How to Backdoor the Knowledge Distillation

Chen Wu
Computer Science and Engineering
Pennsylvania State University
cvw5218@psu.edu

Qian Ma
Information Science and Technology
Pennsylvania State University
qfm5033@psu.edu

Prasenjit Mitra
Information Science and Technology
Pennsylvania State University
pum10@psu.edu

Sencun Zhu
Computer Science and Engineering
Pennsylvania State University
sxz16@psu.edu

Abstract

Knowledge distillation has become a cornerstone in modern machine learning systems, celebrated for its ability to transfer knowledge from a large, complex teacher model to a more efficient student model. Traditionally, this process is regarded as secure, assuming the teacher model is clean. This belief stems from conventional backdoor attacks relying on poisoned training data with backdoor triggers and attacker-chosen labels, which are not involved in the distillation process. Instead, knowledge distillation uses the outputs of a clean teacher model to guide the student model, inherently preventing recognition or response to backdoor triggers as intended by an attacker. In this paper, we challenge this assumption by introducing a novel attack methodology that strategically poisons the distillation dataset with adversarial examples embedded with backdoor triggers. This technique allows for the stealthy compromise of the student model while maintaining the integrity of the teacher model. Our innovative approach represents the first successful exploitation of vulnerabilities within the knowledge distillation process using clean teacher models. Through extensive experiments conducted across various datasets and attack settings, we demonstrate the robustness, stealthiness, and effectiveness of our method. Our findings reveal previously unrecognized vulnerabilities and pave the way for future research aimed at securing knowledge distillation processes against backdoor attacks.

1 Introduction

Knowledge distillation (KD) has emerged as a fundamental technique in modern machine learning, enabling the transfer of knowledge from a large, complex teacher model to a more efficient student model without significant loss of performance [1]. This process has been instrumental in deploying deep learning models on resource-constrained devices, such as mobile phones and embedded systems [2], by reducing computational requirements while maintaining high accuracy. The knowledge transfer has been extended to other tasks, such as mitigating adversarial attacks [3], private model compression [4], knowledge communication in the heterogeneous federated learning [5, 6, 7].

Traditionally, the security of knowledge distillation has been presumed robust, especially when the teacher model is clean and trustworthy. This confidence stems from the nature of conventional backdoor attacks, which typically rely on poisoning the training data with backdoor triggers and attacker-specified labels [8]. Even when training data is polluted by the attacker with backdoored input samples and altered output labels, the implant of the backdoor behavior to the student model cannot be guaranteed. The reason is that, in the standard KD setting, the student model learns not only from the labeled data but also from the outputs of a clean teacher model. The clean teacher model will not react to the backdoored input samples and does not generate the desired altered output expected by the attacker. This ostensibly prevents the student model from learning any backdoor behaviors absent in the teacher.

However, this assumption overlooks the potential vulnerabilities naturally exist in almost all deep neural network models [9] - the adversarial examples. A clean (backdoor-free) teacher model can still predict the carefully designed adversarial input examples as another label. This technique can make small perturbations to the original input images that remain almost imperceptible to the human vision system to be wrongly predicted by a deep neural network model [10]. At the same time, the distillation process heavily relies on the teacher model's output to train the student model, since the student's ouputs are supposed to align with the teacher's. If this training dataset is compromised, it opens a novel attack vector that can be exploited without altering the teacher model. Despite the critical implications, the security risks associated with poisoning the distillation dataset in the context of a clean teacher model remain under-explored.

In this paper, we challenge the prevailing assumption of inherent security in knowledge distillation by introducing a novel adversarial backdoor attack. Our method strategically poisons the distillation dataset with adversarial examples embedded with backdoor triggers. Unlike traditional backdoor attacks, our approach does not require manipulating the training (distillation) process or compromising the teacher model. Instead, it subtly alters the distillation data to induce the student model to learn the backdoor behavior through the knowledge distillation process.

Our contributions are summarized as follows:

- **Novel Attack Methodology**: We propose the first adversarial backdoor attack that successfully implants backdoor behaviors into the student model during knowledge distillation with a clean teacher model. This challenges the assumption that knowledge distillation is secure when the teacher model is uncompromised.
- **Extensive Experimental Validation**: We conduct comprehensive experiments across various datasets (e.g., CIFAR-10, CIFAR-100) and attack settings to evaluate the effectiveness, robustness, and stealthiness of our proposed attack. The results demonstrate high attack success rates with minimal impact on the student model's performance on clean data.
- **Revealing Unrecognized Vulnerabilities**: Our findings expose previously unrecognized vulnerabilities in the knowledge distillation process, highlighting the need for revised security considerations and the development of robust defense mechanisms against such attacks.

2 Related Work

2.1 Adversarial Attacks

Deep learning networks perform a wide variety of computer vision tasks with remarkable performances. However, as discovered by Szegedy et al. [10], they are susceptible to adversarial attacks in the form of small perturbations to images that remain almost imperceptible to the human vision system. Such attacks can cause the neural network classifiers to completely change their predictions about the image. This vulnerability of the deep neural networks serves as the cornerstone of our proposed backdoor attack method.

In the scope of this paper, we have utilized but not limited to the following popular adversarial attack method. The Fast Gradient Sign Method (FGSM) was proposed by [11] to efficiently compute an adversarial perturbation for a given image by solving the cost function to increase the loss of the classifier on the resulting image. The Basic Iterative Method (BIM) is an iterative adversarial attack proposed by [12]. This method extends the one-step method of perturbing images by taking a single large step in the direction that increases the loss of the classifier to iteratively taking multiple small steps and adjusting the direction after each step. The CW attack was proposed by [13] to fight against the defensive distillation against adversarial perturbations [3]. It was a set of three adversarial attacks making the perturbations quasi-imperceptible by restricting their ℓ_2, ℓ_∞ and ℓ_0 norms.

