Revisiting Data-Free Knowledge Distillation with Poisoned Teachers

Junyuan Hong [*1] **Yi Zeng** [*2] **Shuyang Yu** [*1] **Lingjuan Lyu** [3] **Ruoxi Jia** [2] **Jiayu Zhou** [1]

Abstract

Data-free knowledge distillation (KD) helps transfer knowledge from a pre-trained model (known as the teacher model) to a smaller model (known as the student model) without access to the original training data used for training the teacher model. However, the security of the synthetic or out-of-distribution (OOD) data required in data-free KD is largely unknown and under-explored. In this work, we make the first effort to uncover the security risk of data-free KD w.r.t. untrusted pre-trained models. We then propose Anti-Backdoor Data-Free KD (ABD), the first plug-in defensive method for data-free KD methods to mitigate the chance of potential backdoors being transferred. We empirically evaluate the effectiveness of our proposed ABD in diminishing transferred backdoor knowledge while maintaining compatible downstream performances as the vanilla KD. We envision this work as a milestone for alarming and mitigating the potential backdoors in data-free KD. Codes are released at https://github.com/illidanlab/ABD.

1. Introduction

In recent years, deep learning (DL) has witnessed tremendous success in solving real-world challenges (Yu et al., 2022; Yamada et al., 2020; Zhang et al., 2020) by training huge models on giant data (Dosovitskiy et al., 2020; Tolstikhin et al., 2021; Wang et al., 2022c). Yet the performance-favored large model size has hindered their deployment to resource-limited (Beyer et al., 2022) and communication-limited (Tan et al., 2022) systems that meanwhile require responsive inferences, e.g., on tiny sensors, and frequent sharing of model parameters, e.g., federated learning (Konečný et al., 2016; Zhu et al., 2021).

To tailor the highly performant large models for the budget-constrained devices, knowledge distillation (KD) (Hinton et al., 2015) and more recently data-free KD (Chawla et al., 2021; Ye et al., 2020; Fang et al., 2022), has emerged as a fundamental tool in the DL community. Data-free KD, in particular, can transfer knowledge from a pre-trained large model (known as the *teacher model*) to a smaller model (known as the *student model*) without access to the original training data of the teacher model. The non-requirement of training data generalizes KD to broad real-world scenarios, where data access is restricted for privacy and security concerns. For instance, many countries have strict laws on accessing facial images (Parkhi et al., 2015), financial records (Shah et al., 2022), and medical information (Antonelli et al., 2022). Recently, data-free KD also empowers federated learning on heterogeneous clients (Zhu et al., 2021; Seo et al., 2022) and on low-bandwidth communication networks (Zhu et al., 2022; Zhang et al., 2022b;a).

Despite the benefits of data-free KD and the vital role it has been playing, a major security concern has been overlooked in its development and implementation: *Can a student trust the knowledge transferred from an untrusted teacher?* The untrustworthiness comes from the non-trivial chance that pre-trained models could be retrieved from non-sanitized or unverifiable sources, for example, third-party model vendors (Liu et al., 2022) or malicious clients in federated learning (Bagdasaryan et al., 2020). One significant risk is from the *backdoor* pre-implanted into a teacher model (Jia et al., 2022), which alters model behaviors drastically in the presence of predesigned triggers but remains silent on clean samples. As traditional attacks typically require to poison training data (Gu et al., 2019; Souri et al., 2021; Barni et al., 2019; Zeng et al., 2022b), it remains unclear if student models distilled from a poisoned teacher will suffer from the same threat without using the poisoned data.

In this paper, we take the first leap to uncover the *data-free backdoor transfer* from a poisoned teacher to a student through comprehensive experiments on 10 backdoor attacks. We evaluated one vanilla KD using clean training data (Hinton et al., 2015) and three training-data-free KD method which use synthetic data (ZSKT (Micaelli & Storkey, 2019) & CMI (Fang et al., 2021)) or

[*]Equal contribution [1]Michigan State University, Michigan, USA [2]Virginia Tech, Virginia, USA [3]Sony AI, Japan. Correspondence to: Lingjuan Lyu <lingjuan.lv@sony.com>, Jiayu Zhou <jiayuz@msu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

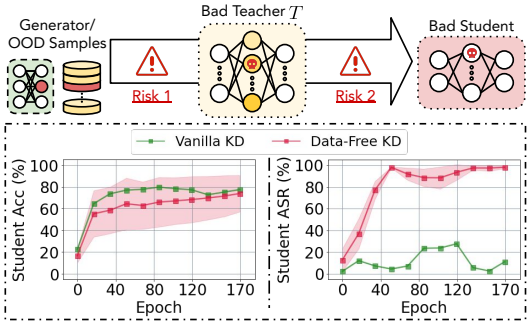

Figure 1. Data-free KD may transfer the backdoor knowledge (high Attack Success Rate, or ASR, from poisoned teachers to the student). The experiment is conducted on CIFAR-10 with a Trojan WM-poisoned teacher model (Liu et al., 2018). Vanilla KD denotes the KD with 10,000 clean in-distribution samples. The data-free KD is averaged with three methods that are based on synthetic or OOD data. The shadowed region is the standard deviation. We depict the student model's performance on clean samples (Student Acc) and samples patched with the backdoor trigger (Student ASR).

out-of-distribution (OOD) data as surrogate distillation data (Asano & Saeed, 2021). To highlight the risks, we showcase the result of distilling a poisoned pre-trained WideResNet-16-2 (Zagoruyko & Komodakis, 2016) as the teacher on CIFAR-10 in Fig. 1. Our main observations in Section 3 are summarized as follows and essentially imply two identified risks in data-free KD: (1) Vanilla KD does not transfer backdoors by using clean in-distribution data, while all three training-data-free distillations suffer from backdoor transfer by 3 to 8 types of triggers out of 10 with a more than 90% attack success rate. Contradicting the two results indicates the *poisonous nature of the surrogate distillation data* in data-free KD; (2) The successful attack on distillation using trigger-free out-of-distribution (OOD) data demonstrates that triggers are not essential for backdoor injection, but the *poisoned teacher supervision* is.

Upon observing aforementioned two identified risks, we propose a plug-in defensive method, Anti-Backdoor Data-Free KD (ABD), that works with general data-free KD frameworks. ABD aims to suppress and remove any backdoor knowledge being transferred to the student, thus mitigating the impact of backdoors. The high-level idea of ABD is two-fold: **Shuffling Vaccine (SV)** during distillation: suppress samples containing potential backdoor knowledge being fed to the teacher (mitigating backdoor information participates in the KD); Student **Self-Retrospection (SR)** after distillation: synthesize potential learned backdoor knowledge and unlearns them at later training epochs (the backstop to unlearn acquired malicious knowledge). We believe ABD is a significant step towards making data-free KD secure and the downstream student trustworthy. The main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to uncover the security risk of data-free KD regarding untrusted pre-trained models on 10 backdoor types and 4 diverse distillation methods.
- We identify two potential causes for the backdoor infiltrating from the teacher to the student via data-free KD, that may inspire defense methods.
- To mitigate the data-free backdoor transfer, we propose ABD, the first plug-in defensive method for data-free KD methods.
- To evaluate the effectiveness of the ABD, we conduct extensive experiments on 2 benchmark datasets and 10 different attacks to show ABD's efficacy in diminishing the transfer of malicious knowledge.

2. Background

In this section, we introduce the preliminaries on backdoor attacks and data-free KD, and then we define the threat model considered in the paper.

