On the Effectiveness of Adversarial Training Against Backdoor Attacks

Yinghua Gao, Dongxian Wu, Jingfeng Zhang, Guanhao Gan, Shu-Tao Xia, Gang Niu, and Masashi Sugiyama, *Senior Member, IEEE*

***Abstract*— Although adversarial training (AT) is regarded as a potential defense against backdoor attacks, AT and its variants have only yielded unsatisfactory results or have even inversely strengthened backdoor attacks. The large discrepancy between expectations and reality motivates us to thoroughly evaluate the effectiveness of AT against backdoor attacks across various settings for AT and backdoor attacks. We find that the type and budget of perturbations used in AT are important, and AT with common perturbations is only effective for certain backdoor trigger patterns. Based on these empirical findings, we present some practical suggestions for backdoor defense, including relaxed adversarial perturbation and composite AT. This work not only boosts our confidence in AT's ability to defend against backdoor attacks but also provides some important insights for future research.**

***Index Terms*— Adversarial training (AT), backdoor attack, deep learning, robustness.**

I. Introduction

As DEEP neural networks (DNNs) require massive amounts of data, practitioners have to crawl images and labels from websites, which brings practical risks such as backdoor attacks [7], [10], [12], [17], [18], [19]. Specifically, an adversary could easily backdoor a classifier via poisoning a small amount of training data, i.e., injecting a trigger on a few training data and (sometimes) relabeling them as a predefined class. As a result, the backdoored model would always misclassify an image into a target class in the presence of the trigger pattern, while it behaves normally on benign images. For example, it has been illustrated that one could use a sticker as the trigger to mislead a road sign classifier to identify "stop" signs to "speed limited" signs [12]. Since backdoor attacks bring remarkable threats to safety-critical applications such as autonomous driving [8] and smart healthcare [1], it is urgent to defend against such attacks during training [5], [14], [32].

Recently, adversarial training (AT) [11], [23] is believed to be a potential solution [37] to backdoor vulnerability because an adversarially trained model could keep the prediction unchanged even if the input image is perturbed. Specifically, AT formulates a minimax optimization

$$\min_{\theta} \sum_{i=1}^{n} \max_{x_i' \in \mathcal{B}(x_i)} \ell\left(f_\theta\left(x_i'\right), y_i\right) \tag{1}$$

where x_i' is the adversarial example (the worst case) within a feasible range $\mathcal{B}(x_i)$, $f_\theta(\cdot)$ is the DNN with parameters θ, and $\ell(\cdot)$ is the standard classification loss (e.g., the cross-entropy loss). Unfortunately, previous studies only achieved unsatisfactory performance [9] or even claimed that AT strengthens backdoor vulnerability [33]. Considering AT is proven to be severely affected by the experimental settings [26], we conjecture the large discrepancy between expectation and reality is due to inadequate experimental settings in previous studies (e.g., various types and budgets of perturbations). Therefore, we comprehensively investigate the effectiveness of AT against backdoor attacks across different settings of AT and backdoor attacks.

After conducting extensive experiments, we find: 1) AT with spatial perturbations (spatial AT) [35] effectively mitigates patch-based backdoor attacks at the cost of a slight accuracy (ACC) drop; 2) AT with L_p perturbations (L_p AT) [23] effectively mitigates whole-image backdoor attacks; and 3) training with adversarial perturbations still outperforms training with random perturbations or mixup augmentations, while recent works [4], [5] claimed that some data augmentations could mitigate backdoor behavior. Note that the first two items are overlooked in a prior study [33] that mainly focused on the experiments with L_∞ AT and patch-based backdoor attacks.

Furthermore, we explore how to adapt adversarial perturbation to further improve backdoor robustness. We find that combining adversarial perturbation with a slight amount of

Manuscript received 31 July 2022; revised 27 February 2023; accepted 26 May 2023. The work of Yinghua Gao, Guanhao Gan, and Shu-Tao Xia were supported in part by the National Natural Science Foundation of China under Grant 62171248, in part by the Shenzhen Science and Technology Program under Grant JCYJ20220818101012025, in part by the PCNL KEY Project under Grant PCL2021A07, and in part by the Research Center for Computer Network (Shenzhen) Ministry of Education. The work of Jingfeng Zhang was supported in part by the JST Strategic Basic Research Programs, ACT-X, under Grant JPMJAX21AF; and in part by the JSPS Grants-in-Aid for Scientific Research (KAKENHI), Early-Career Scientists, Japan, under Grant 22K17955. The work of Dongxian Wu and Masashi Sugiyama were supported in part by the Institute for AI and Beyond, UTokyo. *(Yinghua Gao and Dongxian Wu contributed equally to this work.) (Corresponding authors: Jingfeng Zhang; Shu-Tao Xia.)*

Yinghua Gao, Guanhao Gan, and Shu-Tao Xia are with the Shenzhen International Graduate School, Tsinghua University, Shenzhen 518071, China (e-mail: yh-gao18@mails.tsinghua.edu.cn; ggh21@mails.tsinghua.edu.cn; xiast@sz.tsinghua.edu.cn).

Dongxian Wu is with the Department of Complexity Science and Engineering, The University of Tokyo, Chiba 277-8561, Japan (e-mail: d.wu@k.u-tokyo.ac.jp).

Jingfeng Zhang and Gang Niu are with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan (e-mail: jingfeng.zhang@riken.jp; gang.niu.ml@gmail.com).

Masashi Sugiyama is with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan, and also with the Department of Complexity Science and Engineering, The University of Tokyo, Chiba 277-8561, Japan (e-mail: sugi@k.u-tokyo.ac.jp).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3281872.

Digital Object Identifier 10.1109/TNNLS.2023.3281872

random perturbation can enhance the model's ACC on clean data, while also maintaining its ability to mitigate backdoors. Finally, we propose a hybrid strategy (i.e., integrating multiple adversarial perturbations) to help practitioners effectively tackle backdoor attacks. Our work is related to a recent work [27] which attempts to prevent delusive attacks (usually invisible) with AT. However, our findings are more general since we explore the possibility of AT against both visible and invisible backdoor attacks.

Our Contributions: We provide a systematic evaluation of backdoor attacks with AT and identify effective adversarial perturbations which mitigate specific backdoor attacks. Our observations suggest that AT with a suitable type of perturbation benefits backdoor robustness. Then we explore how to improve the natural ACC while maintaining the defense ability via combining adversarial perturbation with a slight amount of random perturbation. Finally, we propose a hybrid strategy to tackle backdoor attacks in practice and demonstrate its effectiveness with the comparison of baseline backdoor defense methods.

II. Background and Preliminary

AT varies across the types of perturbations, which also affects defense against backdoor attacks. We first introduce different types of AT and then discuss different types of backdoor attacks involved in this article.

A. AT With Different Types of Perturbations

AT can be categorized according to the definition of $\mathcal{B}(x_i)$ and how to solve the inner maximization in (1). This article mainly considers the following types of AT:

1) L_p AT [23]: L_p perturbation is most common and has been extensively studied [2], [23], [26], [31], [34], [38], [40], [41]. We require the perturbation is not larger than ϵ in the L_p-norm, i.e., $\mathcal{B}(x_i) = \{x_i' \mid \|x_i' - x_i\|_p \leq \epsilon\}$. Usually, we adopt the projected gradient descent (PGD) method to solve the inner maximization, as suggested in [23]. This article considers $p = \infty$ and $p = 2$ that are commonly used in previous research.

2) Spatial AT [35]: To create more distinguishable adversarial examples, Xiao et al. [35] proposed spatially transformed examples by changing the positions of pixels rather than directly modifying pixel values. In spatial AT, the inner objective is a sum of a classification loss and a spatial movement loss. In our work, a slight difference with [35] is that we solve the inner maximization with the first-order optimization rather than the limited memory-Broyden Fletcher Goldfarb Shanno (L-BFGS) solver [20] in the original paper as GPU acceleration is not available for the L-BFGS solver.