2.2 Knowledge Distillation

The idea of model compression was first proposed by [14] to transfer the information from a large model or an ensemble of models into a small model without a significant drop in accuracy. This learning process was later formally popularized as knowledge distillation by [1]. The main idea of this process is to let the student model mimic the teacher model behaviors in order to obtain a competitive or even superior performance. Inspired by this idea, recent works have extended the knowledge distillation method in different applications. Mutual learning was proposed by [15], and it allows an ensemble of students to learn collaboratively and teach each other throughout the training process. [16] found that the student network performance degrades when the gap between student and teacher is large. They introduced multi-step knowledge distillation, which employed an intermediate-sized network to bridge the gap between the teacher and student model. Furthermore, the knowledge transfer has been extended to other tasks, such as mitigating adversarial attacks [3], private model compression [4], knowledge communication in the heterogeneous federated learning [5, 6, 7]. Researchers [17] have also leveraged only the unlabeled data to prevent the transmission of backdoor behaviors from the polluted teacher model to the student model. At the same time, some researchers [18] studied the security risk of data-free knowledge distillation that transfers backdoors from a poisoned teacher model to a student model. However, to

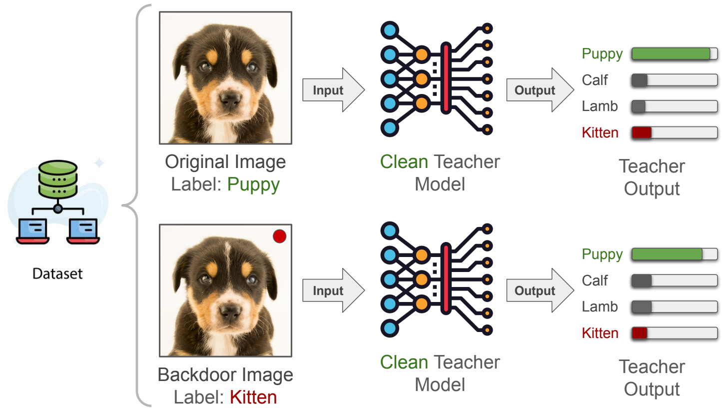


Figure 1: Clean teacher model will predict the correct label instead of the backdoored label given the backdoored image.

the best of our knowledge, no existing work has studied the security risk of knowledge distillation with a clean teacher model, which is the focus of this work.

3 Methodology

3.1 Background and Problem Definition

The knowledge distillation process typically allows the student model to mimic the teacher model behaviors to obtain the "knowledge" from the teacher model. Existing works [18] have explored the vulnerabilities of this process to propagate malicious behaviors (e.g., backdoors) from a bad teacher model. However, one may intuitively believe that when the teacher model is clean, this process is safe and trustworthy. In this work, we reveal the vulnerabilities behind this widely used knowledge distillation process and demonstrate that we can implant backdoor behaviors to the student model even with a clean teacher model.

We define the threat model as given a well-trained clean teacher model M_t, a randomly initialized student model M_s, and an open source dataset D. From the attacker's perspective, the goal is to inject any designed backdoor behaviors into the student model M_s. This backdoor behavior is described as that for any data x that should be correctly classified as label y will be incorrectly classified as another label y' after we add a backdoor pattern p to the original data x. So, the backdoored model will classify $x' = x + p$ as y' while keeping classifying x as y.

The knowledge distillation training process tries to match the softened outputs of student $y_s = softmax(a_s/\tau)$ and teacher $y_t = softmax(a_t/\tau)$ via a Kullback-Leibler divergence loss [19].

$$\mathcal{L}_{KD} = \tau^2 KL(y_s, y_t)$$

where the hyperparameter temperature τ is introduced to add more control on the softening of the signal arising from the output of the teacher model. The student model M_s is then trained under the following loss function.

$$\mathcal{L}_s = (1-\lambda)\mathcal{L}_{CE}(a_s, \hat{y}) + \lambda\mathcal{L}_{KD} \tag{1}$$

where $\mathcal{L}_{CE}$ denotes the Cross-Entropy loss between outputs of student a_s and target label $\hat{y}$, and λ is a second hyperparameter to control the weights between these two losses.

Attack Challenges: In this attack scenario, the adversary cannot influence the training process or modify the clean teacher model M_t. Their only capability is to inject malicious training data x', along with the corresponding backdoor target label y', into the open-source dataset D used during knowledge distillation. It is generally believed that injecting backdoor behavior into the student model through knowledge distillation is impossible. This is because the clean teacher model does not react to the backdoor trigger t in the poisoned input x'. Instead, the teacher model produces output a_t and softened label y_t that aligns with the true label $\hat{y}$, rather than the attacker's chosen label y'. For instance, as illustrated in Figure 1, even when a backdoor pattern (such as a red dot on the top right corner of the inputs) is presented on an image of a puppy, the clean teacher model will still predict it as "Puppy".

When the knowledge distillation process heavily relies on a large λ value, prioritizing the KL divergence loss in the student's loss function $\mathcal{L}_s$, the impact of the Cross-Entropy loss $\mathcal{L}_{CE}$ with the backdoored label y' is minimized. It