Backdoor attacks in pre-trained models. Backdoor attacks are an emerging security threat to DL systems when untrusted data/models/clients participate in the training process (Li et al., 2020c). Backdoor attacks have developed from using sample-independent visible triggers (Gu et al., 2019; Chen et al., 2017; Liu et al., 2018) to more stealthy and powerful attacks with sample-specific (Li et al., 2021a) or visually imperceptible triggers (Li et al., 2020a; Nguyen & Tran, 2020; Zeng et al., 2021; Wang et al., 2022b). More advanced attacks with clean labels ensure the manipulated features are semantically consistent with corresponding labels to better evade manual inspections (Turner et al., 2019; Souri et al., 2021; Zeng et al., 2022b). The above backdoor attacks can be easily deployed to obtain a poisoned model, T^*, by minimizing:

$$\mathbb{E}_{(\boldsymbol{x},y)\sim D}\left[\underbrace{L(T(\boldsymbol{x}),y)}_{\text{clean task}}+\underbrace{L(T(\boldsymbol{x}+\delta),t)}_{\text{backdoor task}}\right], \quad (1)$$

where L is the cross-entropy loss; the clean task denotes the model performance on samples drawn from the clean distribution, D, without triggers (correctly classifying x as label y); and the backdoor task denotes the malicious behavior of the model on observing samples patched with a trigger δ (classifying $x+\delta$ as the target label t). In this paper, we consider a case where the teacher model in KD is potentially inserted with a backdoor, and we focus on analyzing and resolving the associated security risks.

Data-free knowledge distillation (KD). Without ambiguity, KD in this work refers to offline response-based KD (Hinton et al., 2015). Given a teacher model, $T(\cdot)$, and some in-distribution samples (of the same distribution as the training data for the teacher), $\boldsymbol{x}\sim D$, a typical optimization goal of response-based KD is to obtain student model

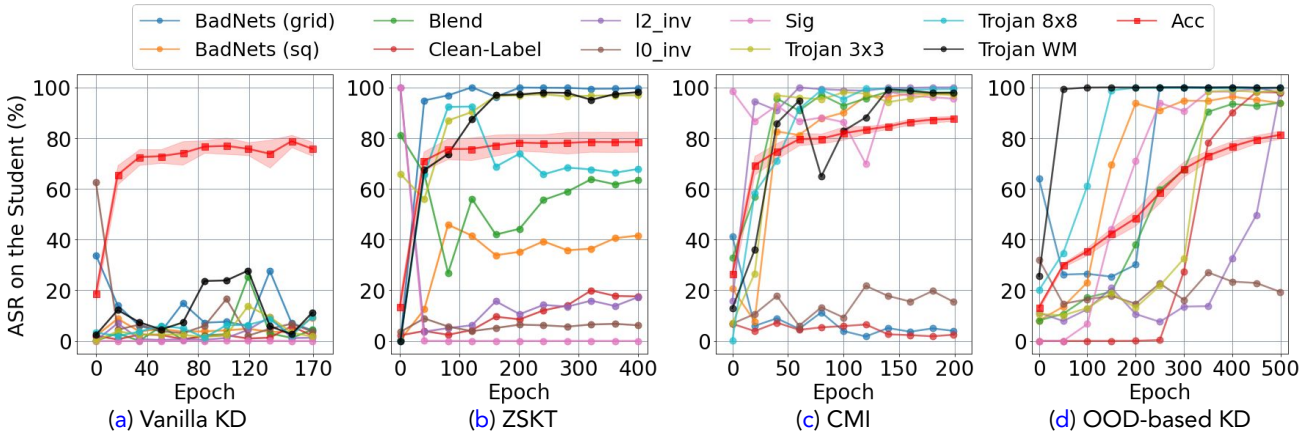

Figure 2. Backdoor Attack Success Rates (ASRs) of the distilled student model using the vanilla KD with clean in-distribution samples (a) and data-free KD using synthetic (b, c) or OOD (d) samples. The clean accuracy (Acc) of each figure is plotted with standard deviations among different attack-poisoned CIFAR-10. We run each KD method with different but sufficient training epochs to ensure convergence. Existing data-free KD methods may lead to the transfer of backdoor knowledge when poisoned teachers' participation.

parameters, θ, that minimizes the Kullback-Leibler divergence loss, $D_{KL}(\cdot)$, between the output softmax-logits of the teacher and the student, $S(\cdot|\theta)$:

$$\theta = \arg\min_\theta \mathbb{E}_{\boldsymbol{x}\sim D}\left[D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}|\theta\right)\right)\right]. \quad (2)$$

With data unavailability becoming more common in real-world DL settings due to privacy, legality, security, and confidentiality concerns, the development or implementation of KD has thus shifted to data-free settings. The key difference between data-free KD and vanilla KD is that the samples used for KD are synthetic (Chen et al., 2019; Micaelli & Storkey, 2019) or sampled from out-of-distribution (OOD) domains (Asano & Saeed, 2021). Promising implementations of data-free KD have also been demonstrated in advanced federated learning frameworks (Tang et al., 2022; Zhu et al., 2021; Zhang et al., 2022b). For generation-based methods, the dataset D in Eq. (2) is replaced by a trainable data generator or a set of trainable images. Generally, the dataset or generator can be parameterized as P and trained by maximizing the disagreement between the teacher and student models:

$$\max_P \mathbb{E}_{\boldsymbol{x}\sim P}\left[D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}\right)\right)\right].$$

Here, representative implementations include the first adversarial data-free distillation, Zero-Shot Knowledge Transfer (ZSKT) (Micaelli & Storkey, 2019), the state-of-the-art data-free KD methods, CMI (Fang et al., 2021). For OOD-based methods, we utilize the single-image extrapolation (Asano & Saeed, 2021), which extracts patches from a single image as training data for Eq. (2). For simplicity, we also denote the set of data as non-trainable P.

Even though techniques for data-free settings enable KD to be generalized more flexibly to data-constrained environments, no existing work has taken a closer look at the potential security risk of doing so. Given the fast development and emerging implementations of data-free KD for security concerning tasks, it is crucial to understand this security risk and study the countermeasures.

Threat model: Knowledge of attacker and defender. For the purpose of risk evaluation and defense, we consider a standard security threat model where an un-trustworthy party participates in the teacher-training process. The **attacker** performs attacks *only* by publicly releasing well-trained models inserted with backdoors or directly transmitting the poisoned model to the user (e.g., in federated learning). A user wishes to deploy the model's knowledge for further use but may require a different model structure due to size/memory constraints (Wu et al., 2022), or client heterogeneity (Zhu et al., 2021). For the **defender**, the original training data used by the attacker is unavailable for knowledge transfer, and the goal is to develop a practical countermeasure to diminish the chance of transferring backdoor knowledge in data-free KD without additional knowledge requirements, i.e., the defender only has access to the teacher model. Because of the data-free assumption, existing defenses requiring a clean dataset, for instance, (Li et al., 2021c; Zeng et al., 2022a; Wang et al., 2022a) are typically excluded in our scenarios.

3. Data-Free Can Steam Security Risks

General Threats on Data-free Distillation. For an empirical evaluation of the existing data-free KD methods, We consider 10 different backdoor attacks, BadNets with grid (grid) (Gu et al., 2019), BadNets (sq), Blend (Chen et al., 2017), Clean-label (Turner et al., 2019), l2-invisible (l2_inv) (Li et al., 2020a), l0-invisible (l0_inv), Sig (Barni

et al., 2019), Trojan Square 3×3 (Trojan 3×3), Trojan Square 8×8 (Trojan 8×8), and Trojan watermark (Trojan WM) (Liu et al., 2018). These attacks are then deployed to train 10 poisoned teacher models (WideResNet-16-2 (Zagoruyko & Komodakis, 2016)) with attack settings referred to in their original papers. The poisoned teacher models' performances and attack visualizations are provided in Figure 5. We further deploy different KD methods on these well-trained teacher models and obtain the respective student models for evaluation. The results of these data-free KD methods are then compared to the vanilla KD, which uses 10,000 clean, in-distribution CIFAR-10 samples. We depict the attack success rate (ASR) of these attacks on the distilled student models in Figure 2. The clean accuracy (Acc) on benign samples is similar when each method is run till converge (vanilla KD takes 170 epochs to converge. It takes 400, 200, and 500 epochs for ZSKT, CMI, and OOD-based methods to converge, respectively). We combined the Acc of all the models distilled by the respective KD method into a single red line with standard deviation.