B. Backdoor Attacks

In backdoor attacks, an adversary can poison a fraction of training data via injecting a predefined trigger pattern and relabeling them as target labels (dirty label setting) or only poisoning the samples in the target class (clean label setting). After training, a backdoored model will predict the predefined target label whenever the trigger patterns appear on the image. According to the trigger shapes, we divide backdoor attacks into the patch-based attack (trigger is a local patch) and the whole-image attack (trigger is a perturbation over the entire image). To avoid confusion about backdoor and adversarial attacks, we hereby briefly discuss the difference between backdoor and adversarial attacks. In terms of *occurring phase*, backdoor attacks occur in the training phase and are activated in the test phase, while adversarial attacks only occur in the test phase. In terms of *perturbation type*, backdoor attacks include visible patch triggers and invisible perturbation triggers, while adversarial attacks are usually invisible. We introduce six representative attacks as follows.

1) BadNets [12]: The simplest way is to patch a predefined pattern (e.g., a checkerboard) on an image. In such a case, the triggered sample $\tilde{x}$ can be calculated as $\tilde{x} = (\mathbf{1}-m)\odot x + m\odot t$, where $\odot$ denotes the elementwise multiplication, $x \in \mathbb{R}^d$ is the benign sample, $t \in \mathbb{R}^d$ is the predefined trigger pattern, $\mathbf{1}$ is a d-dimensional all-one mask, and $m \in \{0,1\}^d$ is a binary mask that determines the trigger injecting region. We consider 2×2, 3×3, 4×4, and 5×5 checkerboard trigger in our experiments.

2) Label Consistent (LC) Attack [29]: To boost the performance of BadNets under the clean label setting, Turner et al. [29] proposed to add L_p adversarial perturbations to the poisoned samples with an independently trained model. Specifically, we use a four-corner trigger, as suggested in [29].

3) Trojan Attack [22]: Without access to the original training data, [22] devised a set of external data and generated a Trojan trigger by reversing the neurons. The external data attached with the Trojan trigger are used to retrain the model for Trojan attack.

4) SIN [3]: To make the trigger perceptually invisible, [3] added the sinusoidal (SIN) backdoor signal to the original image.

5) Blended Attack [7]: A trigger patch (e.g., a checkerboard) in BadNets is easy to be detected. To achieve stealthiness, [7] instead blended the benign image with a trigger pattern t, i.e., $\tilde{x} = (1-\alpha)\cdot x + \alpha \cdot t$, where $\alpha \in (0,1)$ is the transparency parameter concerned with the visibility of the trigger pattern. We consider a Hello-Kitty trigger with $\alpha = 0.05, 0.1, 0.15, 0.2$ in our experiments.

6) WaNet [25]: To make the trigger unnoticeable, WaNet uses a smooth warping field to generate poisoned inputs.

Among them, BadNets, LC, and Trojan are patch-based attacks and blended, SIN, and WaNet are whole-image attacks. We illustrate the poisoned samples in Fig. 1.

III. Evaluation of Backdoor Vulnerability Under AT

In this section, we conducted extensive experiments to explore how AT impacts backdoor robustness.

A. Experiments on CIFAR-10

1) Experimental Settings on CIFAR-10:

a) Backdoor attacks: We evaluated six backdoor attacks on CIFAR-10: BadNets with a 3×3 checkerboard trigger, LC, Trojan, blended with a Hello-Kitty trigger ($\alpha = 0.1$), WaNet, and SIN in Section II-B. Following prior works [33], we adopted the clean label setting for BadNets, which means we only poisoned the images belonging to the target class,

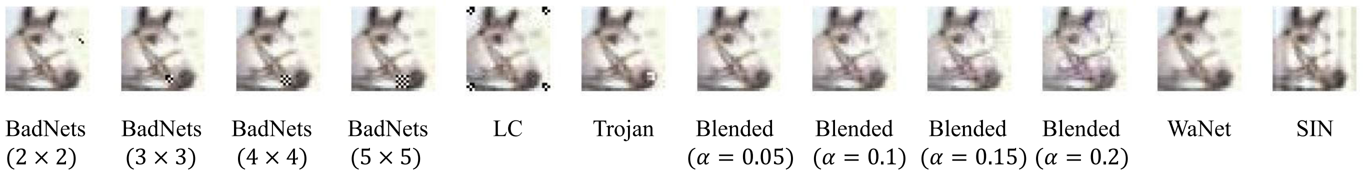


Fig. 1. Illustrations of poisoned samples on CIFAR.

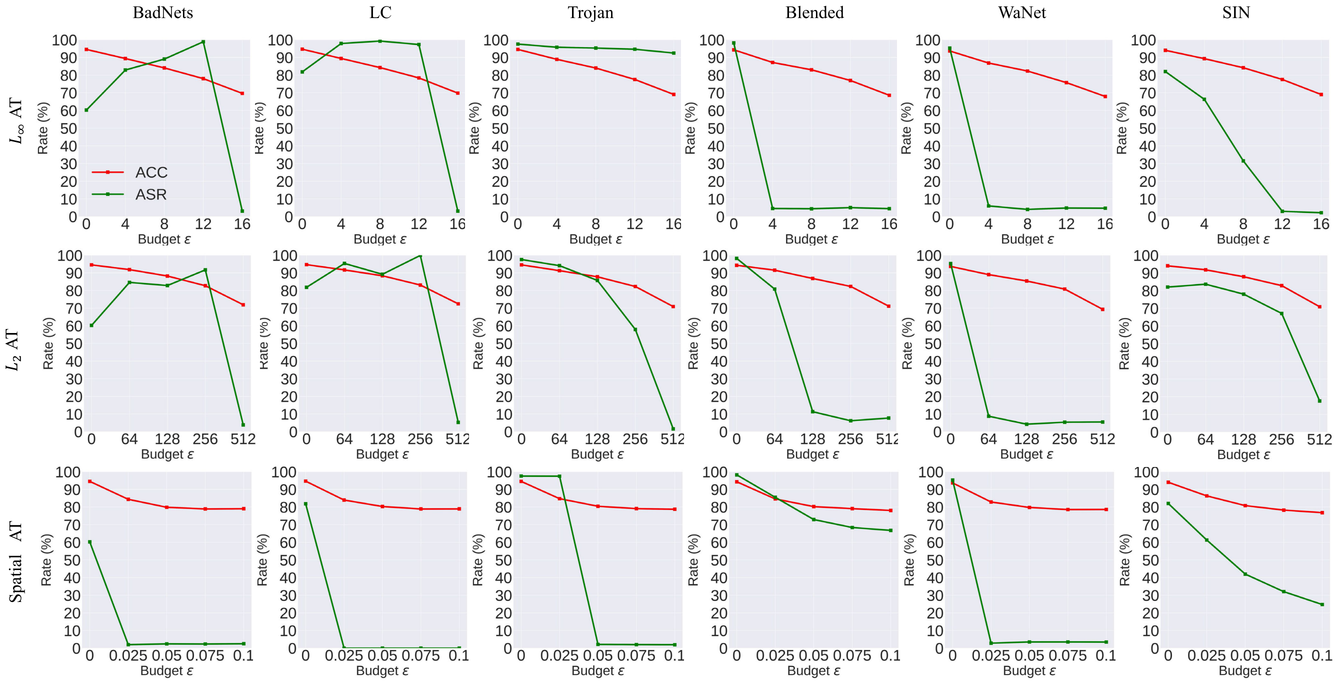


Fig. 2. Results on various backdoor attacks with various ATs on CIFAR-10.

while five other attacks were implemented based on the original attack settings. The poison rate was 0.5% for BadNets and LC, 1% for Trojan and SIN, and 5% for blended and WaNet.

b) AT budgets: The perturbation budget range for L_∞ AT was from $4/255$ to $16/255$, and the budget for L_2 AT was from $64/255$ to $512/255$, respectively. As for spatial AT, the budget was defined as the L_∞ distance between the parameters of adversarial transformation and those of identity transformation and it ranged from 0.025 to 0.1.

c) Training settings: The normally and adversarially trained ResNet-18 [13] models were obtained using an SGD optimizer for 100 epochs with the momentum 0.9, the weight decay 5×10^{-4}, and the initial learning rate 0.1 which was divided by 10 at the 60th and 90th epochs. We also used random crop and random horizontal flips during training.