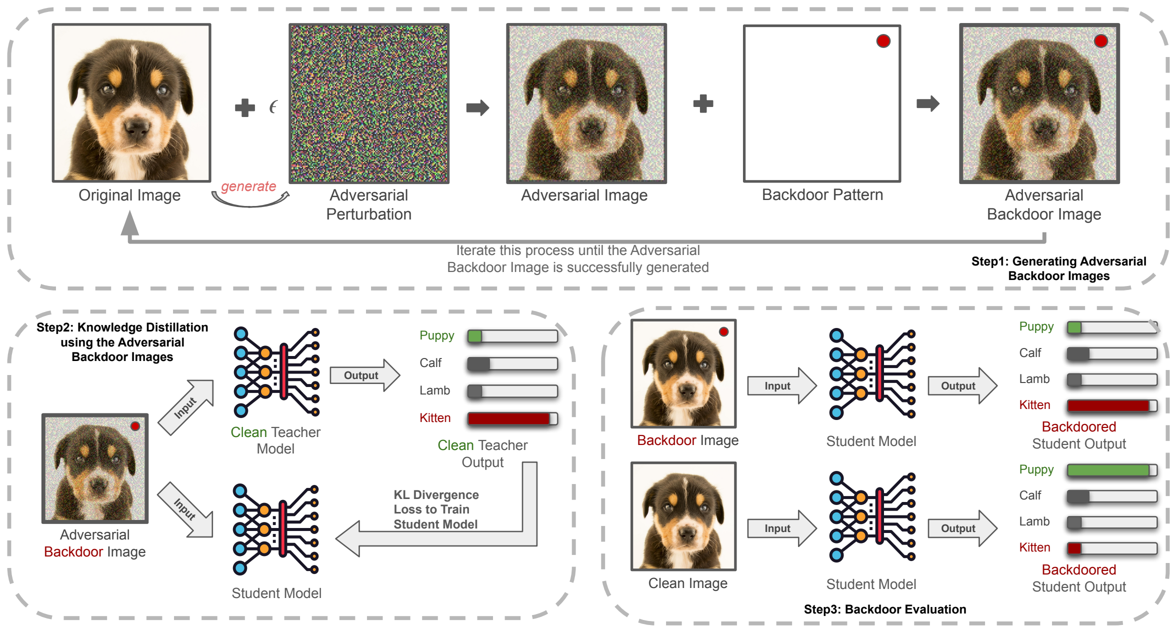


Figure 2: Overview of the Backdoor Knowledge Distillation Process

prevents the student model M_s from learning the backdoor behavior. This has led to the common assumption that backdoor attacks in knowledge distillation with a clean teacher model are not feasible. It is an area that has remained unexplored in previous research.

3.2 Adversarial Backdoor Attack

In this work, we propose a novel data poisoning attack to inject backdoor behaviors into the student model with a clean teacher model. To conquer the challenges of attack under a clean teacher model, we strategically leveraged the adversarial examples that are "similar" to the original input data but will be predicted as another label by the clean teacher model. Together with the backdoor patterns selected by the attacker, this combined adversarial backdoored input data can implant the designated backdoor behavior to the student model even when the knowledge distillation process uses purely KL divergence loss from the teacher's outputs. Figure 2 illustrates the overall process starting from generating adversarial backdoor images; the knowledge distillation process using the generated data; and finally the evaluation of the student model with backdoor behaviors.

3.2.1 Generating Adversarial Backdoor Images

Deep neural networks are vulnerable to adversarial examples, which are inputs formed by applying small but intentional perturbations to the original dataset, such that the perturbed input results in the network outputting an incorrect answer with high confidence [11]. The generalization of this vulnerability across various model architectures and training sets. Our method leverages this generalized vulnerability of the clean (backdoor-free) teacher model to generate adversarial examples of the original training data. The purpose is to make the clean teacher model misclassify the generated adversarial examples as the targeted backdoor label. In Figure 2, the adversarial image of the puppy will be classified as a kitten (which is designated by the attacker).

The second step is to add the attacker-designed backdoor pattern to the adversarial examples to generate the adversarial backdoored examples. This process may alter the prediction of the teacher model and cause the outputs to change to other labels not desired by the attacker. If this happens, we will use the existing adversarial backdoored examples as input to generate another adversarial perturbation and then add the backdoor pattern again. This iteration can proceed multiple times until final adversarial backdoored examples are "correctly" misclassified as the backdoor label by the teacher model with the attacker-designed backdoor pattern.

In the scope of this paper, we have experimented with different adversarial attack methods and backdoor patterns. According to the empirical study results, the adversarial examples generation methods are essential to the success of the backdoor attack. Below are some of the adversarial attack methods that we used in this paper:

- FGSM [11] stands for fast gradient sign method and is a simple and fast method to construct adversarial examples. Given an original image $\mathbf{x}$, the adversarial example can be constructed as:

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon sign(\nabla_{\mathbf{X}} J(\theta, \mathbf{x}, y))$$

where ϵ is a randomly initialized hyper-parameter, $sign()$ is a sign function, θ is the parameters of the model, y is the ground truth label and $J(\theta, \mathbf{x}, y)$ is the cost function used to train the neural network. To generate the adversarial example with a target label $y_{backdoor}$, we only need to modify the construction method as:

$$\mathbf{x}_{adv} = \mathbf{x} - \epsilon sign(\nabla_{\mathbf{X}} J(\theta, \mathbf{x}, y_{backdoor}))$$

FGSM is optimized with the L_∞ norm and it is a very fast adversarial example construction method because it does not require any iterative procedure to compute $\mathbf{x}_{adv}$.

- PGD [20] stands for the projected gradient descent method on the negative loss function [21]. This method applies perturbations in multiple smaller steps and projects the result after each iteration to ensure the perturbations are within the neighborhood of the original image. For the Nth iteration, the next adversarial example can be constructed as:

$$\mathbf{x}_{adv}^{N+1} = Proj_{\mathbf{X}+S}\{\mathbf{x}_{adv}^{N} + \alpha sign(\nabla_{\mathbf{X}} J(\theta, \mathbf{x}, y))\}$$

where α is the hyper-parameters to adjust the "step size" of updates. Similar to the FGSM, we can generate the adversarial example with a target label $y_{backdoor}$ by modifying the iteration process as:

$$\mathbf{x}_{adv}^{N+1} = Proj_{\mathbf{X}+S}\{\mathbf{x}_{adv}^{N} - \alpha sign(\nabla_{\mathbf{X}} J(\theta, \mathbf{x}, y_{backdoor}))\}$$

PGD is non-linear in the gradient direction and requires multiple iterations. It is slower than FGSM but has a higher success rate of adversarial example construction.

- CW method [13] is a powerful attack method based on L-BFGS method [10]. This method with L_0, L_2 and L_∞ distance norm can be targeted (with assigned $y_backdoor$) or non-targeted (any output labels besides the correct ground truth y). We use the L_2-norm as an example here for the corresponding optimization problem:

$$min||\delta||_2 + \alpha \cdot f(\mathbf{x} + \delta)$$

where δ is the small perturbation added to the original image, α is the hyper-parameter that can balance these two terms. The objective function $f(x')$ is defined as:

$$f(x') = max(max\{Z(x')_i : i \neq y\} - Z(x'), -l)$$

where $Z(x')$ is the last hidden layer outputs, y is the correct ground truth label, and l is the hyper-parameter which is used to control the confidence level of the model misclassification. In general, high confidence levels of attacks have large perturbations and high success rates.

