC-EM originates from breaking the degeneracy of the regularization process, the feature collision method can directly be applied under the classical ℓ_∞ norm restriction by simply modifying the loss to $\mathcal{L}_{\text{fc}}$ during the poison generation. Therefore, we also evaluate the performance of FC-EM under ℓ_∞ restriction with budget 0.08 in Table 11.

Results demonstrate that, even under ℓ_∞ restriction, FC-EM outperforms other methods, with the lowest test accuracy. This superiority may stem from FC-EM inducing stronger linear separability, as proved in Theorem 5.3, suggesting

enhanced poison power. However, it comprises with larger Chamfer and Hausdorff distance, potentially leading to a loss of imperceptibility. Furthermore, when FC-EM is applied under Chamfer restriction, it consistently outperforms its ℓ_∞ norm counterpart, highlighting the superiority of poisoning under distance regularization for 3D point clouds availability attacks.

Table 11. Quantitative results on the test accuracy (Acc) and distances (Chamfer distance ($\mathcal{D}_c$) and Hausdorff distance ($\mathcal{D}_h$)) of our FC-EM attack under Chamfer and ℓ_∞ restriction.

Restriction	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$
Chamfer	27.67%	8.4	13.5
ℓ_∞	59.52%	42.0	18.7

D.3. Poison Ratios

We assess FC-EM when only a portion of the dataset is available for poisoning. Results presented in Table 12 reveal that availability attacks degenerate significantly, even when only 1% of data remains clean. This limitation seems inherent to availability attacks, akin to the observed challenges in 2D image methods that effectiveness decreases when clean data is mixed (Huang et al., 2020; Fowl et al., 2021; Sandoval-Segura et al., 2022b).

Availability attacks are typically formulated with the assumption that all training data can be poisoned, serving as a privacy-preserving mechanism. It is reasonable to expect that if clean (private) data is leaked, the efficacy of attacks will diminish. However, developing a novel attack resilient to a small part of privacy leakage remains an intriguing problem, which may be left as the future work.

Table 12. Quantitative results on the test accuracy (Acc) of our FC-EM under different poison ratios.

Ratio	99%	95%	90%	80%	60%	30%
Acc(%)	65.24	75.73	79.42	82.29	85.62	87.07

D.4. Early Stopping

We provide the learning process of the victim model when trained on FC-EM poisoned dataset in Figure 6. It illustrates the rapid learning capability, with both training and validation accuracy approaching 100% in the initial epochs. Notably, in (Sandoval-Segura et al., 2022a), it was observed that availability attacks often lose some power when early-stopping is applied. However, FC-EM exhibits strong resistance to early-stopping, signifying that it effectively deceive models from learning useful features, rather than initially acquiring and later being forgotten.
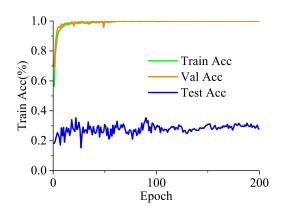

Figure 6. Training, Validation and Test accuracy(%) of FC-EM poisoned dataset on ModelNet40. It illustrates that both training and validation accuracy approaching 100% in the initial epochs, but the test accuracy remains low constantly.

D.5. Experiments on More Network Structures

There are more advanced 3D recognition network architectures(Xu et al., 2021b;a; Zhao et al., 2021; Ma et al., 2022). To further validate the effectiveness of FC-EM, we evaluate it alongside other availability attacks on two modern point cloud classification networks, PCT (Zhao et al., 2021) and PointMLP (Ma et al., 2022). Results are shown in Table 13. The results reveal that the performance of availability attacks on these networks aligns with those observed on PointNet, PointNet++, and DGCNN, with our FC-EM showing consistent efficacy and decent distances. This suggests that FC-EM's performance is largely independent of the underlying model architecture, making it applicable to both traditional and modern networks.

Table 13. Quantitative results of baselines and our FC-EM on ModelNet40 on PCT and PointMLP. Our FC-EM achieves the lowest test accuracy (Acc) and exhibits suitable imperceptibility (measured by Chamfer distance ($\mathcal{D}_c$) and Hausdorff distance ($\mathcal{D}_h$)), causing the strongest availability attack.