From Figure 2, we find that all the evaluated data-free KD approaches have transferred some of the attack's malicious knowledge from the poisoned teachers to the student. Based on the difference in the data used for knowledge distillation, we now highlight two potential risks that may lead to the transfer of backdoor knowledge. Additional results on other dataset-setting are presented in Appendix B.

Potential Risk in Bad Synthetic Input Supply. Noting the Attack Success Rate (ASR) result on the students with vanilla KD using clean in-distribution samples is utterly different from the data-free settings' results. The key difference between the vanilla KD and data-free KD is the data supply, i.e., the input taken in by the teacher model. We hereby highlight a potential risk associated with the input supplied to the teacher in data-free KD. In particular, we find the poisoned teacher's participation in the synthetic data generation may lead to the generation of poisoned samples. We can assume the student starts without backdoor knowledge, i.e., $S(\boldsymbol{x}+\delta_t) = S(\boldsymbol{x})$[1] for any $\boldsymbol{x}$. We may simplify the data generation by maximizing the error by the student: $\boldsymbol{x}_p = \arg\max_{\boldsymbol{x}} D_{KL}\left(T\left(\boldsymbol{x}\right) \| S\left(\boldsymbol{x}\right)\right)$. We may reformulate $\boldsymbol{x}$ as $\boldsymbol{x} = \boldsymbol{x}_0 + \delta$ and assume $\boldsymbol{x}_0 \in \{\boldsymbol{x} | T(\boldsymbol{x}) \neq t\}$. We assume $\delta \in C_{<\epsilon}$, which is a potential backdoor within a bounded constraint, e.g., $\|\delta\| \leq \epsilon$. Note that though there is no constrained optimization in practice, the small learning rate and uncontrolled optimization may converge into a pitfall. Thus, equivalently, $\delta_p = \arg\max_{\delta \in C_{<\epsilon}} D_{KL}\left(T\left(\boldsymbol{x}_0+\delta\right) \| S\left(\boldsymbol{x}_0+\delta\right)\right)$. Note that

$$
\begin{aligned}
& D_{KL}\left(T\left(\boldsymbol{x}_0+\delta_t\right) \| S\left(\boldsymbol{x}_0+\delta_t\right)\right) \\
= & D_{KL}\left(T\left(\boldsymbol{x}_0+\delta_t\right) \| S\left(\boldsymbol{x}_0\right)\right) \geq D_{KL}\left(T\left(\boldsymbol{x}_0\right) \| S\left(\boldsymbol{x}_0\right)\right)
\end{aligned}
$$

[1]We omit θ when $S(\cdot)$ is deployed for evaluation for simplicity

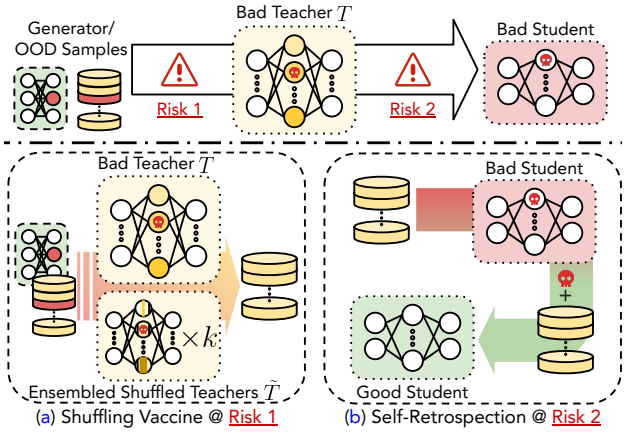

Figure 3. Risks of standard data-free KD with generator/OOD samples and the proposed ABD. Upper part: the two identified risks in data-free KD. Lower parts: the proposed ABD. (**a**) **Shuffling Vaccine** diminish the chance of bad input supplied for distillation; (**b**) student **Self-Retrospection** at a later stage of training to confront the potential learned backdoor from the teacher.

Therefore, there is a chance to generate δ_p from the above maximization, i.e., $P[\|\delta_t - \delta_p\| \leq \epsilon] > 0$. In other words, there is a potential risk associated with the input supplied to the teacher for distillation in data-free KD.

Potential Risk in Bad Supervision. On the other hand, the process of sampling data from OOD does not have the poisoned teacher's participation. However, we still find attacks that can infiltrate the teacher to the student via OOD-based KD. We hereby highlight another potential risk associated with the output logits of the teacher. That is, the returned soft labels may contain backdoor knowledge and thus lead to bad students.

4. Anti-Backdoor Data-Free KD

On observing the significant risks, we propose a plug-in anti-backdoor fixture for securing the existing data-free distillation method as formulated in Eq. (2). Our method is composed of two sequential strategies aimed to mitigate the two potential risks discussed in Section 3: **Shuffling Vaccine (SV)** before distillation optimization to diminish the chance of potential backdoored samples' participation and **Self-Retrospection (SR)** of the student at a later training stage of the student model to confront the bad supervision. An overview of our method and the relation to the two identified risks is illustrated in Fig. 3.

4.1. Shuffling Vaccine (SV)

Our method is inspired by the recent advance on backdoor model suspection (Cai et al., 2022). Previous work has revealed (Tran et al., 2018; Hayase et al., 2021) the sparse nature of backdoor activations. Most images will activate

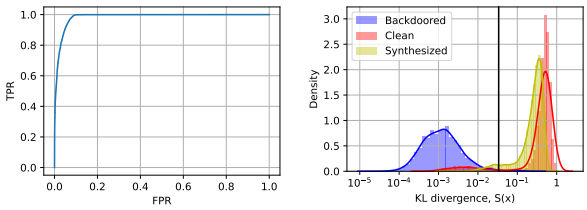

Figure 4. (a) ROC curve of $\mathcal{S}(x)$ colored by clean or backdoored samples. The corresponding AUC is 0.984. (b) Comparing $\mathcal{S}(x)$ where the black vertical line represents the 3σ boundary of the backdoored samples. A portion of the synthetic images falls into the danger zone.

different feature channels in deep layers of a network. In contrast, backdoor triggers will singly but significantly light a few channels such that other semantic features will be weakened layer by layer. As the backdoor activation is sparse, Cai *et al.* proposed *Channel Shuffling*, which amplified the nature by shuffling channels to suspect if a model is ever poisoned. The intuition is that the backdoor only relies on a few channels and shuffling may not destroy the connection. Instead, the prediction path for clean images will be ruined since a high ratio of semantic features will be compromised.

In this work, we novelly repurpose the *Channel Shuffling* to detect suspicious samples that may rely on some shortcuts in the networks. We hypothesize that if a sample can be stably predicted as one class under Channel Shuffling, then the sample is prone to be poisonous. Formally, we derive a shuffled model $\tilde{T}$ by shuffling the last few layers of T and then define a score metric as:

$$\mathcal{S}(x;\tilde{T}) = \log D_{KL}(\tilde{T}(x)\|T(x)),$$

where a smaller value indicates a higher risk of poisoning. In Fig. 4, we show that triggered samples have much lower $\mathcal{S}(x)$ than clean samples, showing that the metric is effective for detecting poison samples; Therefore, we apply the metric to suppress the generation or usage of suspicious samples, such that we can mitigate backdoor transfer in data-free distillation.