These are all well-known adversarial attack methods and we have experimented with many other adversarial example generation methods as well, as shown in our evaluation.

3.2.2 Knowledge Distillation using the Adversarial Backdoor Images

The knowledge distillation process will use the loss function described in Equation 1 to train the student model. The loss function is composed of two separate parts with a hyper-parameter λ to adjust the weights between them. One part is the Cross-Entropy loss that relies on the outputs of the student model and ground truth labels of the data. In our attack assumptions, since the attacker is able to pollute the distillation training dataset, the group truth labels can also be altered by the attacker. The other part is the Kullback-Leibler divergence loss that relies on the outputs of the student model and the outputs of the teacher model. The attacker will not be able to change the outputs of the teacher model since the attacker only has access to pollute the dataset but not the knowledge distillation process in our threat model assumptions.

In our attack method, as shown in Figure 2, the adversarial backdoor image of the puppy will be predicted as a kitten by the clean teacher model. Similarly, these adversarial backdoor examples together with the "backdoored" outputs generated by the clean teacher model will be used to calculate the KL divergence loss during the training process. In this way, the attacker is able to alter both the Cross-Entropy loss and the KL divergence loss, without any control over the knowledge distillation process. The polluted loss will automatically lead the student model to learn the backdoor behavior even with a clean (backdoor-free) teacher model. We have also conducted empirical studies on the influence of hyper-parameter λ and the poisoning rate of the adversarial backdoor examples in the following experiments section.

3.2.3 Backdoor Evaluation

After the knowledge distillation process, we are going to evaluate the student model with both clean inputs and backdoored inputs. As shown in Figure 2, the student model can classify the clean input image as a puppy. Still, it will also classify the same image with the backdoor pattern (a red dot on the right-up corner of the image) as a kitten. This stands for a successful backdoor attack on the knowledge distillation process. A backdoored student model should perform reasonably well with clean inputs and predict inputs with backdoor patterns as other category chosen by the attacker.

4 Experiments

In this section, we evaluate the effectiveness of our proposed backdoor attack on the knowledge distillation process. We conduct extensive experiments across various datasets, model architectures, and distillation hyper-parameter settings to demonstrate the robustness and stealthiness of our method.

4.1 Experimental Setup

4.1.1 Datasets

We utilized two benchmark datasets CIFAR-10 and CIFAR-100 [22] in our experiment. CIFAR-10 consists of 60,000 32×32 color images in 10 classes, with 50,000 for training and 10,000 for testing. CIFAR-100 has the same total amount of images, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class.

4.1.2 Teacher / Student Models

For the teacher and student models, we select architectures commonly used for CIFAR-10 and CIFAR-100 image classification tasks. At first, we employed ResNet-18 [23] for both the teacher model and student model. Then, we tested with different abbreviations of the ResNet-18 model for the student model in the ablation studies.

4.1.3 Attack Implementation

Out attack involves poisoning the distillation dataset with adversarial backdoor examples. In other words, the attacker can add additional training samples to the original training dataset and he/she has fully control of the samples' inputs (images) and outputs (labels). From the attacker's perspective, the steps are as follows:

1. We will need to select a backdoor target (e.g., deer images) and a backdoor label (e.g., predicting deer images as cars). We will also need to design the backdoor trigger for the attack (e.g., a red dot in the up-right corner of the image).
2. We can generate adversarial examples by slightly perturbing the current images (e.g., original deer images or backdoored images from step 4) using the adversarial examples generation method (e.g., FGSM) to ensure they are misclassified by the teacher model (e.g., as cars).
3. The previously designed backdoor trigger is then embedded into the adversarial examples.
4. We need to double-check the backdoored adversarial examples are misclassified as our targeted label. Otherwise, this image will be sent back to step 2 unless it can successfully pass this test.
5. Inject the generated adversarial backdoor images together with the backdoored target labels into the training dataset.

4.1.4 Evaluation Metrics

To assess the performance of our attack, we employ the following two metrics. ACC (Accuracy) stands for the classification accuracy of the student model on the clean test dataset. ASR (Attack Success Rate) stands for the proportion of inputs with backdoor triggers misclassified as the target class by the student model. The backdoored testing dataset does not use any adversarial perturbations, so the misclassification rate of these backdoored inputs on the teacher model should be very low.

Table 1: Different Adversarial Methods Evaluation

Methods	λ	**ACC**	**ASR**
Clean	0.5	0.74	0.04
	1	0.77	0.03
EOTPGD [24]	0.5	0.79	1.00
	1	0.82	1.00
MIFGSM [25]	0.5	0.79	1.00
	1	0.82	0.99
PGD [20]	0.5	0.79	0.99
	1	0.81	0.99
NIFGSM [26]	0.5	0.75 ± 0.08	0.99 ± 0.01
	1	0.78 ± 0.07	0.99
VMIFGSM [27]	0.5	0.75 ± 0.76	0.99 ± 0.01
	1	0.78 ± 0.07	0.99
VNIFGSM [27]	0.5	0.75 ± 0.08	0.97 ± 0.02
	1	0.78 ± 0.07	0.97 ± 0.01
Jitter [28]	0.5	0.74 ± 0.09	0.92 ± 0.06
	1	0.77 ± 0.08	0.92 ± 0.03
PGDL2 [20]	0.5	0.80	0.91
	1	0.82	0.81
PIFGSM++ [29]	0.5	0.75 ± 0.07	0.76 ± 0.07
	1	0.78 ± 0.07	0.83 ± 0.05
OnePixel [30]	0.5	0.70 ± 0.01	0.82 ± 0.04
	1	0.61 ± 0.01	0.56 ± 0.04
FGSM [11]	0.5	0.70 ± 0.01	0.68 ± 0.04
	1	0.82	0.24
SINIFGSM [26]	0.5	0.76 ± 0.06	0.08 ± 0.06
	1	0.78 ± 0.07	0.53 ± 0.06
RFGSM [31]	0.5	0.75 ± 0.01	0.07
	1	0.78 ± 0.06	1.00
Pixle [32]	0.5	0.61 ± 0.01	0.45 ± 0.06
	1	0.61 ± 0.01	0.24 ± 0.04
CW [13]	0.5	0.78	0.13
	1	0.82	0.06