	PCT			PointMLP		
Poison	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$
Clean	92.02%	-	-	93.84%	-	-
EM	80.31%	24.2	11.2	72.20%	26.7	13.9
AP	67.18%	17.5	13.5	61.43%	18.5	11.6
AP-T	67.02%	18.4	14.0	45.34%	16.7	12.1
REG-EM	80.96%	4.5	5.6	85.86%	3.4	3.0
REG-AP	72.89%	6.2	12.7	87.56%	7.3	10.5
REG-AP-T	71.64%	7.8	13.3	69.41%	2.7	7.6
FC-EM (**ours**)	**34.32%**	5.5	7.5	**21.88%**	6.4	6.7

E. Degeneracy of Adversarial Poisons

Error-maximization poisoning approach, known as Adversarial Poisons (Fowl et al., 2021), was designed by training a clean source model and crafting poisons using 250 steps of PGD on the loss-manimization objective. It is worth noting that, for common adversarial training/attacks, people often use a much smaller steps of PGD, like PGD-10 and PGD-20, to obtain decent defense/attack performance. Differently, AP conducts a larger step of PGD attack, in other word, it use stronger adversarial attacks to achieve availability attacks.

Unfortunately, when we add distance regularization on such stronger attack approach, i.e., we directly extend AP to REG-AP, the distance regularization term will force the poison δ to be a weaker adversarial attack rather than a stronger one. Intuitively, distance regularization will lend δ become smaller while maintaining the effective attack. Therefore, REG-AP tends to find a successful adversarial attack under a smaller distance, rather than find a stronger adversarial attack within a attack restriction region.

Precisely, the following theorem demonstrates that, when using AP attacks, if data regularization is conducted, the convergent point is either not an adversarial point, or a weakest adversarial point that possesses the smallest distance. Therefore, choose larger attack step to obtain stronger adversarial poisons will fail if distance regularization exists.

Theorem E.1. *Let $\mathcal{L}_{\mathrm{cls}}$ be the 0-1 loss. The optimal point δ^* of the untargeted optimization (4) satisfies either (1)* $\delta^* = \arg\min_{\delta \in \Delta} \mathcal{L}_{\mathrm{dis}}(x + \delta, x)$, *where* $\Delta = \{\delta \mid f_\theta(x + \delta) \neq y\}$, *or (2)* $f_\theta(x + \delta^*) = y$. *Similarly, the optimal point δ^* of the targeted optimization (4) satisfies (1)* $\delta^* = \arg\min_{\delta \in \Delta} \mathcal{L}_{\mathrm{dis}}(x + \delta, x)$, *where* $\Delta = \{\delta \mid f_\theta(x + \delta) = \tau(y)\}$, *or (2)* $f_\theta(x + \delta^*) \neq \tau(y)$.

Proof. We only prove the untargeted case, the targeted one can be proved similarly. If the optimal point δ^* satisfies that $f_\theta(x + \delta) \neq y$, as $\mathcal{L}_{\mathrm{cls}}$ is the 0-1 loss, $\mathcal{L}_{\mathrm{cls}}(x + \delta, y, \theta^*) = 1$ for all $\delta \in \Delta$.

Therefore, $\delta^* = \arg\min_{\Delta} \mathcal{L}_{\mathrm{dis}}(x + \delta, x)$. For any successful adversarial attacks δ, optimization (4) will force to find the smallest magnitude of perturbation δ such that $\mathcal{L}_{\mathrm{dis}}$ achieves minimum. $\square$

F. Feature Disentanglement of Adversarial Poisons

We also add similarity loss for generation of AP poisons. It is worth noting that when generating poisons by maximizing feature collision loss, one could cause inner-class features disperse while inter-class features close. Therefore, the feature

used for classification will be disentangle under maximization, we called it *Feature Disentanglement Error-Maximization* (FD-AP) poison. Moreover, AP process generates poison noises that exhibits non-robust features in incorrect class, but $\mathcal{L}_{\text{fc}}$ itself does not include label information, we maximize both $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{fc}}$ in practice for effective poisons.

