A More experiments and technical details

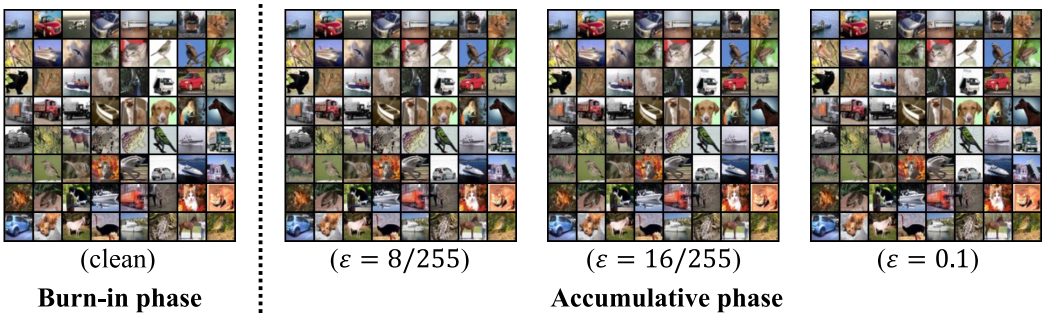
We provide more empirical results and technical details to support our conclusions in the main text.

A.1 Model architectures with group normalization (GN)

Table 1: Classification accuracy (%) of the simulated online learning models by using model architectures with group normalization (GN) on CIFAR-10 (we substitute BN layers in ResNet-18 with GN layers). We perform ablation studies on different tricks used in the accumulative phase.

Method	Acc. before trigger	Acc. after trigger	Δ
Clean trigger	82.99	83.58	$+0.59$
+ **accumulative phase**	80.84±1.08	76.16±0.76	-4.68±0.78
+ weight momentum	80.70±0.44	73.35±2.09	**-7.36±1.70**
Poisoned trigger			
+ $\epsilon = 8/255$	82.99	79.33	-3.66
+ **accumulative phase**	81.82±0.16	77.04±0.19	-4.78±0.34
+ weight momentum	81.35±0.81	75.40±0.75	-5.95±0.60
+ optimizing $\mathcal{P}$	80.00±0.98	75.26±0.74	-5.74±0.33
+ weight momentum	80.51±1.48	73.50±1.70	**-7.02±0.22**
+ $\epsilon = 16/255$	82.99	78.46	-4.53
+ **accumulative phase**	81.49±0.39	76.19±0.94	-5.30±0.56
+ weight momentum	81.05±0.99	73.44±1.72	-7.60±0.83
+ optimizing $\mathcal{P}$	80.75±0.84	72.75±1.31	-8.00±1.24
+ weight momentum	80.62±0.80	71.52±1.69	**-9.10±1.34**
+ $\epsilon = 0.1$	82.99	77.61	-5.38
+ **accumulative phase**	80.76±0.35	72.57±1.06	-8.20±0.72
+ weight momentum	80.40±0.75	70.64±1.64	-9.76±0.90
+ optimizing $\mathcal{P}$	80.05±0.47	70.49±1.31	-9.56±1.14
+ weight momentum	80.05±0.47	68.88±1.42	**-11.17±1.27**

A.2 Visualization of perturbed images



(clean)

Burn-in phase

($\varepsilon = 8/255$) ($\varepsilon = 16/255$) ($\varepsilon = 0.1$)

Accumulative phase

Figure 1: We provide visualization on the perturbed accumulative poisoning samples under $\epsilon = 8/255$, $\epsilon = 16/255$, and $\epsilon = 0.1$, respectively. As seen, the crafted adversarial noises are hardly perceptible, especially in large-scale scenes that a human observer cannot easily distinguish the noise patterns.

A.3 Technical details

Our methods are implemented by Pytorch [41], and run on GeForce RTX 2080 Ti GPU workers. The experiments of ResNet-18 are run by a single GPU. We assume that the poisoning adversaries have white-box accessibility to the model states, including the random seeds in the federated learning case. The CIFAR-10 dataset [23] consists of 60,000 32x32 colour images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images. We perform RandomCrop with four paddings and RandomHorizontalFlip in training as the data augmentation.