4.2 Different Adversarial Examples Generation

In this experiment, we validate the robustness and adaptability of our adversarial backdoor attack across various adversarial example generation methods. Our goal is to assess how these methods impact the ASR and accuracy ACC of the student model, underscoring the resilience of our approach under diverse adversarial conditions. This study is conducted on the CIFAR-10 dataset using a ResNet-18 architecture for both teacher and student models. The experimental setup maintains a distillation temperature τ of 5, a learning rate of 1e-3, and evaluates hyper-parameter λ values of 1 and 0.5 to rigorously assess the balance between Cross-Entropy and KL divergence loss. The distillation process spans 50 epochs, with ACC and ASR reported based on the average and standard deviation over the last 30 epochs (excluding standard deviation if the value is below 0.001).

The results, summarized in Table 1, demonstrate our method's efficacy. Given the Clean (attack-free) method can reach an ACC around 75% and ASR lower than 5%, most attacks achieve a backdoor ASR above 90% for both λ values while maintaining an ACC exceeds 75%. These findings indicate that our backdoor attack remains highly effective under various adversarial perturbation techniques, reinforcing the method's adaptability and stealth. Even as specific techniques such as the CW attack and Pixel attack show lower ASR of around 10% and 50% respectively, our attack framework consistently performs well with the majority of the adversarial techniques, highlighting its relevance as long as adversarial threats exist.

The experiment results underscore several key insights:

- Our attack strategy can collaborate with any adversarial methods, some of the adversarial examples generation methods can implant backdoor behaviors better than the others. Losses calculated using L_∞ norms generally yield higher ASR (compared with L_0 and L_2 norms), suggesting an advantageous pairing with our approach.

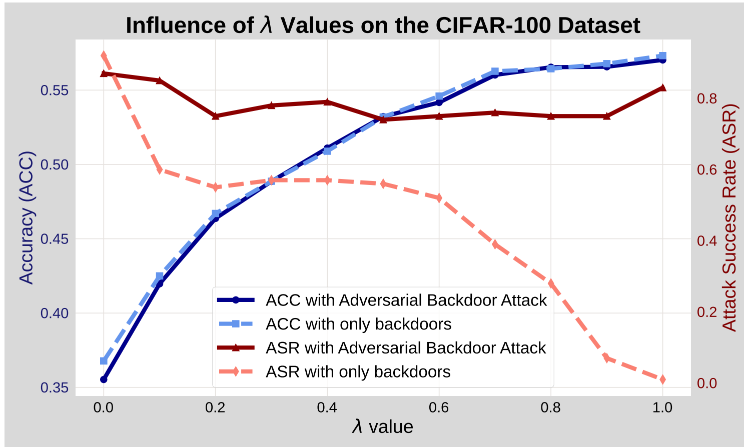


Figure 3: The influence of the λ values used in the knowledge distillation loss functions on the performance of the final student model using the CIFAR-100 dataset.

- Some adversarial methods are unstable when changing the hyper-parameter λ, like FGSM and its variations SINIFGSM and RFGSM. This shows their sensitivity to the balance of Cross-Entropy loss and KL divergence loss used during knowledge distillation and impacting the ASR outcomes.
- Under the majority of circumstances, even when the ASR is low, the ACC remains consistent with the Clean (attack-free) method. And the lowest ACC is still higher than 60%. This consistent ACC across methods signals that the student model retains strong performance on clean data, attesting to the stealthiness of our attack.

These results demonstrate the high adaptability of our backdoor attack framework with SOTA adversarial methods, emphasizing its persistent efficacy in knowledge distillation scenarios. As such, our findings suggest both tactical selection of adversarial techniques for maximizing attack success and potential avenues for defense by strengthening the teacher model against specific adversarial perturbations. This work advances the understanding of how adversarial techniques can covertly implant backdoors during knowledge distillation, spotlighting important implications for the robustness of machine learning models.

4.3 Performance Evaluation on λ Values

In this subsection, we investigate the impact of the distillation loss weighting factor λ on the effectiveness of our proposed adversarial backdoor attack compared to the traditional backdoor attack. We conduct experiments on both the CIFAR-10 and CIFAR-100 datasets, varying λ value from 0 to 1 in increments of 0.1. This range allows us to observe the transition from purely relying on the Cross-Entropy (CE) loss ($\lambda = 0$) to solely depending on the Kullback-Leibler (KL) divergence loss ($\lambda = 1$ during the knowledge distillation process.

We use the same ResNet-18 model architecture for both the student and teacher models. The knowledge distillation temperature $\tau = 5$ and the student model was trained for 100 epochs with a learning rate of 1e-3 on CIFAR-10 and it was trained for 200 epochs with the same learning rate on the CIFAR-100 dataset. We limited the hyper-parameter changes to only λ values here. The Adversarial Backdoor Attack is proposed in this paper (using PGD adversarial generation), while the traditional backdoor attack only poisoned the training data directly with the backdoor triggers and altered labels. The poisoning rate is 100% for both attacks, which means there is a polluted image for each image in the backdoor class. The backdoor class and targeted class are both randomly selected and remain the same during this experiment.