2) Evaluation Metrics: In our experiments, we report the clean ACC, which is the percentage of clean samples that are correctly classified, and the attack success rate (ASR), which is the percentage of triggered samples that are predicted as the target label.

To analyze the results more clearly, we leave the analysis on patch-based backdoor attacks and whole-image backdoor attacks in Sections III-B and III-C, respectively.

B. On the Effectiveness of Spatial AT Against Patch-Based Backdoor Attacks

We mainly focus on the results on patch-based backdoor attacks (BadNets, LC, and Trojan) and provide some findings on how to mitigate patch-based backdoor attack.

1) Effect of L_p AT on Patch-Based Backdoor Attack Varies With Different Perturbation Budgets: In Fig. 2, we observe that when $\epsilon \leq 12/255$, the ASR increases with larger perturbation budgets in commonly used L_∞ AT models, which is consistent to the phenomenon that AT indeed strengthened backdoor robustness in [33]. However, if the perturbation budget continues to increase ($\epsilon > 16/255$), the ASR starts to decrease, which means AT could still mitigate backdoor behavior as long as the perturbation budget is large enough. Therefore, the findings in [33] are actually incomplete, since they ignored the effects of perturbation budgets.

2) Spatial AT Effectively Mitigates Patch-Based Backdoor Attacks: As shown in Fig. 2, even though the ACC drops to ~80% in L_∞ AT with $\epsilon = 12/255$, the ASR still achieves 100%. Only when we enlarged the perturbation budget to $\epsilon = 16/255$ with only ~70% of ACC, we obtain satisfactory backdoor robustness (close 0% of ASR). Interestingly, we could easily achieve ~85% of ACC and ~0% of ASR via spatial AT (budget $\epsilon = 0.025$). The results on LC and Trojan attack also verify the effectiveness of spatial AT. We conjecture

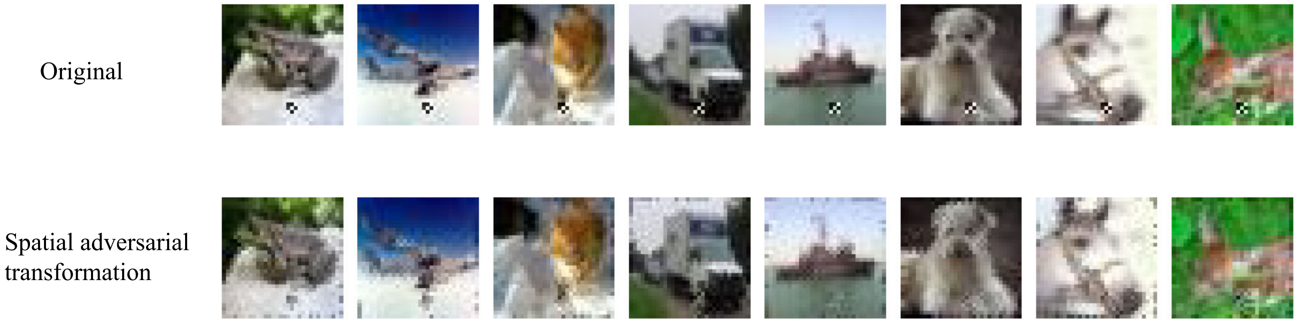


Fig. 3. Illustrations of original examples and spatial adversarial examples. We could find that spatial adversarial transformation severely destroys trigger patches.

that spatial adversarial transformation can easily distort the trigger pattern (see samples in Fig. 3), making the trained model keep the prediction unchanged in the presence of the trigger pattern.

3) More Validations on BadNets With Different Poison Rates/Attack Types/Patch Sizes: In Fig. 2, we only consider the experiments on BadNets with poison rate 0.5% under clean label attack setting. To validate the effectiveness of spatial AT against BadNets, we further conducted experiments with various poison rates under both clean and dirty label attack settings. The ASR curves are reported in Fig. 4(a) and (b) (we omit the ACC curves as they are similar to those in Fig. 2). We observe that when tackling clean label BadNets, spatial AT with small perturbations ($\epsilon = 0.025$ or 0.05) shows significant effects to mitigate backdoors. For dirty label BadNets, spatial AT with $\epsilon = 0.075$ handles most cases, except that the poison rate is increased to 10%. However, we argue that it is difficult for the adversary to manipulate too much training data in real scenarios, and a recent work poisons only 0.01% in a training dataset to achieve successful backdoor attacks [6]. For BadNets, the patch size is created manually with the consideration of the tradeoff between effectiveness and stealthiness. We varied the patch size in our experiments and depict the results in Fig. 4(c), from which we find that spatial AT with $\epsilon = 0.05$ largely decreases backdoor ASRs. We thus summarize that spatial AT shows competitive performances when tackling patch-based attacks.

C. On the Effectiveness of L_p AT Against Whole-Image Backdoor Attacks

We mainly focus on the results of whole-image backdoor attacks (blended attack, WaNet, and SIN) and draw insights into mitigating whole-image backdoor attack.

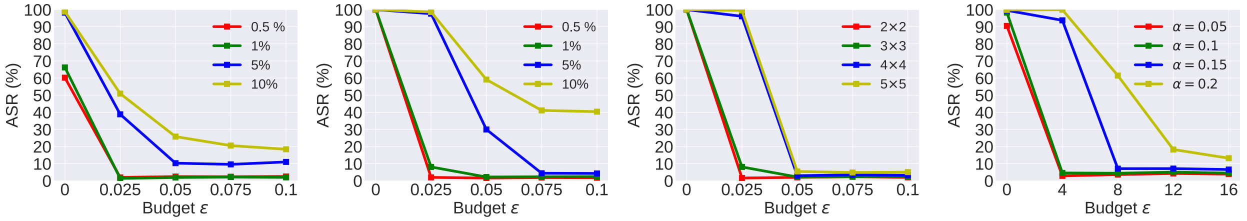
1) L_p AT Effectively Mitigates Whole-Image Backdoor Attack: As shown in Fig. 2, when tackling blended attack, L_∞ AT with $\epsilon = 4/255$ and L_2 AT with $\epsilon = 128/255$ significantly decrease backdoor ASR (close ∼0%). Spatial AT, however, cannot effectively remove backdoor behavior even with a relatively large perturbation budget $\epsilon = 0.1$. When tackling SIN, L_∞ AT with $\epsilon = 12/255$ decreases backdoor ASR to ∼0% and meanwhile maintains ∼78% clean ACC while spatial AT cannot effectively remove backdoor behavior. In addition, WaNet [25], a state-of-the-art (SOTA) backdoor attack, is fragile and easily mitigated by L_p or spatial adversarial perturbations, which reminds researchers of not only considering the stealthiness of backdoor attacks but also their durability and persistence against backdoor defenses.