Computational complexity. Empirically, we set the mini-batch as 100, and use 10-steps PGD attacks to execute poisoning. The running time for the vanilla poisoning attack is 2.33 seconds per batch, and for our accumulative poisoning attack is 2.47 seconds per batch.

B More backgrounds

This section introduces more backgrounds on poisoning attacks and backdoor attacks, and details on the adversarial attacks that we use to craft accumulative poisoning samples in our methods. Finally, we describe the commonly used anomaly detection methods against adversarially crafted samples, following previous settings [40].

B.1 Poisoning attacks and backdoor attacks

There is extensive prior work on poisoning attacks, especially in the offline settings against SVM [3], logistic regression [36], collaborative filtering [27], feature selection [54], clustering [8], and neural networks [9, 21, 22, 38, 50]. Poisoning attacks in real-time data streaming are studied on online SVM [4], autoregressive models [1, 7], bandit algorithms [20, 31, 33], and classification [26, 52, 57].

Compared to poisoning attacks, backdoor attacks draw attention in more recent researches. These progresses involve backdoor attacks on self-supervised learning [42], point clouds [29, 51, 53], language models [28], graph neural networks [55], real physical world [56], brain computers [37], geenrative models [44], and image classification [14, 43, 45].

B.2 Adversarial attacks

In online learning case, the poisoning capability is constrained under ℓ_p-bounded threat models, where the perturbation δ is required to be bounded by a preset threshold ϵ under ℓ_p-norm, i.e., $\|\delta\|_p \leq \epsilon$. Below we introduce the details of several adversarial attacks that can be used in our experiments.

Fast gradient sign method (FGSM) [16] generates an adversarial example under the ℓ_∞ norm as

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot \operatorname{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}_{\mathrm{A}}(\boldsymbol{x}, y)), \tag{1}$$

where $\boldsymbol{x}$ is the original clean input, y is the input label, $\boldsymbol{x}^{adv}$ is the crafted adversarial input, and $\mathcal{L}_{\mathrm{A}}$ is the adversarial objective. The sign function sign is used according to the first-order approximation under ℓ_∞-norm [39, 48].

Projected gradient descent (PGD) [34] extends FGSM by iteratively taking multiple small gradient updates as

$$\boldsymbol{x}_{t+1}^{adv} = \operatorname{clip}_{\boldsymbol{x},\epsilon}\big(\boldsymbol{x}_t^{adv} + \eta \cdot \operatorname{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}_{\mathrm{A}}(\boldsymbol{x}_t^{adv}, y))\big), \tag{2}$$

where $\operatorname{clip}_{\boldsymbol{x},\epsilon}$ projects the adversarial example to satisfy the ℓ_∞ constraint and η is the step size. Note that PGD involves a random initialization step as $\boldsymbol{x}_0^{adv} \sim \mathcal{U}(\boldsymbol{x} - \epsilon, \boldsymbol{x} + \epsilon)$.

Momentum iterative method (MIM) [10] integrates a momentum term into basic iterative method (BIM) [24] with the decay factor $\mu = 1.0$ as

$$\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} \mathcal{L}_{\mathrm{A}}(\boldsymbol{x}_t^{adv}, y)}{\|\nabla_{\boldsymbol{x}} \mathcal{L}_{\mathrm{A}}(\boldsymbol{x}_t^{adv}, y)\|_1}, \tag{3}$$

where the adversarial examples are updated by

$$\boldsymbol{x}_{t+1}^{adv} = \operatorname{clip}_{\boldsymbol{x},\epsilon}(\boldsymbol{x}_t^{adv} + \alpha \cdot \operatorname{sign}(\boldsymbol{g}_{t+1})). \tag{4}$$

MIM has good performance as a transfer-based attack in the black-box setting.

B.3 Anomaly detection in the adversarial setting

Recently, many defense methods are proposed against poisoning attacks [5, 6, 15, 35, 46, 47, 49] and against backdoor attacks [2, 11, 13, 17–19, 30]. Since we apply adversarial attacking methods (i.e., PGD) to craft accumulative poisoning samples, we exploit related detection methods in the adversarial literature.