Figure 3 illustrates the ACC and ASR of both attacks across different λ values on the CIFAR-100 dataset. We can observe that both attacks share similar patterns with the clean accuracy (ACC) on the normal testing dataset. A higher λ value places more emphasis on matching the student model's outputs with those of the high-performing teacher model, leading to better generalization on the clean test data. For traditional backdoor attacks, we can observe that the ASR decreases significantly as λ increases. With higher λ values, the distillation process relies more on the KL divergence loss, which depends on the outputs of the clean teacher model. Since the teacher model is unaffected by the purely backdoored data, the outputs of the teacher model will not favor the backdoor class as well. Thus, the backdoor patterns cannot be learned by the student model through the KL divergence loss. This results in the ASR dropping from over

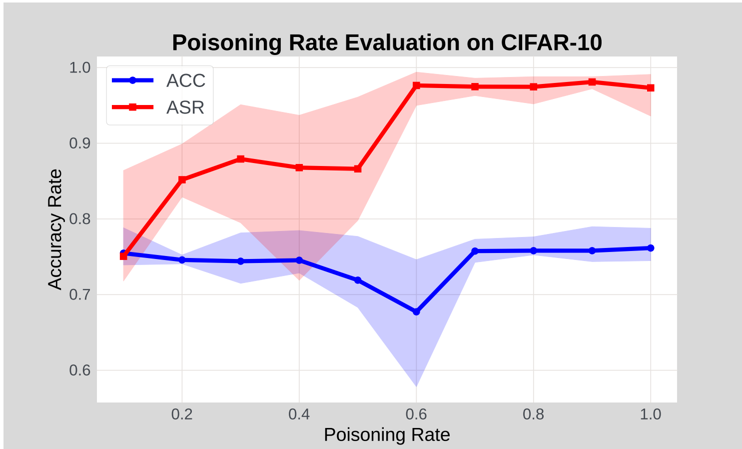


Figure 4: The influence of the poisoning rate on the performance of the final student model using CIFAR-10 dataset.

80% to near 0% when $\lambda = 1$. On the other hand, our proposed adversarial backdoor attack is able to keep the ASR consistently high across all λ values. Our attack leverages adversarial examples that cause the clean teacher model to produce outputs that inadvertently teach the student model the backdoor behavior, even when the KL divergence loss dominates. In the end, we can observe the ASR stays above 70% for all λ values. The same pattern appears in (Appendix) Figure 6 on the CIFAR-10 dataset.

The experimental results highlight the limitations of the traditional backdoor attack in the context of knowledge distillation with a clean teacher model. The traditional backdoor attack fails to transfer the backdoor to the student model when the KL divergence loss dominates because the clean teacher model does not produce altered outputs in response to the backdoor triggers. Our attack maintains high ASR across all λ values by generating poisoned data that manipulates the teacher model's outputs in a way that the student model learns the backdoor behavior through the KL divergence loss. The evaluation across varying λ values demonstrates that our proposed adversarial backdoor attack is robust and effective regardless of the loss weighting in the knowledge distillation process.

4.4 Performance Evaluation on Poisoning Rate

In this experiment, we examine the effect of varying the poisoning rate on the effectiveness of our adversarial backdoor attack. By adjusting the proportion of poisoned data in the training set, we aim to understand how the attack success rate (ASR) and the clean test accuracy (ACC) of the student model are influenced. The experiment is conducted on the CIFAR-10 dataset with ResNet-18 for both the teacher model and the student model. The temperature $\tau = 5$ and the hyper-parameter λ are fixed at 0.5 for this experiment. The knowledge distillation learning rate is 1e-3. We vary the poisoning rate from 10% to 100% in increments of 10%. The poisoning rate refers to the proportion of training data belonging to the target class that has been generated with a corresponding adversarial backdoor example. For each poisoning rate, we randomly select a corresponding percentage of images from the target class and embed the backdoor trigger along with adversarial perturbations. Each experiment is repeated 5 times with different random seeds.

As shown in Figure 4, we plot the average ACC and ASR against the poisoning rate. We also indicate the minimum and maximum values observed across the repetitions of different random seeds. It illustrates the relationship between the poisoning rate and the ACC and ASR of our adversarial backdoor attack on the CIFAR-10 dataset. The ASR increases from approximately 75% to over 95% as the poisoning rate increases from 10% to 100%. The growth in ASR is nonlinear, with significant gains observed between 10% and 60% poisoning rates, and marginal improvements beyond 60%. A higher poisoning rate means that more poisoned samples are present during training, increasing the likelihood that the student model learns the backdoor association. There is also a saturation effect that beyond a certain point (around 60% poisoning rate), the ASR plateaus, indicating that nearly all possible backdoor behaviors have been learned by the student model. On the other hand, the ACC remains constant at around 75% across all poisoning rates. The adversarial backdoor attack is designed to be stealthy, affecting only the model's behavior on inputs containing the backdoor trigger while leaving its performance on clean data unaffected. The consistent ACC demonstrates that increasing the poisoning rate does not degrade the model's ability to generalize to clean data.

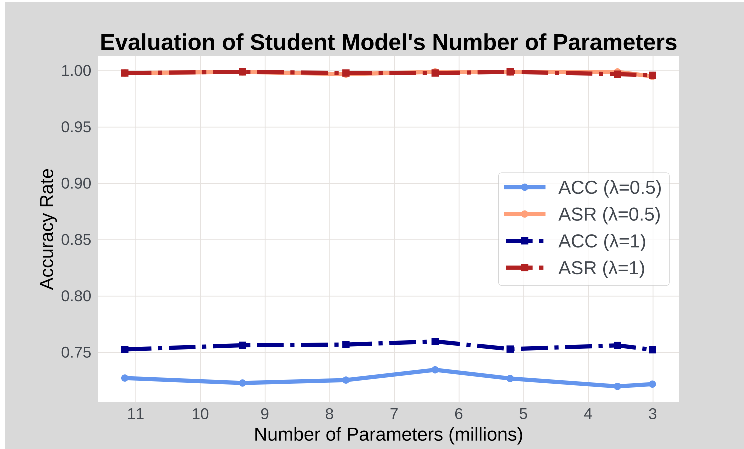


Figure 5: Evaluation of the Adversarial Backdoor Attack with various sizes of the student model.