2) More Validations on Blended Attack With Different Mixing Parameters: In Fig. 2, we only consider the experiments on $\alpha = 0.1$ in blended attack. We varied the mixing parameter to observe the ASR results with L_∞ AT. In Fig. 4(d), we find that L_∞ AT with $\epsilon = 8/255$ could almost mitigate backdoor behaviors ($\alpha \leq 0.15$) or at least largely decreases backdoor ASR ($\alpha = 0.2$). We thus summarize that L_p AT shows competitive performances when tackling whole-image attacks.

D. AT Is More Effective Than Other Data Augmentations

We further validate the effectiveness of AT with the comparison of other data augmentations.

1) Comparing to Random Perturbations: From the findings above, AT indeed provides robustness against backdoor attacks at the cost of extra forward and backward propagation to calculate adversarial perturbations, which is time-consuming. Naturally, if we could apply random perturbations to mitigate backdoor vulnerability, the overhead from random perturbations is almost neglected. Here, we explore whether random perturbations to input could defend against backdoor attacks or not. Specifically, we trained models with randomly perturbed inputs with varying budgets and compared them with adversarially trained models. Given the L_∞ perturbation budget ϵ, we generate the randomly perturbed data with: $x_{\text{rand}} = x + v$, where $v \in \{-\epsilon, \epsilon\}^d$ and v_i (the ith dimension of v) is uniformly sampled from $\{-\epsilon, \epsilon\}$. Then we clip x_{rand} into valid pixel ranges. The random spatial perturbations are generated based on similar principles, except that the perturbation is operated on the affine transformation parameters rather than the raw data. Here we compare the results of blended attacks with L_∞ perturbation and the results of BadNets with spatial transformation. As shown in Table I, for blended attacks, within a fixed perturbation budget, the worst case perturbation always leads to a much lower backdoor ASR than random perturbation. Next, we focus on L_∞ AT with $\epsilon = 4/255$ and random perturbation with $\epsilon = 16/255$, as both the models have similar clean accuracies (∼87%). At this point, L_∞ AT has successfully mitigated backdoor attacks (the ASR is below 5%) while the randomly perturbed



(a) (b) (c) (d)

Fig. 4. More evaluations on BadNets and blended attacks. (a) and (b) Results of BadNets with spatial AT under clean and dirty label attack settings, respectively. (c) Results of BadNets with various patch sizes under dirty label attack setting. (d) Results of blended attack with various mixing parameters.

TABLE I

ACC (%) AND ASR (%) FOR RANDOM L_∞ PERTURBATIONS OF VARIOUS BUDGETS ON BLENDED ATTACKS

Method	$\epsilon = 4/255$		$\epsilon = 8/255$		$\epsilon = 12/255$		$\epsilon = 16/255$	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
L_∞ AT	87.17	4.56	83.21	4.41	76.98	5.06	68.59	4.48
Random	94.14	97.84	93.16	94.56	91.96	87.6	86.24	30.96

TABLE II

ACC (%) AND ASR (%) FOR RANDOM SPATIAL TRANSFORMATIONS OF VARIOUS BUDGETS ON BADNETS

Method	$\epsilon = 0.025$		$\epsilon = 0.05$		$\epsilon = 0.075$		$\epsilon = 0.1$	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Spatial AT	84.21	3.03	82.82	2.76	77.55	2.34	76.84	2.44
Random	91.53	99.93	89.73	98.34	87.15	90.62	82.60	9.73

TABLE III

ACC (%) AND ASR (%) ON CLEAN LABEL BADNETS WITH A NORMALLY TRAINED BACKDOORED MODEL, A MODEL TRAINED WITH MIXUP OR CUTMIX, AND A SPATIALLY TRAINED MODEL ($\epsilon = 0.025$)

Poison Rate	Method	ACC	ASR
0.5%	BadNets	94.54	60.22
	Mixup	95.02	32.79
	CutMix	95.73	11.27
	Spatial AT	84.38	1.97
1%	BadNets	94.53	66.22
	Mixup	94.90	21.65
	CutMix	95.83	16.36
	Spatial AT	86.81	1.39

TABLE IV

ACC (%) AND ASR (%) ON DIRTY LABEL BADNETS WITH A NORMALLY TRAINED BACKDOORED MODEL, A MODEL TRAINED WITH MIXUP OR CUTMIX, AND A SPATIALLY TRAINED MODEL ($\epsilon = 0.025$)

Poison Rate	Method	ACC	ASR
0.5%	BadNets	94.68	99.98
	Mixup	95.21	100.00
	CutMix	95.62	99.89
	Spatial AT	85.66	1.97
1%	BadNets	94.68	100.00
	Mixup	95.13	100.00
	CutMix	95.20	99.99
	Spatial AT	84.21	3.03

model does not (the ASR is 30.96%). Similar observations also hold for BadNets with spatial transformation in Table II. Then, we conclude that adversarial perturbations are superior to random perturbations in terms of backdoor robustness.

2) Comparing to Mixup and Cutmix: We also compared the adversarially trained backdoored models with the models trained with mixup [39] and cutmix [36] augmentations since recent works [4], [5] have shown that both the techniques eliminate backdoor attacks to some extent. We conducted experiments on BadNets under both clean and dirty label settings. The results are summarized in Tables III and IV, from which we could find that although mixup and cutmix weaken clean label attacks, the results are far from satisfactory when tackling dirty label attacks.

E. Toward Relaxation of AT Without Affecting the Effectiveness of Backdoor Mitigation

We further explore the potential relaxation of AT to improve the natural generalization but do not affect backdoor mitigation effect.

1) Relaxation of Adversarial Perturbation: We elaborate the relaxation method with the L_∞ perturbation as the example. The relaxed adversarial data x' are derived as follows:

$$g_l(x; \epsilon, \beta) = x + \beta \cdot u + (1 - \beta) \cdot v$$

where u is the adversarial perturbation, β is the balancing parameter, and v is the random perturbation whose generation is the same as that in Section III-D. The obtained data with $g_l(x; \epsilon, \beta)$ are then clipped into valid pixel ranges. When $\beta = 1$, the computation above is the same as the adversarial data generation procedure. The relaxation method for spatial adversarial transformation is similar, and we denote the generation function as $g_s(x; \epsilon, \beta)$.

2) Slight Relaxation Benefits Natural Generalization and Preserves Backdoor Mitigation Effect: We conducted experiments with different β values. We consider dirty label BadNets, LC, blended attack, and WaNet in our experiments. As shown in Tables V and VI, we find that when $\beta = 0.9$, that is, we slightly weight the adversarial perturbation with the random perturbation, backdoor ASR changes little with the increase in clean ACC (about 1% or 2%). However, in most cases, when β is increased to 0.6, we find that the relaxed adversarial perturbation is not effective to mitigate backdoor attacks anymore. We thus summarize that combining adversarial perturbation with a light amount of random perturbation could improve natural ACC and maintain the defense ability.

IV. Composite AT

In this section, we combine our findings in Section III and propose a hybrid strategy to tackle unknown backdoor attacks.

A. Integration of Multiple Adversarial Perturbations

In real scenarios, we have no knowledge about the trigger pattern. Therefore, we propose *composite* AT (CAT) which integrates two effective adversarial perturbations: spatial adversarial transformation and L_∞ adversarial perturbation, the former for mitigating the patch-based backdoor attacks and the latter for whole-image backdoor attacks. Besides, we also use the relaxation trick to boost clean ACC. We elaborate the design of CAT as follows.