(1) Suppresing backdoor generation. For methods like ZSKT and CMI, distillation is built upon a synthetic dataset by contrasting the teacher and student models. Given a teacher model T and a shuffled model $\tilde{T}$, we define a new regularization term, $R(\boldsymbol{x};\tilde{T},T)$ on synthetic samples $\boldsymbol{x}$:

$$\max_{P} \mathbb{E}_{\boldsymbol{x}\sim P}\left[D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}\right)\right) + \alpha R(\boldsymbol{x};\tilde{T},T)\right],$$
$$R(\boldsymbol{x};\tilde{T},T) := \phi(T\left(\boldsymbol{x}\right)\|\tilde{T}\left(\boldsymbol{x}\right))D_{KL}\left(T\left(\boldsymbol{x}\right)\|\tilde{T}\left(\boldsymbol{x}\right)\right),$$

where $\phi(T\left(\boldsymbol{x}\right)\|\tilde{T}\left(\boldsymbol{x}\right))$ will yield 1 if the predicted labels of $T(\boldsymbol{x})$ and $\tilde{T}(\boldsymbol{x})$ are the same. Considering the randomness of shuffling, we use an ensemble of three shuffled teachers as $\tilde{T}$.

(2) Suppressing suspicious distillation. For OOD distillation, there is no way to control the sample generation and thus calls for a different defense method. One straightforward way is to drop the suspicious samples, which however may reduce the data size and result in overfitting. Instead, we introduce a soft constraint on the distillation to better trade-off model utility and security.

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}\sim P}\left[(1-\phi+\tfrac{1}{\alpha}\phi)D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}\right)\right)\right],$$

where ϕ is the output of $\phi(T\left(\boldsymbol{x}\right)\|\tilde{T}\left(\boldsymbol{x}\right))$. If ϕ is activated, then the sample loss will be shrunk by α.

4.2. Self-Retrospection (SR)

To mitigate the potential risk associated with bad supervision, we propose a post-hoc treatment to use the student's knowledge to confront potential backdoor knowledge that has been learned from the teacher. More specifically, we use the student's own knowledge to synthesize potential backdoor knowledge being learned and confront the model update from the teacher's supervision with the following Self-Retrospection (SR) task:

$$\theta^* = \arg\min_{\theta} \max_{\delta\in C_{<\epsilon}} \frac{1}{n}\sum_{i=1}^{n} D_{KL}\left(S\left(\boldsymbol{x}|\theta\right)\|S\left(\boldsymbol{x}+\delta|\theta\right)\right),$$

noting that δ in the outer loop is a function that depends on θ in the inner loop, i.e., $\delta(\theta)$. The intuition of the formulation is to synthesize a universal noise that will result in most of the samples' output logits greatly changed compared to the output of these samples without the noise. The amount of change is then depicted by the KL divergence, as a larger value of KL divergence depicts a stronger variation between the outputs with or without the noise being patched.

To resolve the proposed bi-level optimization, inspired by Zeng et al. (2022a); Rajeswaran et al. (2019), we approximate $\nabla\delta(\theta)$ with a suboptimal solution of δ^*. In particular, one can approximate $\nabla\delta(\theta)$ with an iterative solver of limited rounds (e.g., conjugated gradient algorithm (Rajeswaran et al., 2019), or fixed-point algorithm (Grazzi et al., 2020)) along with the reverse mode of automatic differentiation (Griewank & Walther, 2008). With a successful estimation of $\nabla\delta(\theta)$, we then can plug it into the process of computing the complete hypergradient for student SR:

$$\nabla\psi(\theta) = \nabla_2 D_{KL}(\delta(\theta),\theta) + (\nabla\delta(\theta))^{\top}\nabla_1 D_{KL}(\delta(\theta),\theta)$$

where we simplified $D_{KL}\left(S\left(\boldsymbol{x}|\theta\right)\|S\left(\boldsymbol{x}+\delta|\theta\right)\right)$ as a function to δ and θ, where δ is a variable dependent on θ, i.e., $D_{KL}(\delta(\theta),\theta)$, and $\nabla_1(\cdot)$ or $\nabla_2(\cdot)$ denotes the partial derivatives w.r.t. the first variable or the second variable respectively. We summarize the whole process of one

round of student SR in Algorithm 1, where our synthesized student SR hypergradeint is used to confront the original gradient acquired from $D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}|\theta\right)\right)$ for student update thus mitigates the potential risk of bad supervision.

Algorithm 1 One Round of KD with Self-Retrospection

Input: $T(\cdot)$ (Teacher model);
$S(\cdot;\theta)$ (Student model with parameters θ);
Parameters: n_δ (Number of steps);
$\eta, \gamma > 0$ (Step size);

$\mathcal{L}_S \leftarrow D_{KL}\left(T\left(\boldsymbol{x}\right)\|S\left(\boldsymbol{x}|\theta\right)\right)$
$\delta \sim \mathcal{N}(0, \sigma^2\mathbf{I}^d)$
for $1, 2, \ldots, n_\delta$ **do**
  $\mathcal{L}_\delta \leftarrow -D_{KL}\left(S\left(\boldsymbol{x}|\theta\right)\|S\left(\boldsymbol{x}+\delta|\theta\right)\right)$
  $\delta \leftarrow \delta - \gamma\frac{\partial\mathcal{L}_\delta}{\partial\delta}$
end
Estimate $\nabla\delta^\top$ by assuming δ is suboptimal with iterative solver
Compute $\nabla\tilde{\psi}(\theta)$ with $\nabla\delta^\top$ pluged in
$\theta \leftarrow \theta - \eta\left(\frac{\partial\mathcal{L}_S}{\partial\theta} + \nabla\tilde{\psi}(\theta)\right)$

4.3. Overall Pipeline

Vaccine verification and search. Due to the random nature of shuffling and backdoor mechanisms, there is a chance that the Shuffling Vaccine is not able to detect triggers. Therefore, we verify the functionality of Shuffling Vaccines before using them. The challenge of verification is lacking known clean and poisoned samples. For this purpose, we first run data-free KD to cache some surrogate data as set D_s and check if the $\mathcal{S}(x)$ distribution has a large tail. The intuition has been illustrated in Fig. 4. According to the three-sigma rule of thumb, a normal distribution should have 0.3% samples of values smaller than $\mu - 3\sigma$ where μ and σ are mean and standard deviations, respectively. Thus, we check the existence of the tail by computing the tail ratio, defined as

$$\tau(\tilde{T}; D_s) := \frac{|\{x \in D_s | \mathcal{S}(x;\tilde{T}) < \mu - 3\sigma\}|}{|D_s|},$$

where μ and σ denote the mean and standard deviation of $\{\mathcal{S}(x;\tilde{T})|x \in D_s\}$. We threshold $\tau(\tilde{T}; D_s)$ by 0.02 to choose a shuffle model with a large tail. If $\tilde{T}$ does not satisfy the condition, we will repeat shuffling for 8 times until giving up. The setting of the threshold is further ablated in Appendix D.

We summarize our whole pipeline in the Algorithm 2. We first use Shuffling Vaccines if a proper vaccine can be found. If a vaccine is not found, we will do normal data-free KD and use the student SR as a post-hoc treatment at a later learning stage when the student is well-trained till converged. If a vaccine is found, we may ask the user to determine if a sacrifice of clean accuracy is worth it for better security and activates the student SR on demand. In our setting, we activate SR if the clean accuracy drops lower than 5% using SV.

Time complexity analysis. SV is utilized to obtain an ensemble of effective shuffled models, and the forward pass of these models is used to suppress backdoor information. Compared to vanilla data-free KD for each epoch that includes SV, we introduce an additional $\tilde{O}(n \cdot \tilde{O}(\theta_T))$ time complexity, where $\tilde{O}(\theta_T)$ represents the time complexity of using the teacher model, θ_T, in a single forward pass on a batch of data. n is the number of shuffle models used in the ensemble. For SR, based on our Algorithm 1 design (total n_δ rounds) and assuming the fixed-point algorithm (Grazzi et al., 2020) as the iterative solver with ϑ iterations for computing $\nabla\tilde{\psi}(\theta)$, the time complexity is $\tilde{O}(n_\delta \cdot \vartheta \cdot \tilde{O}(\theta))$, where $\tilde{O}(\theta)$ is the time complexity of training the student model, θ, via backpropagation on a batch of data (similar to the forward pass of one epoch given the same quantity of samples (Zeng et al., 2022a)). In practice, we adopted $\vartheta = 5$, and for most one-target attack cases, $n_\delta = 10$ is sufficient for algorithm convergence. Both techniques only introduce linear additional computational costs on the order of the size of the teacher or student model. In practice, we find in our experiment, the overall ZSKT+ABD empirical time cost with our settings on the WRNs to be only 1.03 times higher than the vanilla ZSKT, evaluated on CIFAR-10 with ZSKT and BadNets (grid) trigger.