Kernel density (KD). In Feinman et al. [12], KD applies a Gaussian kernel $K(z_1, z_2) = \exp(-\|z_1 - z_2\|_2^2/\sigma^2)$ to compute the similarity between two features z_1 and z_2. There is a hyperparameter σ controlling the bandwidth of the kernel, i.e., the smoothness of the density estimation. For KD, we restore $1,000$ correctly classified training features in each class and use $\sigma = 10^{-2}$.

Local intrinsic dimensionality (LID). In Ma et al. [32], LID applies K nearest neighbors to approximate the dimension of local data distribution. The empirical estimation of LID is calculated as

$$\text{LID}(x) = -\left(\frac{1}{K}\sum_{i=1}^{K}\log\frac{r_i}{r_K}\right)^{-1}, \tag{5}$$

where r_i is the distance from x to its i-th nearest neighbor. Note that we actually apply negative LID (i.e., $-\text{LID}(x)$) to make sure that a lower metric value indicates outliers. Instead of computing LID in each mini-batch, we allow the detector to use a total of $10,000$ correctly classified training data points, and treat the number of K as a hyperparameter. For LID, we restore a total of $10,000$ correctly classified training features and use $K = 600$.

Gaussian-based detection. Gaussian mixture model (GMM) [58] and Gaussian discriminative analysis (GDA) [25] are two commonly used Gaussian-based detection methods. Both of them fit a mixture of Gaussian model in the feature space of trained models. The main difference is that GDA uses all-classes data to fit a covariance matrix, and GMM fits conditional covariance matrices. We calculate the mean and covariance matrix on all correctly classified training samples.

C The capacity of the poisoning attackers

In the federated learning cases, we denote G as the aggregated gradient from the clients, and G_n as the gradient from the n-th client, where $n \in [1, N]$. We considered three threat models in our paper:

C.1 No gradient clip: $G = \sum_n G_n$

In this case, we can only modify a *single* client, e.g., the k-th client, to achieve arbitrary poisoned aggregated gradient G^{poi}. Specifically, by the simple trick of recovered offset, we poison G_j to be $\mathcal{A}(G_j)$, where

$$\mathcal{A}(G_j) = G^{\text{poi}} - \sum_{n \neq j} G_n,$$

such that $\mathcal{A}(G_j) + \sum_{n \neq j} G_n = G^{\text{poi}}$. In our simulation experiments, we set batch size be 100, and treat each data point as a client. So we can only poison the gradient on a singe data point/client, i.e., the poisoning ratio is 1%. Ideally, in this case the poison ratio

$$\frac{1}{\text{number of clients}}$$

can be arbitrarily close to zero for large number of clients. The empirical results correspond to the "No clip" column.

C.2 Gradient clip after aggregation: $G = \textbf{Clip}_\eta(\sum_n G_n)$

Similar to the derivations above, in this case we can still poison a *single* client, e.g., the j-th client, such that

$$\textbf{Clip}_\eta\left(\mathcal{A}(G_j) + \sum_{n \neq j} G_n\right) = \textbf{Clip}_\eta\left(G^{\text{poi}}\right),$$

where the poisoning ratio is still 1% in our simulation experiments, and can be arbitrarily close to zero as discussed above. The empirical results correspond to the "ℓ_2-norm clip bound" and "ℓ_∞-norm clip bound" columns.

C.3 Gradient clip before aggregation: $G = \sum_n \mathbf{Clip}_\eta(G_n)$

In this case, assuming that we can modify M clients in the index set S, i.e., $|S| = M$ and the per-batch poison ratio is $\frac{M}{N}$. Specifically, for any $m \in S$, we poison G_m to be $\mathcal{A}_m(G_m)$. Under the gradient clip constraint, we want to achieve G^{poi} as well as we can, so we optimize the objective

$$\min_{\mathcal{A}_m, m\in S} \left\| G^{\text{poi}} - \left(\sum_{m\in S} \mathbf{Clip}_\eta\left(\mathcal{A}_m(G_m)\right) + \sum_{n\notin S} \mathbf{Clip}_\eta\left(G_n\right) \right) \right\|.$$