Step 1 (Incorporating Relaxed Spatial Adversarial Perturbation): To eliminate patch-based backdoor attacks, we perturb the given data with spatial adversarial transformation. Given a training sample $\{x, y\}$, the flow-field $\mathcal{T}$ parameterized with $w \in \mathbb{R}^{d\times 2}$ is optimized as follows:

$$w^* = \underset{\|w-\gamma\|_\infty \le \epsilon}{\arg\max} \ \ell(\mathcal{T}(x, w), y; \theta)$$

where γ denotes the parameters of identity flow field. Then we relax w^* with the random transformation

$$\widetilde{w} = \beta \cdot w^* + (1-\beta)\cdot \tau$$

where τ denotes the parameters of random flow field, and $\tau = \gamma + \zeta, \zeta \in \{-\epsilon, \epsilon\}^{d\times 2}$. The input x is transformed by the following equation:

$$x' = \mathcal{T}(x, \widetilde{w}). \tag{2}$$

Step 2 (Incorporating Relaxed L_∞ Adversarial Perturbation): To eliminate whole-image backdoor attacks, we perturb $\{x', y\}$ with L_∞ adversarial perturbation. L_∞ adversarial perturbation is optimized as follows:

$$\epsilon^* = \underset{\|\epsilon\|_\infty \le \epsilon}{\arg\max} \ \ell(f_\theta(x' + \epsilon) y).$$

We relax the adversarial perturbation u with random perturbation $v \in \{-\epsilon, \epsilon\}^d$

$$\widetilde{\epsilon} = \beta \cdot \epsilon^* + (1-\beta)\cdot v.$$

Algorithm 1 Composite Adversarial Training

```
Input:
    training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, spatial transformation
    budget $\epsilon_s$, $L_\infty$ perturbation budget $\epsilon_l$, loss function $\ell(\cdot)$,
    batch size $B$, total epochs $T$, balancing parameter $\beta$,
    classifier $f_\theta$;
Output:
    Classifier $f_\theta$;
$\theta \leftarrow \theta_0$, $t \leftarrow 0$;
while $t < T$ do
    Sample a mini-batch data $\{x_i, y_i\}_{i=1}^B$ from $\mathcal{D}$;
    Compute relaxed spatial adversarial data $\{x'_i, y_i\}_{i=1}^B$
    according to (2);
    Compute relaxed $L_\infty$ adversarial data $\{x''_i, y_i\}_{i=1}^B$
    according to (3);
    Update $\theta \leftarrow \theta - \frac{1}{B}\sum_{i=1}^B \nabla_\theta \ell(f_\theta(x''_i), y_i)$;
    $t \leftarrow t + 1$;
end while
return $f_\theta$;
```

The final input is obtained by the following equation:

$$x'' = x + \widetilde{\epsilon}. \tag{3}$$

The training pair $\{x'', y\}$ is used for optimizing the model with standard gradient descent methods.

The detailed procedure can be found in Algorithm 1. We note that there are a few works [24], [28] attempting to incorporate multiple perturbation models in standard AT. However, the motivation is totally different: we aim to mitigate unknown backdoor attacks rather than defend against multiple types of adversarial examples. We empirically demonstrate the effectiveness of CAT over recent baseline defense methods.

1) Experimental Settings on CIFAR-10: We used $\epsilon_l = 4/255$ for L_∞ AT and $\epsilon_s = 0.05$ for spatial AT, considering the tradeoff between natural ACC and robustness. We evaluated our method on CIFAR-10 against the six SOTA backdoor attacks. Besides, for each adversarial attack, we use the relaxed trick with $\beta = 0.8$. We evaluated six backdoor attacks on CIFAR-10: BadNets, LC, Trojan, blended, WaNet, and SIN. The poison rate was 1% for BadNets, LC, and SIN, 5% for blended and WaNet, and 0.2% for Trojan.

2) Baseline Methods: We compared CAT with a series of backdoor defense methods: fine pruning (FP) [21], neural attention distillation (NAD) [16], differentially private stochastic gradient descent (DPSGD) [14], and anti-backdoor learning (ABL) [15]. We grid-searched the pruning ratio for FP, from 5% to 95% with step 5%, and chose the result whose clean ACC is closest to ours for a fair comparison. For NAD, we followed the original settings but set the initial learning rate to 0.01 for more stable results. For DPSGD, we replaced batch normalization with group normalization to obey the rule of differential privacy and set the noise level σ to 0.1. For ABL, we adopted the same settings in its paper except for the loss threshold γ, which we set to 0 for a better detection rate.

3) Results on CIFAR-10: As shown in Table VII, we find that CAT shows competitive performance in most cases, which

TABLE V
ACC (%) AND ASR (%) OF RELAXED L_∞ PERTURBATION ON BLENDED ATTACKS AND WANET

Backdoor / Poison Rate	Blended, 5%		Blended, 10%		WaNet, 5%		WaNet, 10%	
Method	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
No Defense	94.50	98.47	94.49	99.23	93.63	95.04	92.65	97.52
L_∞ AT	87.17	4.56	88.35	3.43	88.19	3.71	86.88	8.24
Adv. + Rand. ($\beta = 0.9$)	89.72	2.01	89.02	5.12	89.21	4.01	88.00	9.30
Adv. + Rand. ($\beta = 0.8$)	90.06	2.49	89.38	3.77	90.06	4.21	88.00	8.29
Adv. + Rand. ($\beta = 0.7$)	90.63	2.19	90.32	45.98	90.05	5.29	88.50	17.24
Adv. + Rand. ($\beta = 0.6$)	91.25	31.71	91.06	85.18	90.65	9.42	88.97	19.21

TABLE VI
ACC (%) AND ASR (%) OF RELAXED SPATIAL ADVERSARIAL TRANSFORMATION ON BADNETS AND LC

Backdoor / Poison Rate	BadNets, 1%		BadNets, 2%		LC, 1%		LC, 2%	
Method	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
No Defense	94.68	100.00	95.00	100.00	94.47	88.64	94.90	95.08
Spatial AT	82.82	2.76	82.77	2.66	82.87	2.26	82.70	3.14
Adv. + Rand. ($\beta = 0.9$)	83.44	2.16	83.58	2.48	83.73	1.92	83.62	3.67
Adv. + Rand. ($\beta = 0.8$)	84.53	2.52	84.84	74.66	84.83	1.82	84.68	39.28
Adv. + Rand. ($\beta = 0.7$)	87.47	98.55	87.00	96.77	87.25	94.21	86.80	97.89
Adv. + Rand. ($\beta = 0.6$)	88.74	99.96	87.29	99.68	87.99	98.09	87.55	99.10

TABLE VII
ACC (%) AND ASR (%) OF VARIOUS BACKDOOR DEFENSE METHODS ON CIFAR-10. THE LOWEST ASR IS INDICATED IN BOLDFACE AND THE SECOND-LOWEST ASR IS INDICATED WITH AN UNDERLINE

	No defense		FP		NAD		DPSGD		ABL		CAT	
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	94.68	100.00	94.14	99.92	89.35	10.19	71.40	99.90	79.58	94.28	79.89	**3.47**
LC	94.47	88.64	85.46	**0.00**	79.81	0.05	71.24	99.40	83.44	**0.00**	80.28	6.89
Blended	94.50	98.47	94.21	71.21	89.45	10.02	70.41	67.93	78.57	47.70	80.24	**4.93**
WaNet	93.63	95.04	88.53	92.93	84.87	**2.28**	69.22	57.60	80.04	98.59	79.95	3.04
Trojan	94.54	97.56	94.11	85.07	89.48	1.97	71.40	1.76	84.19	28.93	79.76	**0.94**
SIN	94.03	82.01	84.95	**0.00**	82.83	2.17	73.82	64.61	78.39	2.35	79.75	54.89
Avg. Drop (↓)	-	-	**4.08**	35.43	8.34	**89.17**	23.06	28.42	13.61	48.31	14.33	81.26

TABLE VIII
ACC (%) AND ASR (%) OF VARIOUS BACKDOOR DEFENSE METHODS ON CIFAR-100. THE LOWEST ASR IS INDICATED IN BOLDFACE AND THE SECOND-LOWEST ASR IS INDICATED WITH AN UNDERLINE