We formulate the FD-AP process as below:

$$\max_{\delta} \mathbb{E}_{(x,y)\in D}\left[\mathcal{L}_{\text{cls}}(x+\delta,y;\theta^*) + \zeta \cdot \mathcal{L}_{\text{fc}}(x+\delta,y;\theta^*,t) - \beta \cdot \mathcal{L}_{\text{dis}}(x+\delta,x)\right], \text{ (Untargeted)}$$

$$\text{or} \quad \min_{\delta} \mathbb{E}_{(x,y)\in D}\left[\mathcal{L}_{\text{cls}}(x+\delta,\tau(y);\theta^*) + \zeta \cdot \mathcal{L}_{\text{fc}}(x+\delta,\tau(y);\theta^*,t) + \beta \cdot \mathcal{L}_{\text{dis}}(x+\delta,x)\right], \text{ (Targeted)}$$

$$s.t. \quad \theta^* \in \arg\min_{\theta} \mathbb{E}_{(x,y)\in D}\left[\mathcal{L}_{\text{cls}}(x,y;\theta)\right], \tag{7}$$

where ζ is the hyperparameter controlling the strength of feature collision loss, the detailed algorithm is provided in Algorithm 2.

Experimental results in Table 14 show that our FD-AP method slightly outperform all of the other AP-based methods, without losing too much distance compared to REG-AP(-T), and strictly be better than AP(-T). However, FD-AP performs suboptimally compared with FC-EM. The main reason may result from that feature similarity operation needs more epoch to optimize, like in contrastive learning (Chen et al., 2020), they often require 1000 epochs rather than 100/200 epoch for standard training. But FD-AP instead, only generate poisons once with more PGD steps, which may undermine the utility of feature similarity. A better approach to convert AP-based methods to 3D point clouds is an interesting question. We may leave it as the future work.

Algorithm 2 Feature Disentanglement 3D Adversarial Poison Attack (FD-AP)

Input: A 3D point cloud training dataset $D = \{(x_i, y_i)\}_{i=1}^N$. A model with initialized parameters θ. Distance loss function $\mathcal{L}_{\text{dis}}$ and regularization strength β. Feature collision loss function $\mathcal{L}_{\text{fc}}$, strength ζ and temperature t. Classifier parameters α_θ and T_θ. Attack parameters α_a and T_a. Label group transformation τ for FD-AP-Targeted.

Output: Poisoned dataset $D_\delta = \{(x_i + \delta_i, y_i)\}_{i=1}^N$

$\delta_i \leftarrow 0, i = 1, 2, \cdots, N$

for $t_\theta = 1, \cdots, T_\theta$ **do**

    Sample a mini batch $B = \{x_{b_j}, y_{b_j}\}_{j=1}^{N_B}$

    $\theta \leftarrow \theta - \alpha_\theta \cdot \nabla_\theta \mathbb{E}_{(x_{b_j}, y_{b_j})\in B}\left[\mathcal{L}_{\text{cls}}(x_{b_j} + \delta_{b_j}, y_{b_j}; \theta)\right]$

for $i = 1, 2, \cdots, N$ **do**

    for $t_a = 1, \cdots, T_a$ **do**

        $\delta_N \leftarrow \delta_N + \alpha_a \cdot \nabla_{\delta_N} \mathbb{E}_{(x_N, y_N)\in D}\left[\mathcal{L}_{\text{cls}}(x_N + \delta_N, y_N; \theta) + \zeta \cdot \mathcal{L}_{\text{fc}}(x_N + \delta_N, y_N; \theta, t) - \beta \cdot \mathcal{L}_{\text{dis}}(x_N + \delta_N, x_N)\right]$

        (Untargeted)

        $\delta_N \leftarrow \delta_N - \alpha_a \cdot \nabla_{\delta_N} \mathbb{E}_{(x_N, y_N)\in D}\left[\mathcal{L}_{\text{cls}}(x_N + \delta_N, \tau(y_N); \theta) + \zeta \cdot \mathcal{L}_{\text{fc}}(x_N + \delta_N, \tau(y_N); \theta, t) + \beta \cdot \mathcal{L}_{\text{dis}}(x_N + \delta_N, x_N)\right]$

        (Targeted)

Table 14. Quantitative results on the test accuracy (Acc) on PointNet under various Error-maximum (AP) based poison methods for ModelNet40. Our proposed FD-AP-T obtains the lowest test accuracy (Acc) and gains good imperceptibility (measured by Chamfer distance ($\mathcal{D}_c$) and Hausdorff distance ($\mathcal{D}_h$)), outperforming all other AP-based availability attacks.

Poison	Acc	$\mathcal{D}_c(\times 10^{-4})$	$\mathcal{D}_h(\times 10^{-3})$
AP	69.37%	27.3	17.0
AP-T	61.95%	26.7	16.8
REG-AP	75.32%	5.5	11.5
REG-AP-T	72.61%	5.1	11.5
FD-AP	61.71%	11.8	17.1
FD-AP-T	52.88%	7.0	13.4