Under the mild condition that η is small, the optimal solution of the above objective is $\forall m \in S$,

$$\mathcal{A}_m(G_m) = G^{\text{poi}} - \sum_{n\notin S} \mathbf{Clip}_\eta\left(G_n\right).$$

and

$$\sum_{m\in S} \mathbf{Clip}_\eta\left(\mathcal{A}_m(G_m)\right) = \sum_{m\in S} \mathbf{Clip}_\eta \left(G^{\text{poi}} - \sum_{n\notin S} \mathbf{Clip}_\eta\left(G_n\right) \right) = \mathbf{Clip}_{M\eta} \left(G^{\text{poi}} - \sum_{n\notin S} \mathbf{Clip}_\eta\left(G_n\right) \right).$$

As we can see, poisoning more clients (i.e., larger M) can be regarded as relaxing the gradient clip constraint (i.e., relax from η to $M\eta$).

References

[1] Scott Alfeld, Xiaojin Zhu, and Paul Barford. Data poisoning attacks against autoregressive models. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[2] Hassan Ali, Surya Nepal, Salil S Kanhere, and Sanjay Jha. Has-nets: A heal and select mechanism to defend dnns against backdoor attacks for data collection scenarios. *arXiv preprint arXiv:2012.07474*, 2020.

[3] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning (ICML)*, 2012.

[4] Cody Burkard and Brent Lagesse. Analysis of causative attacks against svms learning from data streams. In *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*, pages 31–36, 2017.

[5] Javier Carnerero-Cano, Luis Muñoz-González, Phillippa Spencer, and Emil C Lupu. Regularization can help mitigate poisoning attacks... with the right hyperparameters. In *ICLR Workshop on Security and Safety in Machine Learning Systems*, 2021.

[6] Jian Chen, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. De-pois: An attack-agnostic defense against data poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 2021.

[7] Yiding Chen and Xiaojin Zhu. Optimal attack against autoregressive models by manipulating the environment. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[8] Antonio Emanuele Cinà, Alessandro Torcinovich, and Marcello Pelillo. A black-box adversarial attack for poisoning clustering. *arXiv preprint arXiv:2009.05474*, 2020.

[9] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *USENIX Security Symposium*, pages 321–338, 2019.

[10] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. Black-box detection of backdoor attacks with limited information and data. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.

[12] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[13] Ji Gao, Amin Karbasi, and Mohammad Mahmoody. Learning and certification under instance-targeted poisoning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

[14] Siddhant Garg, Adarsh Kumar, Vibhor Goel, and Yingyu Liang. Can adversarial weight perturbations inject neural backdoors. In *ACM International Conference on Information & Knowledge Management (CIKM)*, pages 2029–2032, 2020.

[15] Jonas Geiping, Liam Fowl, Gowthami Somepalli, Micah Goldblum, Michael Moeller, and Tom Goldstein. What doesn't kill you makes you robust (er): Adversarial training against poisons and backdoors. In *ICLR Workshop on Security and Safety in Machine Learning Systems*, 2021.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[17] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: Defending against backdoor attacks using robust statistics. In *International Conference on Machine Learning (ICML)*, 2021.

[18] Todd Huster and Emmanuel Ekwedike. Top: Backdoor detection in neural networks via transferability of perturbation. *arXiv preprint arXiv:2103.10274*, 2021.

[19] Charles Jin, Melinda Sun, and Martin Rinard. Provable guarantees against data poisoning using self-expansion and compatibility. *arXiv preprint arXiv:2105.03692*, 2021.

[20] Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Xiaojin Zhu. Adversarial attacks on stochastic bandits. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[21] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, pages 1885–1894. PMLR, 2017.

[22] Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *arXiv preprint arXiv:1811.00741*, 2018.

[23] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *The International Conference on Learning Representations (ICLR) Workshops*, 2017.

[25] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[26] Laurent Lessard, Xuezhou Zhang, and Xiaojin Zhu. An optimal control approach to sequential machine teaching. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2019.

[27] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[28] Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. Hidden backdoors in human-centric language models. *arXiv preprint arXiv:2105.00164*, 2021.

[29] Xinke Li, Zhiru Chen, Yue Zhao, Zekun Tong, Yabang Zhao, Andrew Lim, and Joey Tianyi Zhou. Pointba: Towards backdoor attacks in 3d point cloud. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.