	No defense		FP		NAD		DPSGD		ABL		CAT	
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	76.65	99.72	62.82	48.91	63.97	20.06	24.47	99.87	64.73	**0.00**	55.49	1.59
LC	75.92	98.71	59.44	**0.00**	60.8	0.19	23.84	95.88	67.75	85.95	55.64	1.87
Blended	75.84	94.45	59.76	14.25	57.86	**0.54**	21.37	21.63	65.22	0.69	54.91	7.49
WaNet	74.11	96.03	60.98	57.25	63.12	39.18	20.60	43.10	63.16	80.59	54.53	**3.90**
Avg. Drop (↓)	-	-	14.88	67.13	14.19	82.24	53.06	32.11	**10.42**	55.42	20.49	**93.52**

demonstrates the effectiveness of the composite strategy. Compared with FP and NAD, a major advantage is that CAT does not need extra clean data, which leads to wider applications as clean data may be hard to collect in some areas. Compared with DPSGD and ABL, CAT achieves more stable and better results in terms of backdoor robustness. CAT decreases almost all backdoor ASR lower than 10% except SIN (the ASR of SIN is also largely decreased). We attribute the results to the difference in the technical strategy. Although the three methods (DPSGD, ABL, and CAT) aim to train clean models with poisoned data from scratch, ABL identifies the candidates of poisoned data in the early training stage

TABLE IX

ACC (%) AND ASR (%) OF VARIOUS BACKDOOR DEFENSE METHODS ON IMAGENET SUBSET. THE LOWEST ASR IS INDICATED IN BOLDFACE AND THE SECOND-LOWEST ASR IS INDICATED WITH AN UNDERLINE

	No defense		FP		NAD		ABL		CAT	
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	87.92	89.17	86.14	7.65	88.33	86.89	71.8	79.23	81.22	**1.75**
Blended	88.51	98.42	85.61	77.70	87.72	93.70	72.76	27.33	81.30	**1.67**
Avg. Drop (↓)	-	-	2.34	51.12	**0.19**	3.50	15.94	40.52	6.96	**92.09**

TABLE X

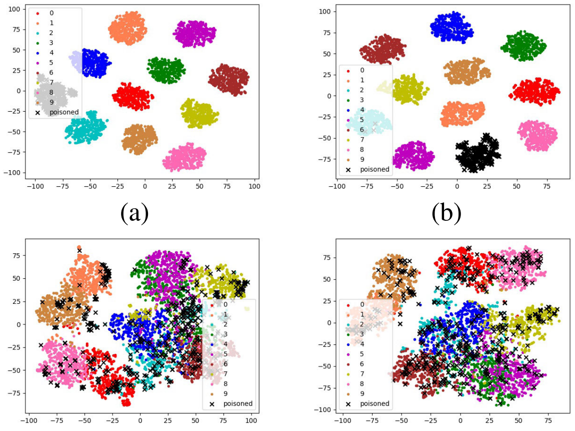
ACC (%) AND ASR (%) OF CAT WITH DIFFERENT β VALUES ON CIFAR-10

	No defense		CAT ($\beta = 1.0$)		CAT ($\beta = 0.9$)		CAT ($\beta = 0.8$)	
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	94.68	100.00	77.82	2.97	79.11	3.11	79.89	3.47
LC	94.47	88.64	78.02	3.34	79.00	4.24	80.28	6.89
Blended	94.50	98.47	78.47	4.87	79.51	4.74	80.24	4.93
WaNet	93.63	95.04	78.25	3.00	78.77	3.44	79.95	3.04
Trojan	94.54	97.56	78.34	1.61	79.04	1.09	79.76	0.94
SIN	94.03	82.01	77.94	53.74	78.75	62.60	79.75	54.89
Avg. Drop (↓)	-	-	16.17	**82.03**	15.28	80.42	**14.33**	81.26

TABLE XI

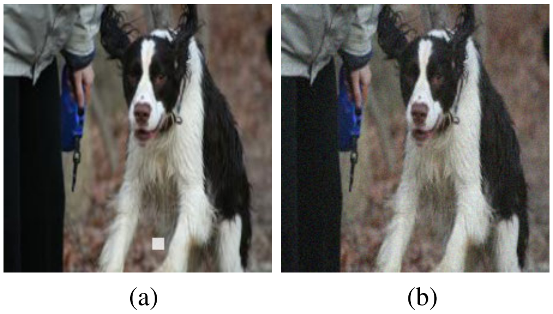
ACC (%) AND ASR (%) OF BADNETS WITH DIFFERENT POISON RATES ON CIFAR-10

	No defense		CAT ($\beta = 1.0$)		CAT ($\beta = 0.9$)		CAT ($\beta = 0.8$)	
Poison rate	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
1%	94.68	100.00	77.82	2.97	79.11	3.11	79.89	3.47
5%	94.41	100.00	78.80	23.81	79.53	19.66	80.22	54.64
10%	94.18	100.00	78.93	32.32	78.91	40.90	79.71	79.22
Avg. Drop (↓)	-	-	15.91	**80.30**	15.24	78.78	**14.48**	54.22



(a) (b) (c) (d)

Fig. 5. t-SNE visualizations of standardly trained models and CAT models. (a) BadNets, ST. (b) Blended, ST. (c) BadNets, CAT. (d) Blended, CAT.



(a) (b)

Fig. 6. Illustrations of poisoned samples on ImageNet. (a) BadNets. (b) Blended.

and forgets them later. However, the ACC of detecting the poisoned data and the gradient ascent used in ABL tend to cause the training instability, which will not happen in CAT as we only perturb the training data with imperceptible noise. DPSGD perturbs gradients with noise to minimize the difference between clean gradients and poisoned ones. The perturbation leads to a significant drop in the clean ACC, and yet does not provide meaningful guarantees. One limitation of CAT is that the adversarial perturbations lead to clean ACC drop, which is universal in AT methods, and we leave the improvements for our future work.

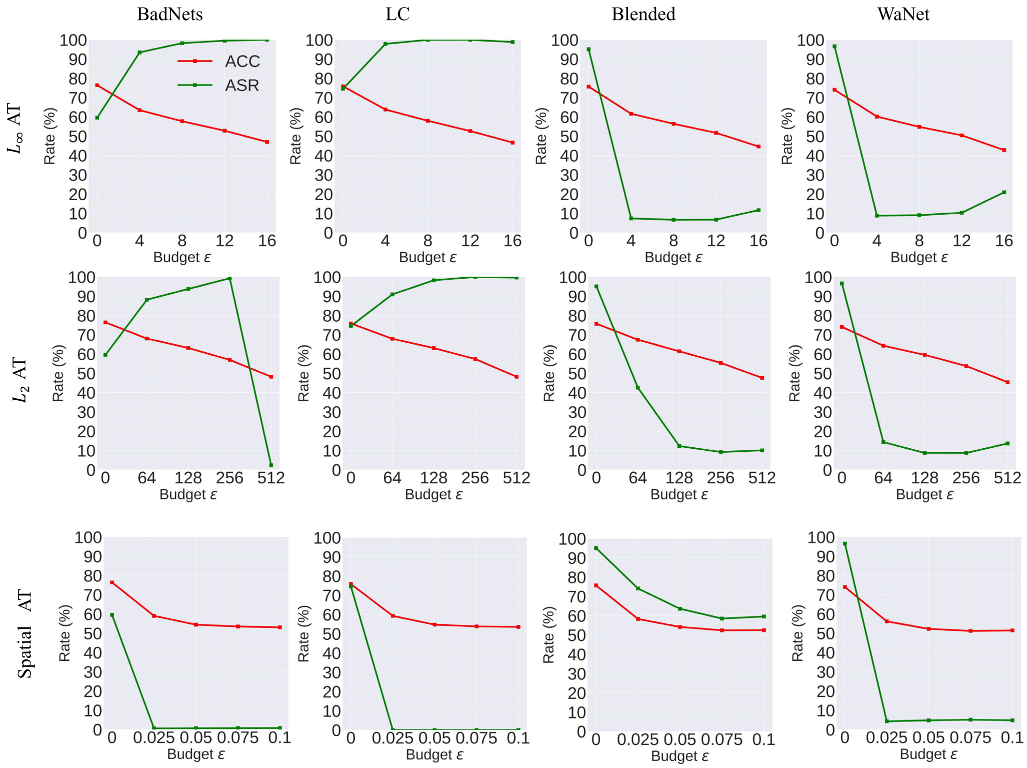


Fig. 7. Results on various backdoor attacks with various ATs on CIFAR-100.