Algorithm 2 Anti-Backdoor Data-Free KD (ABD)

Input: $T(\cdot)$ (Teacher model);
$S(\cdot;\theta)$ (Student model with parameters θ);
Parameters: λ (Starting step for student SR);

Synthesize or obtain a set of OOD samples D_s
Search for $\tilde{T}$ at most 8 trials
if *Found effective* $\tilde{T}$ **then**
  `/* 1. Early Prevention with SV */`
  Data-free KD with SV till step λ
else
  Data-free KD till step λ
end
`/* 2. Later Treatment with SR */`
if *Activates Student SR* **then**
  Data-free KD with student SR
end

5. Experiment

In this section, we evaluate how the proposed ABD can secure data-free KD against backdoor attacks under various data, model, and trigger configurations.

Datasets and models. We use the same datasets, CIFAR-10 (Krizhevsky et al., 2009) and GTSR-B (Stallkamp et al., 2012), as (Zeng et al., 2022a) to evaluate the backdoor defenses. Following the setup of ZSKT (Micaelli & Storkey, 2019), we use WideResNet (Zagoruyko & Komodakis, 2016) for training 10-way or 43-way classifiers on CIFAR-10 and GTSR-B, respectively. We use WRN-16-2 to denote a 16-layer WideResNet with a width factor of 2.

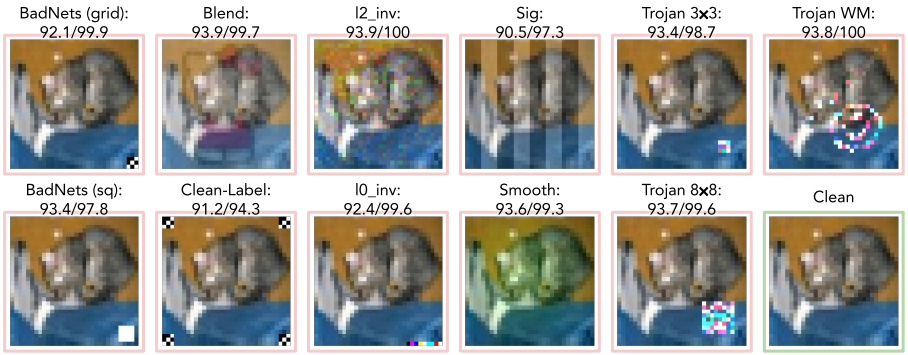

Figure 5. Trigger visualization and teacher model performances on CIFAR-10. The performance (Acc/ASR) of the poisoned teacher using each backdoor attack is provided beneath each trigger's name. We envision the backdoored example for each attack on CIFAR-10.

Backdoor attacks. Prior to distillation, we pre-train a teacher model on a poisoned training dataset and use data-free distillation methods to train a student under the soft supervision of the pre-trained teacher model. The poisoned pre-training dataset contains 10% samples with backdoors injected by different attack manners. For example, the BadNets attack (Gu et al., 2019) injects grid or square patterns to the corner of an image, which are denoted as BadNets (grid) or (sq). Examples of triggers are presented in Fig. 5.

Evaluation metrics. Following the common practice on backdoor defense, e.g., (Li et al., 2020b; 2021b; Zeng et al., 2022a), we use attack success rate (ASR) and clean accuracy (Acc) as the measures evaluating distillation methods. ASR is defined as the portion of backdoored test samples that can successfully mislead the model to predict the target class specified by the attacker. Acc is the classification accuracy measured on a clean test set. A favored method should present a smaller ASR and meanwhile a larger Acc.

Distillation methods. We use ZSKT (Micaelli & Storkey, 2019), CMI (Fang et al., 2021), and OOD (Asano & Saeed, 2021) as the baseline distillation methods. We use 20% clean data for vanilla knowledge distillation (Hinton et al., 2015), denoted as Clean KD. We follow previous work to use their published codes[2][3][4] and hyperparameters. More details in Appendix A.

Defending multiple types of attacks. To evaluate the effectiveness of the proposed defending method, we construct a benchmark against different backdoors. We use WRN16-2 as the teacher and WRN16-1 as the student to predict image classes in CIFAR-10. All the teachers are trained and selected based on the best test accuracy on clean images.

[2] https://github.com/polo5/ZeroShotKnowledgeTransfer
[3] https://github.com/zju-vipa/CMI
[4] https://github.com/yukimasano/single-img-extrapolating

Trigger	Teacher Acc/ASR	Student Acc/ASR		
		ZSKT	ZSKT+ABD	Clean KD
BadNets (grid)	92.1/99.9	71.9/96.9	68.3/0.7	74.6/4.3
Trojan WM	93.8/100	82.7/93.9	78.2/22.5	77.5/11.1
Trojan 3x3	93.4/98.7	80.9/96.8	71.7/33.3	72.9/1.7
Blend	93.9/99.7	77.0/74.4	71.5/23.1	78.0/4.3
Trojan 8x8	93.7/99.6	80.5/57.2	72.6/17.8	75.2/9.3
BadNets (sq)	93.4/97.8	80.8/37.8	77.9/1.9 (s)	76.2/9.1
CL	91.2/94.3	76.8/17.5	67.4/10.2	69.4/2.1
Sig	90.5/97.3	77.9/0.0	72.2/0. (s)	77.4/0.
l2_inv	93.9/100	82.0/0.3	70.7/1.9 (s)	77.2/1.2
l0_inv	92.4/99.6	72.8/8.3	69.4/0. (s)	79.2/3.7

Table 1. Evaluation of data-free distillation on more triggers on CIFAR-10 with WRN16-2 (Teacher) and WRN16-1 (student). (s) indicates Shuffling Vaccine is used instead of student SR.

The results are summarized in Table 1. In most triggers, our method effectively treats or protects the ZSKT from backdoor transfer successfully by reducing ASR lower than 30%. Since there is no free lunch for removing backdoors, reducing ASR also results in lower clean accuracy. Especially, without clean data from the training set of teacher models, removing backdoors is even harder to maintain accuracy as compared to the clean KD. This is because, without data from the same distribution, it is hard to distinguish which kinds of features are needed for benign tasks or backdoor tasks. To effectively suppress the risks of backdoors, the degradation of clean accuracy is the essential cost.

Data-free distillation has almost-free resilience to some backdoors. In Table 1, we observe that many triggers are not strong enough to transfer without data. The failure happens when the triggers are not localized to a small region but spatially spreading, e.g., Sig, CL, and l2_inv. Noticeably, the CL trigger relies on the adversarial samples to transfer, which fails with smoothed decision boundaries defined by the teachers' soft labels (Yuan et al., 2020). Remarkably, the natural resilience is almost free for distillation, since the distillation does not significantly reduce accuracy compared to the strong transferred cases.

Dataset	Teacher Arch (size)	Student Arch (size)	Teacher Trigger	Teacher Acc/ASR	Student Acc/ASR ZSKT	+ABD	Clean KD
GTSR-B	WRN16-2 (0.7MB)	WRN16-1 (0.2MB)	BadNets (grid)	88.1/98.8	87.0/99.5	78.4/13.0	89.8/0.3
CIFAR-10	WRN16-2 (0.7MB)	WRN16-1 (0.2MB)	BadNets (grid)	92.1/99.9	71.9/96.9	68.3/0.7	74.6/4.3
	WRN16-2 (0.7MB)	WRN16-1 (0.2MB)	Trojan WM	93.8/100	82.7/93.9	78.2/22.5	77.5/11.1
	WRN40-2 (2.2MB)	WRN16-1 (0.2MB)	BadNets (grid)	94.5/100	84.2/4.6	76.9/10.7 (s)	72.0/4.7
	WRN16-2 (0.7MB)	WRN16-1 (0.2MB)	Trojan WM	94.5/100	87.6/54.5	82.9/5.8 (s)	71.2/5.3

Table 2. Evaluation of anti-backdoor data-free distillation on different datasets and different model architectures. '(s)' indicates Shuffling Vaccine is used instead of the student's Self-Retrospection.