[30] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[31] Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning (ICML)*. PMLR, 2019.

[32] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Michael E Houle, Grant Schoenebeck, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations (ICLR)*, 2018.

[33] Yuzhe Ma, Kwang-Sung Jun, Lihong Li, and Xiaojin Zhu. Data poisoning attacks in contextual bandits. In *International Conference on Decision and Game Theory for Security*, pages 186–204. Springer, 2018.

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[35] Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, and Jihun Hamm. How robust are randomized smoothing based defenses to data poisoning? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[36] Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

[37] Lubin Meng, Jian Huang, Zhigang Zeng, Xue Jiang, Shan Yu, Tzyy-Ping Jung, Chin-Teng Lin, Ricardo Chavarriaga, and Dongrui Wu. Eeg-based brain-computer interfaces are vulnerable to backdoor attacks. *arXiv preprint arXiv:2011.00101*, 2020.

[38] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 27–38, 2017.

[39] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Hang Su, and Jun Zhu. Boosting adversarial training with hypersphere embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[40] Tianyu Pang, Huishuai Zhang, Di He, Yinpeng Dong, Hang Su, Wei Chen, Jun Zhu, and Tie-Yan Liu. Adversarial training with rectified rejection. *arXiv preprint arXiv:2105.14785*, 2021.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

[42] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. *arXiv preprint arXiv:2105.10123*, 2021.

[43] Ahmed Salem, Michael Backes, and Yang Zhang. Don't trigger me! a triggerless backdoor attack against deep neural networks. *arXiv preprint arXiv:2010.03282*, 2020.

[44] Ahmed Salem, Yannick Sautter, Michael Backes, Mathias Humbert, and Yang Zhang. Baaan: Backdoor attacks against autoencoder and gan-based machine learning models. *arXiv preprint arXiv:2010.03007*, 2020.

[45] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. *arXiv preprint arXiv:2003.03675*, 2020.

[46] Sanjay Seetharaman, Shubham Malaviya, Rosni KV, Manish Shukla, and Sachin Lodha. Influence based defense against data poisoning attacks in online learning. *arXiv preprint arXiv:2104.13230*, 2021.

[47] Uday Shankar Shanthamallu, Jayaraman J Thiagarajan, and Andreas Spanias. Uncertainty-matching graph neural networks to defend against poisoning attacks. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[48] Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz. First-order adversarial vulnerability of neural networks and input dimension. In *International Conference on Machine Learning (ICML)*, 2019.

[49] Jack W Stokes, Paul England, and Kevin Kane. Preventing machine learning poisoning attacks using authentication and provenance. *arXiv preprint arXiv:2105.10051*, 2021.

[50] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning fail? generalized transferability for evasion and poisoning attacks. In *USENIX Security Symposium*, pages 1299–1316, 2018.

[51] Guiyu Tian, Wenhao Jiang, Wei Liu, and Yadong Mu. Poisoning morphnet for clean-label backdoor attack to point clouds. *arXiv preprint arXiv:2105.04839*, 2021.

[52] Yizhen Wang and Kamalika Chaudhuri. Data poisoning attacks against online learning. *arXiv preprint arXiv:1808.08994*, 2018.

[53] Zhen Xiang, David J Miller, Siheng Chen, Xi Li, and George Kesidis. A backdoor attack against 3d point cloud classifiers. *arXiv preprint arXiv:2104.05808*, 2021.

[54] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *International Conference on Machine Learning (ICML)*, pages 1689–1698, 2015.

[55] Jing Xu, Stjepan Picek, et al. Explainability-based backdoor attacks against graph neural networks. *arXiv preprint arXiv:2104.03674*, 2021.

[56] Mingfu Xue, Can He, Shichang Sun, Jian Wang, and Weiqiang Liu. Robust backdoor attacks against deep neural networks in real physical world. *arXiv preprint arXiv:2104.07395*, 2021.

[57] Xuezhou Zhang, Xiaojin Zhu, and Laurent Lessard. Online data poisoning attacks. In *Learning for Dynamics and Control*, pages 201–210. PMLR, 2020.

[58] Zhihao Zheng and Pengyu Hong. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.