4) Experimental on CIFAR-100 and ImageNet Subset: For CIFAR-100, we evaluated four backdoor attacks: BadNets with a 3 × 3 checkerboard trigger, LC, blended with a Hello-Kitty trigger, and WaNet. We adopted the clean label setting for BadNets while three other attacks were implemented based on the original papers. The poison rate was 0.3% for BadNets and LC, 3% for blended, and 5% for WaNet. For all the attacks, class 2 was assigned as the target class. For ImageNet subset, we evaluated two backdoor attacks: BadNets and blended. The poison rate is 5% for BadNets and blended. The results are summarized in Tables VIII and IX, from which we could find that CAT also shows competitive results over baseline methods.

5) Experiments With Various β Values: We also conducted experiments with $\beta = 1.0$ and $\beta = 0.9$. The results are summarized in Table X, from which we could find that CAT with $\beta = 1.0$ and $\beta = 0.9$ also effectively mitigates backdoor attacks but clean accuracies are slightly lower than those of CAT with $\beta = 0.8$.

6) Experiments With Various Poison Rates: We also conducted experiments on BadNets with various poison rates. The results are summarized in Table XI, from which we could find that when poison rate is increased to 5% or 10%, CAT could not completely mitigate backdoor behavior but also largely decrease backdoor ASRs.

7) t-SNE Visualizations: To further understand the model's internal response with respect to backdoor triggers, we provide the t-SNE visualizations [30] of standardly trained (ST) models and CAT models with BadNets and blended attacks on the CIFAR-10 dataset. The training configurations are the same with those in Section III, and the results are illustrated in Fig. 5. As shown in Fig. 5, we could find that the poisoned samples form an individual cluster in ST models, yet fail in CAT models.

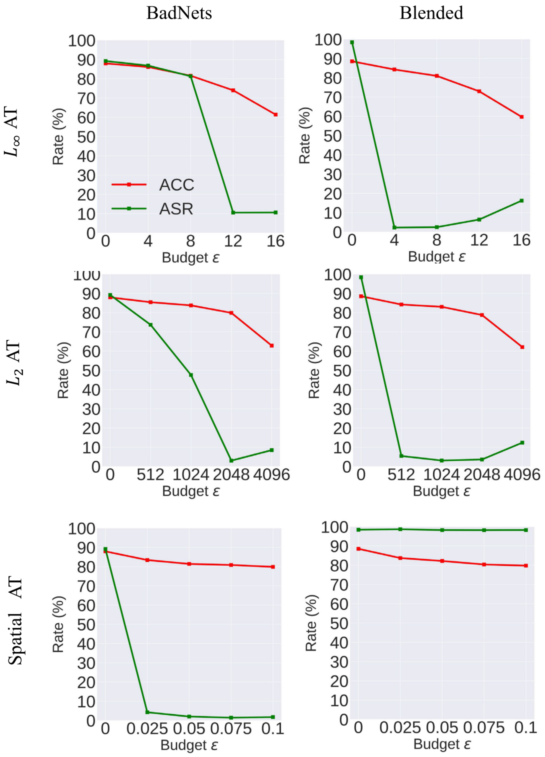


Fig. 8. Results on various backdoor attacks with various ATs on ImageNet subset.

V. Conclusion and Societal Impact

In this work, we conducted thorough experiments to investigate the effects of AT on backdoor attacks. We found that previous studies overlooked the influences of the perturbation type and budget of AT. Furthermore, we demonstrated that AT effectively mitigates backdoor attacks across various cases. We then explored the potential relaxation of adversarial perturbation and proposed composite AT to address unknown backdoor attacks. Through extensive experiments, we provided several important insights when tackling backdoor attacks with AT and itemized them in the table. We believe that our work sheds light on the understanding of the interactions between AT and backdoor attacks and reminds researchers that AT is still an effective defense against backdoor attacks. As for societal impact, our method will benefit backdoor robustness and prevent the adversary from injecting triggers. However, we do not want this article to bring an overly optimistic view of AI safety, since the backdoor attack is only one concern while there are various potential risks including adversarial attacks, privacy breaches, and model extraction.

Messages:
(i) Spatial AT effectively mitigates patch-based backdoor attacks;
(ii) L_p AT effectively mitigates whole-image backdoor attacks;
(iii) Adversarial perturbation is superior to random perturbation and mixup augmentations;
(iv) Adversarial perturbation can be slightly relaxed to improve natural generalization but does not affect backdoor mitigation effects;
(v) We make it possible to defend against unknown backdoor attacks with merely data augmentations (i.e., integrating multiple adversarial perturbations).

Appendix

A. Experiments on CIFAR-100 and ImageNet

1) Experimental Settings for CIFAR-100:

a) Backdoor attacks: For CIFAR-100, we evaluated four backdoor attacks: BadNets with a 3 × 3 checkerboard trigger, LC, blended with a Hello-Kitty trigger ($\alpha = 0.1$), and WaNet. The poison rate was 0.5% for BadNets and LC, 3% for blended, and 5% for WaNet. For all the attacks, class 2 was assigned as the target class.

b) AT budgets: The perturbation budget for L_∞ AT ranges from $4/255$ to $16/255$. The perturbation budget for L_2 AT ranges from $64/255$ to $512/255$. The perturbations budget for spatial AT ranges from 0.025 to 0.1.

2) Experimental Settings for ImageNet:

a) Backdoor attacks: We randomly selected ten classes to create a subset and evaluated two backdoor attacks: BadNets with a 10 × 10 trigger and blended with a random trigger ($\alpha = 0.1$). We adopted dirty label BadNets since we find that the ASR of clean label BadNets is lower than 20%. The poison rate was 5% for BadNets and blended. For all the attacks, class 0 was assigned as the target class. The poisoned samples are illustrated in Fig. 6.

b) AT Budgets: The perturbation budget for L_∞ AT ranges from $4/255$ to $16/255$. The perturbation budget for L_2 AT ranges from $512/255$ to $4096/255$. The perturbation budget for spatial AT ranges from 0.025 to 0.1.

3) Results: The results are summarized in Figs. 7 and 8. We find that L_p AT effectively mitigates whole-image backdoor attacks and spatial AT effectively mitigates patch-based backdoor attacks, which is consistent with the conclusion in CIFAR-10.

References

[1] F. Ali et al., "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion," *Inf. Fusion*, vol. 63, pp. 208–222, Nov. 2020.

[2] Y. Bai, Y. Zeng, Y. Jiang, S.-T. Xia, X. Ma, and Y. Wang, "Improving adversarial robustness via channel-wise activation suppressing," in *Proc. ICLR*, 2021, pp. 1–19.

[3] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in CNNs by training set corruption without label poisoning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 101–105.

[4] E. Borgnia et al., "Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3855–3859.