Architectures and datasets in distillation. In Table 2, we evaluate our method on defending ZSKT against BadNets (grid) and Trojan WM on two datasets and two teacher architectures. There are several intriguing observations. (1) Except for the BadNets from WRN40-2 teacher, all the triggers successfully transfer from the teacher to the student. (2) We notice that the transfer is less effective when the teacher has deeper layers, comparing the WRN40-2 versus WRN16-2 teachers on CIFAR-10. In such under-transfer cases, the anti-backdoor ZSKT even outperforms the clean KD. This may imply that a deeper and over-parameterized model may be more robust in transferring clean knowledge than a few really-clean samples. (3) In all cases, our method effectively defends ZSKT against the tested backdoor attacks. Our method can maintain higher clean accuracy on CIFAR-10, which is composed of more complicated features than the traffic signs in GTSR-B. This observation is surprisingly different from the one in central defense, e.g., in (Zeng et al., 2022a) where CIFAR-10 is a harder dataset to defend. The rationale behind the difference is that GTSR-B encodes simple features sparsely in the same network leaving more space for the trigger features. Therefore, adversarial training by ZSKT will rely more on these triggers to transfer knowledge. Instead, CIFAR-10 models can squeeze out these features to maintain good performance on clean images.

Distillation Method	Teacher Trigger	Teacher Acc/ASR	Student Acc/ASR Baseline	+ABD
ZSKT	Trojan WM	93.8/100	82.7/93.9	78.2/22.5
	BadNets (grid)	92.1/99.9	71.9/96.9	68.3/0.7
CMI	Trojan WM	93.8/100	89.1/99.0	79.8/8.0
	BadNet (sq)	93.8/100	88.3/95.9	83.2/6.0
OOD	Trojan WM	93.8/100	82.3/100	62.3/21.8
	BadeNet (grid)	92.1/99.9	79.8/99.6	78.2/14.5

Table 3. ABD is effective in different data-free distillation methods on CIFAR-10 with WRN16-2 (Teacher) and WRN16-1 (student).

Protecting distillation with different surrogate data. As the data-free knowledge transfer heavily relies on the surrogate data for distillation, we investigate the resilience of different distillation methods and evaluate the backdoor defense on the failure cases. In Fig. 2, we have witnessed the backdoors can transfer successfully through ZSKT, CMI, and OOD. In Table 3, we compare them on the selected transferred triggers. With the same trigger of Trojan WM, all three distillation approaches have >90% ASR. Similar to ZSKT, CMI uses the adversarial loss to find the underfitted samples and therefore transfers knowledge by distilling these synthetic images. Different from ZSKT, CMI introduces more objectives in data optimization to improve the quality of synthetic data, as visually compared in Fig. 6. Unfortunately, improving visual quality cannot eliminate triggers even these triggers could be thought as visually-low-quality features. Interestingly, our method can maintain higher clean accuracy on CMI compared to ZSKT and OOD. Among the three distillations, CMI has traded the least amount of benign accuracy for lower ASR. This implies that image quality is essential for benign accuracy though not robustness.

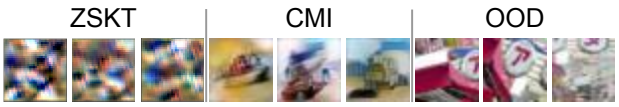

Figure 6. Examples of ZSKT and CMI synthetic images. OOD images are patches of a single large image.

Compared to ZSKT, OOD is more vulnerable to the distributed trigger, Trojan WM, but is more robust against Badnet (grid), by using our defense. This may be attributed to that the local BadNets trigger is less likely to be found in augmented real images than adversarially-generated ones by ZSKT. Instead, Trojan WM has a pattern similar to random noise, that can be approximated by some random pattern in some image augmentations.

Ablation study. A natural question for the proposed defense may be *why we need student SR as a post-hoc treatment, especially when we can use the Shuffling Vaccine.* In Table 4, we compare the model performance by the ablation of Shuffling Vaccine and student Self-Retrospection. We show that shuffling may fail on specific random states and SR can successfully salvage the student from the failure. Therefore, we conclude that the Shuffling Vaccine may fail

on some triggers and require careful selection using synthetic data. When the vaccine fails, the student SR can salvage the case by exploring the strong signal of triggers. Notice that the student SR is less effective if the trigger itself is not strong enough. For example, BadNet has a relatively low ASR without any defense, while SR can only reduce the ASR to 76.2%. This is because the trigger is not the strongest noise-biasing model prediction, and therefore it has a lower chance to be synthesized by the student SR.

SV	SR	BadNets (grid)	Trojan WM
		70.7/87.8	82.7/93.9
✓		67.2/**0.3**	79.0/57.0
	✓	68.3/76.2	79.7/44.1
✓	✓	68.3/**0.7**	78.2/**22.5**
Clean KD		74.6/4.3	77.5/11.1

Table 4. Ablation study of components on CIFAR-10 with two triggers. Report results as Acc/ASR. To show the failure of shuffling, we disable the selection of shuffling here.

Shuffling Vaccine in learning. In Fig. 7, we show how the vaccine suppresses the poisons during distillation and how it fails. To show the failure case of Shuffling Vaccines, we do not make selections here. When defending against BadNets (grid), the Shuffling Vaccine is able to compromise the ASR to almost 0 at the very beginning, which process almost does not affect the convergence on the clean accuracy. On the failure of Shuffling Vaccine, the student Self-Retrospection is applied at the last 5 epochs, which effectively suppresses the Trojan WM backdoors.

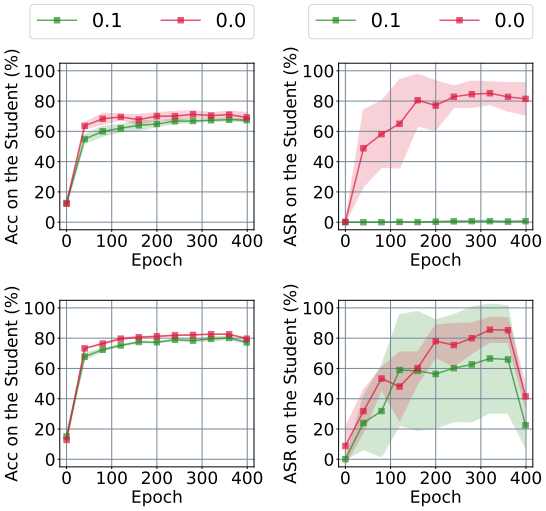

Figure 7. Ablation of the Shuffling Vaccine. The upper two figures are BadNets (grid) and the below two are Trojan WM.

6. Conclusion and Discussions

In this work, we make the first effort to reveal the security risk of data-free KD w.r.t. the untrusted pre-trained models. To mitigate the chance of potential backdoors in the synthetic or OOD data being transferred, we propose ABD, the first plug-in defensive method for data-free KD methods. We empirically demonstrate that ABD can diminish transferred backdoor knowledge while maintaining comparable downstream performances as the vanilla KD.