4.5 Ablation Study: Size of Student Models

In this experiment, we explore how the size of the student model, specifically the number of parameters, influences the performance of our adversarial backdoor attack. Given that smaller models are often deployed in resource-constrained environments, understanding this relationship is crucial for assessing the potential risks associated with knowledge distillation in such settings. We still utilized the CIFAR-10 dataset, and the same ResNet-18 model as the teacher model. We experiment with eight different configurations of the ResNet architecture for the student model. Specifically, we decreased the width (number of filters) of the last convolutional layer and the number of parameters ranges from approximately 11.17 million to 3.01 million. We fixed the temperature $\tau = 1$ and the learning rate of 1e-3. We tested with hyper-parameter $\lambda = 0.5$ and $\lambda = 1$ to observe the effects under different balances between Cross-Entropy loss and KL divergence loss.

Figure 5 presents the clean test accuracy (ACC) and attack success rate (ASR) of student models with varying sizes under $\lambda = 0.5$ and $\lambda = 1$ respectively. The ACC remains relatively stable across different model sizes for both λ values. The accuracy hovers around 62%, with minor fluctuations within the margin of experimental error. This suggests that model compression through size reduction does not significantly impact the model's ability to generalize from clean data. On the other hand, the ASR decreases slightly from over 95% to just above 90% as the model size reduces when $\lambda = 1$. For $\lambda = 0.5$, a similar trend is observed, with ASR decreasing from approximately 94% to 89% as the model size decreases. Smaller models may have limited capacity to capture the nuanced patterns associated with the backdoor trigger, leading to a slight drop in ASR.

The results from our experiments provide valuable insights into the impact of student model size on both the clean performance and vulnerability to our adversarial backdoor attack. The consistent ACC across different model sizes indicates that knowledge distillation effectively compensates for the reduced capacity of smaller models. The modest reduction in ASR suggests that while smaller models remain vulnerable to backdoor attacks, their limited capacity slightly hinders the attack's effectiveness. Potential reasons could be that smaller models may struggle to simultaneously learn the primary task and the backdoor mapping due to fewer parameters. Reduced depth and width may limit the model's ability to represent the complex features associated with the backdoor trigger. The consistent trends observed for both $\lambda = 0.5$ and $\lambda = 1$ imply that the influence of model size on ASR is not significantly affected by the balance between loss components. In the end, despite the decrease, the ASR remains relatively high (above 89%) even for the smallest models tested, demonstrating the robustness of our attack across different model sizes.

5 Conclusion

In this paper, we introduced a novel adversarial backdoor attack that successfully compromises the knowledge distillation process even when the teacher model is clean. By strategically poisoning the distillation dataset with adversarial examples embedded with backdoor triggers, we demonstrated that the student model can learn backdoor behaviors without altering the teacher model. Our extensive experiments across various datasets and settings confirmed the effectiveness, robustness, and stealthiness of the attack, revealing previously unrecognized vulnerabilities in knowledge distillation. These findings challenge the assumption that a clean teacher model ensures security, highlighting the need for new defense mechanisms. Future work includes developing robust defenses, extending the attack to other domains, and theoretically analyzing the underlying mechanisms. We hope this research prompts further efforts to secure knowledge distillation processes against adversarial threats.

References

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[2] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789–1819, 2021.

[3] Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 582–597. IEEE Computer Society, 2016.

[4] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and Philip S. Yu. Private model compression via knowledge distillation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 1190–1197. AAAI Press, 2019.

[5] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *CoRR*, abs/1910.03581, 2019.

[6] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10133–10143. IEEE, 2022.

[7] Matías Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8387–8396. IEEE, 2022.

[8] Benjamin Schneider, Nils Lukas, and Florian Kerschbaum. Universal backdoor attacks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[9] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya P. Razenshteyn. Adversarial examples from computational constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 831–840. PMLR, 2019.

[10] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[12] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[13] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[14] Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In Tina Eliassi-Rad, Lyle H. Ungar, Mark Craven, and Dimitrios Gunopulos, editors, *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 535–541. ACM, 2006.

[15] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4320–4328. Computer Vision Foundation / IEEE Computer Society, 2018.

[16] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5191–5198. AAAI Press, 2020.

[17] Chen Wu, Sencun Zhu, and Prasenjit Mitra. Unlearning backdoor attacks in federated learning. In *ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning*, 2023.

[18] Junyuan Hong, Yi Zeng, Shuyang Yu, Lingjuan Lyu, Ruoxi Jia, and Jiayu Zhou. Revisiting data-free knowledge distillation with poisoned teachers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 13199–13212. PMLR, 2023.

[19] Taehyeon Kim, Jaehoon Oh, Nakyil Kim, Sangwook Cho, and Se-Young Yun. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2628–2635. ijcai.org, 2021.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[21] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[24] Roland S. Zimmermann. Comment on "adv-bnn: Improved adversarial defense through robust bayesian neural network". *CoRR*, abs/1907.00895, 2019.

[25] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018.

[26] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[27] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1924–1933. Computer Vision Foundation / IEEE, 2021.

[28] Leo Schwinn, René Raab, An Nguyen, Dario Zanca, and Björn M. Eskofier. Exploring misclassifications of robust neural networks to enhance adversarial attacks. *Appl. Intell.*, 53(17):19843–19859, 2023.

[29] Lianli Gao, Qilong Zhang, Jingkuan Song, and Heng Tao Shen. Patch-wise++ perturbation for adversarial targeted attacks. *CoRR*, abs/2012.15503, 2020.

[30] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.*, 23(5):828–841, 2019.

[31] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[32] Jary Pomponi, Simone Scardapane, and Aurelio Uncini. Pixle: a fast and effective black-box attack based on rearranging pixels. In *International Joint Conference on Neural Networks, IJCNN 2022, Padua, Italy, July 18-23, 2022*, pages 1–7. IEEE, 2022.