[5] E. Borgnia et al., "DP-InstaHide: Provably defusing poisoning and backdoor attacks with differentially private data augmentations," 2021, *arXiv:2103.02079*.

[6] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," 2021, *arXiv:2106.09667*.

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

[8] S. Ding, Y. Tian, F. Xu, Q. Li, and S. Zhong, "Trojan attack on deep generative models in autonomous driving," in *Proc. Int. Conf. Secur. Privacy Commun. Syst.*, 2019, pp. 299–318.

[9] J. Geiping, L. Fowl, G. Somepalli, M. Goldblum, M. Moeller, and T. Goldstein, "What doesn't kill you makes you Robust(ER): How to adversarially train against data poisoning," 2021, *arXiv:2102.13624*.

[10] M. Goldblum et al., "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," 2020, *arXiv:2012.10544*.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015, pp. 1–11.

[12] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," 2017, *arXiv:1708.06733*.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[14] S. Hong, V. Chandrasekaran, Y. Kaya, T. Dumitraş, and N. Papernot, "On the effectiveness of mitigating data poisoning attacks with gradient shaping," 2020, *arXiv:2002.11497*.

[15] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," in *Proc. NeurIPS*, 2021, pp. 14900–14912.

[16] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," in *Proc. ICLR*, 2021, pp. 1–19.

[17] Y. Li, Y. Bai, Y. Jiang, Y. Yang, S.-T. Xia, and B. Li, "Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection," in *Proc. NeurIPS*, 2022, pp. 1–26.

[18] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," 2020, *arXiv:2007.08745*.

[19] Y. Li, H. Zhong, X. Ma, Y. Jiang, and S.-T. Xia, "Few-shot backdoor attacks on visual object tracking," in *Proc. ICLR*, 2022, pp. 1–11.

[20] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, nos. 1–3, pp. 503–528, Aug. 1989.

[21] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. RAID*, 2018, pp. 273–294.

[22] Y. Liu et al., "Trojaning attack on neural networks," in *Proc. 25th Annu. Netw. Distrib. Syst. Secur. Symp.*, 2018, pp. 1–15.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018, pp. 1–11.

[24] P. Maini, E. Wong, and Z. Kolter, "Adversarial robustness against the union of multiple perturbation models," in *Proc. ICML*, 2020, pp. 6640–6650.

[25] A. Nguyen and A. Tran, "Wanet–imperceptible warping-based backdoor attack," in *Proc. ICLR*, 2021, pp. 1–16.

[26] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of tricks for adversarial training," in *Proc. ICLR*, 2021, pp. 1–21.

[27] L. Tao, L. Feng, J. Yi, S.-J. Huang, and S. Chen, "Better safe than sorry: Preventing delusive adversaries with adversarial training," in *Proc. NeurIPS*, 2021, pp. 16209–16225.

[28] F. Tramer and D. Boneh, "Adversarial training and robustness for multiple perturbations," in *Proc. NeurIPS*, 2019, pp. 1–11.

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

[30] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[31] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *Proc. ICLR*, 2020, pp. 1–11.

[32] M. Weber, X. Xu, B. Karlaš, C. Zhang, and B. Li, "RAB: Provable robustness against backdoor attacks," 2020, *arXiv:2003.08904*.

[33] C.-H. Weng, Y.-T. Lee, and S.-H. B. Wu, "On the trade-off between adversarial and backdoor robustness," in *Proc. NeurIPS*, 2020, pp. 11973–11983.

[34] D. Wu, S.-T. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," in *Proc. NeurIPS*, 2020, pp. 2958–2969.

[35] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," in *Proc. ICLR*, 2018, pp. 1–11.

[36] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6022–6031.

[37] Y. Zeng, S. Chen, W. Park, Z. M. Mao, M. Jin, and R. Jia, "Adversarial unlearning of backdoors via implicit hypergradient," 2021, *arXiv:2110.03735*.

[38] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. ICML*, 2019, pp. 1–12.

[39] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[40] J. Zhang et al., "Attacks which do not kill training make adversarial learning stronger," in *Proc. ICML*, 2020, pp. 11278–11287.

[41] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," in *Proc. ICLR*, 2021, pp. 1–29.

Yinghua Gao received the B.S. degree from the Department of Mathematics, Nankai University, Tianjin, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University, Shenzhen, China.

His research interests are primarily on trustworthy machine learning.

Dongxian Wu received the bachelor's degree in microelectronics from Xidian University, Hangzhou, China, in 2016, and the Ph.D. degree in computer science and technology from Tsinghua University, Shenzhen, China, in 2021.

He is currently a Post-Doctoral Researcher at The University of Tokyo, Tokyo, Japan. He focuses on trustworthy machine learning, especially adversarial learning and data security.

Jingfeng Zhang received the bachelor's degree in computer science from Taishan College, Shandong University, Qingdao, China, in 2016, and the Ph.D. degree in computer science from the School of Computing, National University of Singapore, Singapore, in 2020.

He is currently a Post-Doctoral Researcher at RIKEN-AIP, Tokyo, Japan. He has worked extensively in machine learning security, specifically in adversarial machine learning. His long-term research interest is making artificial intelligence safe for human beings.

Guanhao Gan received the B.S. degree in computer science and technology from the Harbin Institute of Technology, Shenzhen, China, in 2021. He is currently pursuing the M.Eng. degree in electrical engineering and computer technology with SIGS, Tsinghua University, Shenzhen, China.

His main research interests include trustworthy machine learning and deep model copyright protection.

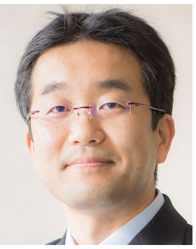
Shu-Tao Xia received the B.S. degree in mathematics and the Ph.D. degree in applied mathematics from Nankai University, Tianjin, China, in 1992 and 1997, respectively.

From March 1997 to April 1999, he was with the Research Group of Information Theory, Department of Mathematics, Nankai University. Since January 2004, he has been with Shenzhen International Graduate School, Tsinghua University, Guangdong, China, where he is currently a Full Professor. His current research interests include coding and information theory, networking, and machine learning.

Gang Niu received the Ph.D. degree in computer science from the Tokyo Institute of Technology, Tokyo, Japan, in 2013.

He is currently an Indefinite-Term Research Scientist at RIKEN Center for Advanced Intelligence Project, Fukuoka, Japan. Before joining RIKEN as a Research Scientist, he was a Senior Software Engineer at Baidu, Beijing, China, and then an Assistant Professor at The University of Tokyo, Tokyo. He has authored more than 90 journal articles and conference papers, including 31 ICML, 17 NeurIPS (one oral and three spotlights), and 11 ICLR (one outstanding paper honorable mention, two orals, and one spotlight) papers. He has coauthored the book titled *Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach* (MIT Press).

Dr. Niu has served as an Area Chair 17 times, including ICLR 2021–2022, ICML 2019–2022, and NeurIPS 2019–2022. He also serves/has served as an Action Editor for TMLR and a Guest Editor of a special issue at MLJ. Moreover, he has served as a Publication Chair for ICML 2022 and co-organized nine workshops, one competition, and two tutorials.



Masashi Sugiyama (Senior Member, IEEE) received the Ph.D. degree in computer science from the Tokyo Institute of Technology, Tokyo, Japan, in 2001.

After experiencing as an Assistant Professor and an Associate Professor at the Tokyo Institute of Technology, Tokyo, he became a Professor at The University of Tokyo, Tokyo, in 2014. Since 2016, he has concurrently served as the Director of RIKEN Center for Advanced Intelligence Project, Fukuoka, Japan. His current research interests include theories and algorithms of machine learning.

Dr. Sugiyama was a recipient of the Japan Academy Medal in 2017 and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in Japan, in 2022.