Limitation and broader impact. We believe our findings could be inspiring for the research in making data-free KD secure and the downstream student trustworthy. As a pilot work, we expect more interesting studies on understanding the poisoning mechanism behind the data-free distillation and defending multiple triggers (Cai et al., 2022) and an ensemble of teacher models for federated learning (Zhang et al., 2022a). When pre-training data are unavailable, a data-free detection strategy could be desired to determine whether the model is ever poisoned. Except for distillation, it is also intriguing to study if pre-trained models can be poisonous under robust training (Hong et al., 2021). Beyond the scope of this paper, the *dark side* of our backdoor mitigation could be the risk on backdoor-based Intelligence-Property (IP) protection (Jia et al., 2020; Li et al., 2019). In IP protection, a backdoor is injected for later verifying the ownership of an unauthorized model distillation. The effectiveness of our mitigation indicates that removing the IP backdoor is possible even without training data.

Acknowledgments

This work is supported partially by Sony AI, NSF IIS-2212174 (JZ), IIS-1749940 (JZ), NIH 1RF1AG072449 (JZ), ONR N00014-20-1-2382 (JZ), a gift and a fellowship from the Amazon-VT Initiative. We also thank anonymous reviewers for providing constructive comments. In addition, we want to thank Haotao Wang from UT Austin for his valuable discussion when developing the work.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. The medical segmentation decathlon. *Nature communications*, 13(1): 1–13, 2022.

Asano, Y. M. and Saeed, A. Extrapolating from a single image to a thousand classes using distillation. *arXiv preprint arXiv:2112.00725*, 2021.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to backdoor federated learning. In

International Conference on Artificial Intelligence and Statistics, pp. 2938–2948. PMLR, 2020.

Barni, M., Kallas, K., and Tondi, B. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 101–105. IEEE, 2019.

Beyer, L., Zhai, X., Royer, d., Markeeva, L., Anil, R., and Kolesnikov, A. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10925–10934, 2022.

Cai, R., Zhang, Z., Chen, T., Chen, X., and Wang, Z. Randomized channel shuffling: Minimal-overhead backdoor attack detection without clean datasets. In *Advances in Neural Information Processing Systems*, 2022.

Chawla, A., Yin, H., Molchanov, P., and Alvarez, J. Data-free knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3289–3298, 2021.

Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., and Tian, Q. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3514–3522, 2019.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fang, G., Song, J., Wang, X., Shen, C., Wang, X., and Song, M. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*, 2021.

Fang, G., Mo, K., Wang, X., Song, J., Bei, S., Zhang, H., and Song, M. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6597–6604, 2022.

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pp. 3748–3758. PMLR, 2020.

Griewank, A. and Walther, A. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Hayase, J., Kong, W., Somani, R., and Oh, S. Spectre: defending against backdoor attacks using robust statistics. In *ICML*, 2021.

Hinton, G., Vinyals, O., Dean, J., et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Hong, J., Wang, H., Wang, Z., and Zhou, J. Federated robustness propagation: Sharing robustness in heterogeneous federated learning. *arXiv preprint arXiv:2106.10196*, 2021.

Jia, H., Choquette-Choo, C. A., Chandrasekaran, V., and Papernot, N. Entangled watermarks as a defense against model extraction. *arXiv preprint arXiv:2002.12200*, 2020.

Jia, J., Liu, Y., and Gong, N. Z. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059. IEEE, 2022.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. Attribute and simile classifiers for face verification. In *ICCV*, pp. 365–372. IEEE, 2009.

Li, S., Xue, M., Zhao, B., Zhu, H., and Zhang, X. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE TDSC*, 2020a.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2020b.

Li, Y., Wu, B., Jiang, Y., Li, Z., and Xia, S.-T. Backdoor learning: A survey. *arXiv:2007.08745*, 2020c.

Li, Y., Li, Y., Wu, B., Li, L., He, R., and Lyu, S. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021a.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. In *NeurIPS*, volume 34, 2021b.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021c.

Li, Z., Hu, C., Zhang, Y., and Guo, S. How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of dnn. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 126–137, 2019.

Liu, H., Jia, J., and Gong, N. Z. Poisonedencoder: Poisoning the unlabeled pre-training data in contrastive learning. *arXiv preprint arXiv:2205.06401*, 2022.

Liu, Y., Ma, S., Aafer, Y., Lee, W.-C., Zhai, J., Wang, W., and Zhang, X. Trojaning attack on neural networks. In *NDSS*, 2018.

Micaelli, P. and Storkey, A. J. Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems*, 32, 2019.

Nguyen, T. A. and Tran, A. T. Wanet-imperceptible warping-based backdoor attack. In *ICLR*, 2020.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. Deep face recognition. 2015.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Seo, H., Park, J., Oh, S., Bennis, M., and Kim, S.-L. Federated knowledge distillation. *Machine Learning and Wireless Communications*, pp. 457, 2022.

Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., Raman, N., Smiley, C., Chen, J., and Yang, D. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*, 2022.

Souri, H., Goldblum, M., Fowl, L., Chellappa, R., and Goldstein, T. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *arXiv:2106.08970*, 2021.

Stallkamp, J., Schlipsing, M., Salmen, J., and Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32: 323–332, 2012.

Tan, A. Z., Yu, H., Cui, L., and Yang, Q. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Tang, Z., Zhang, Y., Shi, S., He, X., Han, B., and Chu, X. Virtual homogeneity learning: Defending against data heterogeneity in federated learning. In *International Conference on Machine Learning*, 2022.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. In *NeurIPS*, pp. 8000–8010, 2018.

Turner, A., Tsipras, D., and Madry, A. Label-consistent backdoor attacks. *arXiv:1912.02771*, 2019.

Wang, H., Hong, J., Zhang, A., Zhou, J., and Wang, Z. Trap and replace: Defending backdoor attacks by trapping them into an easy-to-replace subnetwork. *arXiv preprint arXiv:2210.06428*, 2022a.

Wang, T., Yao, Y., Xu, F., An, S., Tong, H., and Wang, T. An invisible black-box backdoor attack through frequency domain. In *ECCV*, 2022b.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022c.

Wu, C., Wu, F., Lyu, L., Huang, Y., and Xie, X. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):2032, 2022.

Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020.

Ye, J., Ji, Y., Wang, X., Gao, X., and Song, M. Data-free knowledge amalgamation via group-stack dual-gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12516–12525, 2020.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yuan, L., Tay, F. E., Li, G., Wang, T., and Feng, J. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In Richard C. Wilson, E. R. H. and Smith, W. A. P. (eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6.

Zeng, Y., Park, W., Mao, Z. M., and Jia, R. Rethinking the backdoor attacks' triggers: A frequency perspective. In *ICCV*, 2021.

Zeng, Y., Chen, S., Park, W., Mao, Z., Jin, M., and Jia, R. Adversarial unlearning of backdoors via implicit hypergradient. In *ICLR*, 2022a.

Zeng, Y., Pan, M., Just, H. A., Lyu, L., Qiu, M., and Jia, R. Narcissus: A practical clean-label backdoor attack with limited information. *arXiv:2204.05255*, 2022b.

Zhang, J., Chen, C., Li, B., Lyu, L., Wu, S., Ding, S., Shen, C., and Wu, C. Dense: Data-free one-shot federated learning. In *Advances in Neural Information Processing Systems*, 2022a.

Zhang, L., Shen, L., Ding, L., Tao, D., and Duan, L.-Y. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10174–10183, 2022b.

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., and Wu, Y. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

Zhu, Z., Hong, J., and Zhou, J. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, pp. 12878–12889. PMLR, 2021.

Zhu, Z., Hong, J., Drew, S., and Zhou, J. Resilient and communication efficient learning for heterogeneous federated systems. In *Proceedings of Thirty-ninth International Conference on Machine Learning (ICML 2022)*, 2022.

A. Experimental Details

In this section, we provide details of hyper-parameters. To verify shuffling models, we cache 50 batches for ZSKT and 100 batches on OOD as D_s. Shuffling Vaccine is done by randomly changing the order of channels in the last 5 convolutional layers of WideResNet (corresponding to the last stage) and an ensemble of three shuffled models is used. If SV significantly degrades the clean accuracy, we will restart the distillation without SV. The Self-Retrospection treatment is done at the last 3 epochs of CMI/OOD, 800 batches of ZSKT.