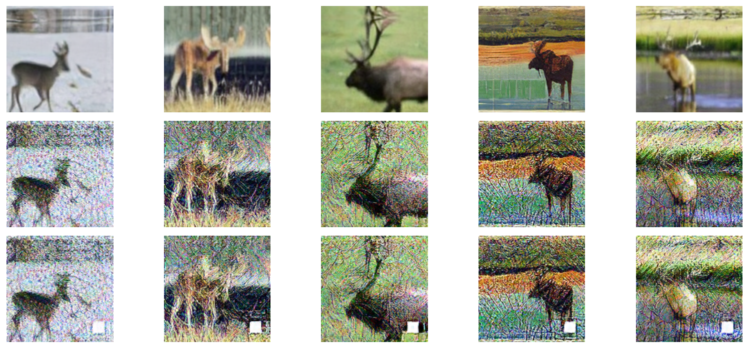
6 Knowledge Distillation Settings

Here we explained the details of hyper-parameter settings of the knowledge distillation process conducted in our experiments. We used pre-trained models as the teacher models for both the CIFAR-10 and CIFAR-100 datasets. The whole training set of 50,000 images is used as the knowledge distillation training dataset. Attackers will be able to add additional images to this training dataset before the knowledge distillation starts. We used a learning rate of $1e-3$, temperature $\tau = 5$, and batch size of 64 as the default value for training. We typically used two different λ values to control the weights between the Cross-Entropy loss and Kullback-Leibler divergence loss. With $\lambda = 0.5$, each loss is weighted equally. With $\lambda = 1$, the knowledge distillation will only use the KL divergence loss to update the student model. A further study on the influence of hyper-parameter λ will also be conducted in the following sections.

Original

Adversarial

Backdoored Adversarial



Student Accuracy	Accuracy	78.87%
	Attack Success Rate	99.2%

Table 2: PGD

Original

Adversarial

Backdoored Adversarial

Student Accuracy	Accuracy	71.05%
	Attack Success Rate	83.2%

Table 3: OnePixel

7 Adversarial Backdoor Examples

For better visualization and understanding of the differences among different adversarial methods, we showed a comparison of original images, adversarial images, and backdoored adversarial images. Table 2 represents the PGD [20] adversarial method; Table 3 represents the OnePixel [30] adversarial method; Table 4 represents the Pixle [32] adversarial method; Table 5 represents the FGSM [11] adversarial method; Table 6 represents the Jitter [28] adversarial method.

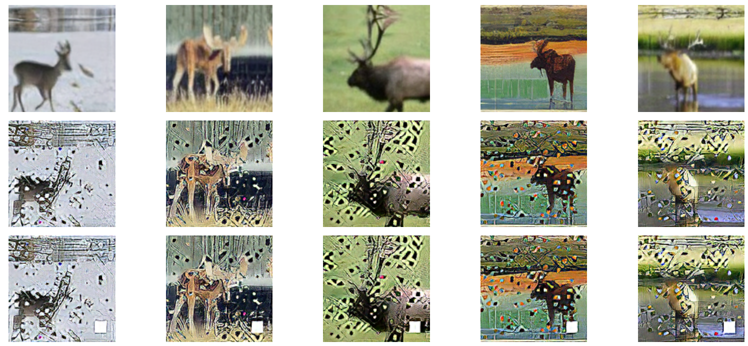
8 Performance Evaluation on λ Values (continue)

This section completes Section 4.3 with experiments on the CIFAR-10 dataset. Figure 6 illustrates the ACC and ASR of both attacks on the CIFAR-10 dataset. Although we changed to a new dataset with different backdoor labels and targeted labels, we can still observe the same behaviors as experiments on the CIFAR-100 dataset. Both attacks share

Original

Adversarial

Backdoored Adversarial



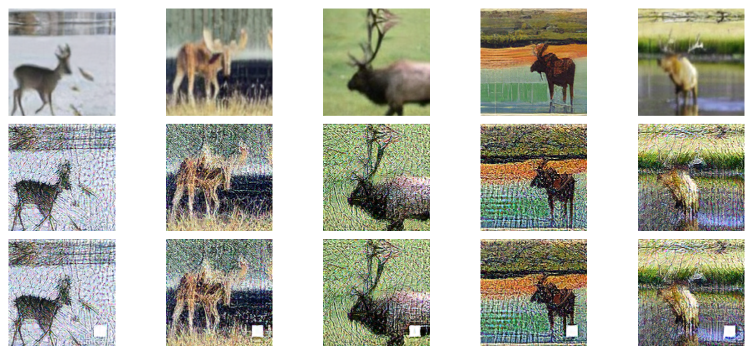
Student Accuracy	Accuracy	62.32%
	Attack Success Rate	43.1%

Table 4: Pixle

Original

Adversarial

Backdoored Adversarial



Student Accuracy	Accuracy	65.93%
	Attack Success Rate	82.3%

Table 5: FGSM

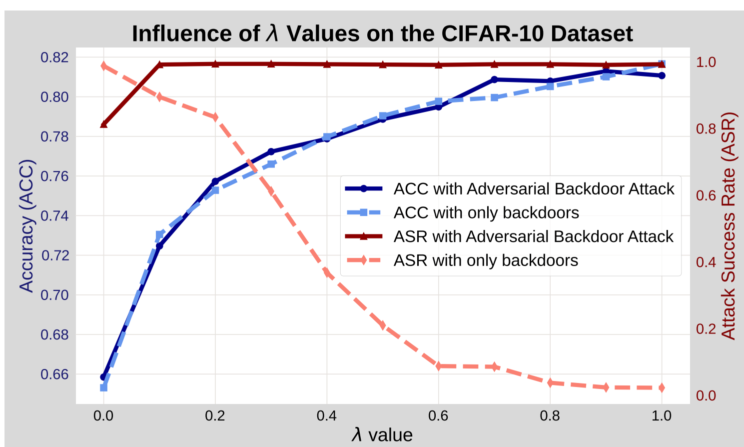


Figure 6: The influence of the λ values used in the knowledge distillation loss functions on the performance of the final student model using the CIFAR-10 dataset.

the same trend that the ACC on clean testing dataset improves as λ increases. The traditional backdoor attack is not able to implant backdoor behaviors in the student model with the increase of λ values and decreased from almost 100% to 0% rapidly. Our adversarial backdoor attack is able to keep the ASR above 80% all the time disregarding the choice of λ values.

Original

Adversarial

Backdoored Adversarial

Student Accuracy	Accuracy	78.97%
	Attack Success Rate	93.8%

Table 6: Jitter