For ZSKT on GTSRB, we tune the KL temperature until maximizing the student's clean accuracy. The preferred temperature will be 0.5, and we remove the feature alignment, which yields better accuracy.

For OOD, we use the pre-sliced $10,000$ patches provided by the authors and augment the patches by random CutMix with 100% probability and $\beta = 0.25$ for the Beta-sampling.

For CMI, we directly use the hyper-parameter set provided by the authors for distillation from WRN16-2 to WRN16-1 on CIFAR10. As commented by the authors, these parameters are very sensitive, and therefore, we do not change them.

All the defense experiments are repeated three times using the seed set $\{0, 1, 2\}$. For pre-training backdoored teachers, we use the published codes (Wang et al., 2022a)[5]. Note that the codes do not normalize the input, but we follow the common practice to normalize the CIFAR10 and GTSRB inputs, following (Zeng et al., 2022a).

B. The Prevalence of the Risk

We further demonstrate the risk of backdoor infiltrating from the teacher model to the student via data-free KD under a different dataset setting, i.e., using the GTSRB dataset. As shown in Figure 8, the observation under the GTSRB-dataset setting is consistent with Figure 2. Beyond the prevalence of the risk, it seems the dataset with lower sample diversity (compared to CIFAR-10, GTSRB data points are traffic signs taken from almost the same angle, which of lower sample diversity within each class) suffers with a higher risk (all the evaluated attacks are transferred to the student under the GTSRB setting) of backdoor being transferred via data-free KD. In addition, we examine the transfer from ResNet34 to ResNet18 on PubFig dataset (Kumar et al., 2009). The ASR can approach 88.6% at the end using trojan_wm trigger when benign accuracy reaches 87.1% after 10,000 epochs.

[5] https://github.com/VITA-Group/Trap-and-Replace-Backdoor-Defense

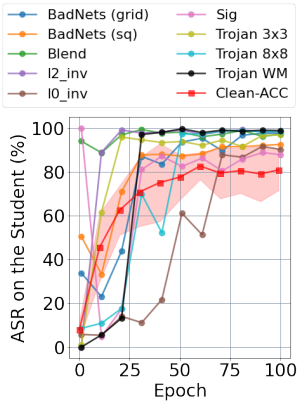

Figure 8. The risk of backdoor infiltrating the student model on the GTSRB-dataset settings (ZSKT as the data-free KD method).

C. Plausible Understanding of the Security Risk w.r.t. Data-free Settings

Why does data-free not lead to poison-free? This question is non-trivial since the distillation samples are either generated via an additional generative network or sampled from OOD; see examples from Fig. 6, which visually do not contain the initial triggers. As most existing backdoor attacks require poisoning of the training samples, it is unclear the main cause of the transfer of backdoor knowledge under data-free settings. As the key difference between vanilla KD and data-free KD methods is the participants of synthetic or OOD samples that of lower confidence w.r.t. the output logits of the teacher, one possible reason is that some of these low-confidence points activate similar neurons of the poisoned teacher as the initial poisoned samples, thus leading to the backdoor knowledge being transferred. An intuition of the presumption is depicted in Fig. 9.

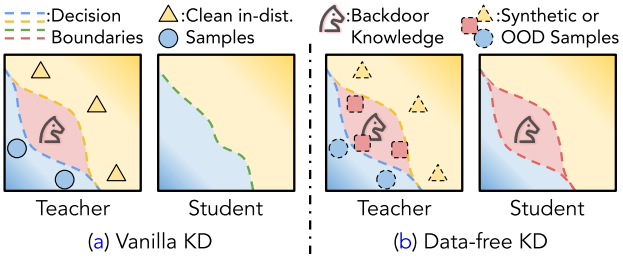

Figure 9. The difference between vanilla KD and data-free KD. The use of synthetic and OOD samples may be of low confidence w.r.t. all vicinity classes (**blue** and **yellow** region) thus activating the backdoor knowledge from the poison teacher (**red** region, which previously cannot be activated via clean in-distribution samples) and leading to the transfer of backdoor knowledge.

Following this idea, we further explore if there's a direct observation indicating that synthetic samples or OOD can activate backdoor knowledge. We train a logistic regression

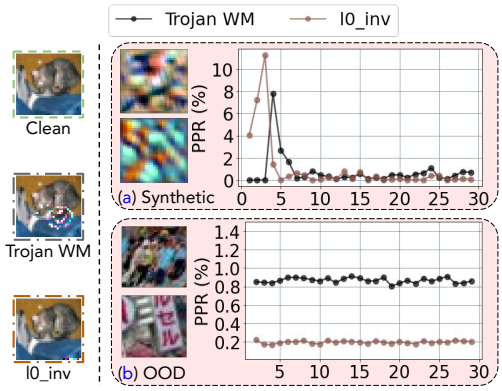

Figure 10. An empirical study of whether synthetic and OOD samples directly activate backdoor knowledge. The chance of a synthetic data point being classified as a poisoned sample is interpreted as the Plausible Poison Ratio (PPR). The left-hand side shows the clean and poisoned examples that were used for this experiment. The right-hand side depicts the visual examples and PPR results of KD methods based on (a) synthetic or (b) OOD samples, respectively.

classifier that takes the teacher model's output logits as the input to see if synthetic and OOD data may activate similar neurons as poisoned samples and how the portion of the sample may affect the transfer of backdoor knowledge. To start, we use the test set of CIFAR-10 to obtain 9,000 output logits of poisoned samples (patched with the initial trigger and labeled as 1) and 10,000 output logits of clean samples (labeled as 0). We then train the logistic regression classifier with the above two categories of logits and apply it to unseen synthetic/OOD samples' output logits from the same epoch and measure the false positive rate of synthetic/OOD data being classified as positioned samples (plausible poison ratio, or PPR). We depict the analysis of the experiment on the Trojan WM (example of an attack that always infiltrates the student) and l0_inv (example of an attack that cannot infiltrate the student) in Figure 10. The insight on OOD samples is quite clear, where we find OOD samples can activate similar neurons as the initial poisoned samples on success attack case (Trojan WM) with 3 to 4 $\times$ higher PPR than the failed attack (l0_inv). This, in a way, aligns with our presumption that the backdoor knowledge is activated with input samples being similar to the initial poisoned data. However, the results of Synthetic data is hard to come to the same conclusion. We suspect that the reason why backdoor attacks can infiltrate data-free KD based on synthetic data is more complicated, and we defer it to future work of exploration.

D. Ablation Study on SV's Threshold

In Fig. 11, we conduct an ablation study on the threshold for selecting shuffle models. We conduct experiments using

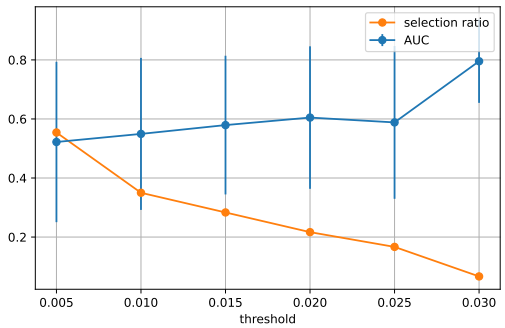

Figure 11. Ablation study on the threshold.

models with Trojan WM and BadNet (grid). We compute the backdoor detection AUC when varying the threshold. A higher AUC is desired for accurately recognizing poison samples. We also report the ratio of shuffle models selected in the sample set. A higher selection ratio means that it is easy to find a desired shuffled model in a limited sample set and therefore is more efficient. In each case, we sample 20 shuffle models. As shown in Fig. 11, the threshold 0.02 can strike a balance between high AUC and a reasonable selection ratio. By the selection ratio 0.22, we only need approximatly 5 samples to find an effective shuffle model in expectation, which is smaller than our sample size 8.