Supplementary Material of "BackdoorBench: A Comprehensive Benchmark of Backdoor Learning"

Baoyuan Wu[1]* **Hongrui Chen**[1] **Mingda Zhang**[1] **Zihao Zhu**[1]
Shaokui Wei[1] **Danni Yuan**[1] **Chao Shen**[2]
[1]School of Data Science, Shenzhen Research Institute of Big Data,
The Chinese University of Hong Kong, Shenzhen
[2]School of Cyber Science and Engineering, Xi'an Jiaotong University

A Additional information of backdoor attack and defense algorithms

A.1 Descriptions of backdoor attack algorithms

In addition to the basic information in Table 1 of the main manuscript, here we describe the general idea of eight implemented backdoor attack algorithms in BackdoorBench, as follows.

- **BadNets** [6]: It was the first work in backdoor learning, which simply inserted a small patch with fixed pattern and location to replace the original pixels in the clean image to obtain a poisoned image.
- **Blended** [3]: It extended BadNets by encouraging the invisibility of the trigger through alpha blending.
- **Label consistent (LC)** [20]: It generated a poisoned image using adversarial attack, by enforcing it to be close to the clean target image in the original RGB space, and close to the clean source image patched with a trigger in the feature space of a pre-trained clean model. Since the poisoned image is labeled as the target class, the mapping from the trigger to the target class could be learned.
- **SIG** [1]: It adopted a sinusoidal signal as the trigger to perturb the clean images of the target class, while not changing their labels, such that achieving the label consistent backdoor attack.
- **Low frequency attack (LF)** [30] : It was built upon an analysis that the triggers in many backdoor attacks bring in high-frequency artifacts, which are easily detectable. Inspired by this analysis, LF developed a smooth trigger by filtering high-frequency artifacts from a universal adversarial perturbation.
- **Sample-specific backdoor attack (SSBA)** [10]: It utilized an auto-encoder to fuse a trigger ($e.g.$, a string) into clean samples to obtain poisoned samples. The residual between the poisoned and the clean sample varied for different clean images, $i.e.$, sample-specific.
- **Input-aware dynamic backdoor attack (Input-aware)** [14]: It was a training-controllable attack by simultaneously learning the model parameters and a trigger generator. When testing, the learned trigger generator generated one unique trigger for each clean testing sample.
- **Warping-based poisoned networks (WaNet)** [15]: It was also a training-controllable attack. A fixed warping function is adopted to slightly distort the clean sample to construct the poisoned sample. The attacker further controlled the training process to ensure that only the adopted fixed warping function can activate the backdoor.

*Corresponds to Baoyuan Wu (wubaoyuan@cuhk.edu.cn).

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

A.2 Descriptions of backdoor defense algorithms

In addition to the basic information in Table 2 of the main manuscript, here we describe the general idea of nine implemented backdoor defense algorithms in BackdoorBench, as follows.

- **Fine-tuning (FT)**: It is assumed that fine-tuning the backdoored model on a subset of clean samples could mitigate the backdoor effect. Note that FT is a widely used approach for transferring pre-trained models to new tasks, but it has been used as a basic component in several backdoor defense methods, such as Fine-pruning (FP) [12], Neural Attention Distillation (NAD) [8].
- **Fine-pruning (FP)** [12]: It is built upon the assumption that *poisoned and benign samples have different activation paths in the backdoored model*. Inspired, FP proposed to firstly prune some inactivated neurons of clean samples, then fine-tune the pruned model based on the subset of benign samples to recover the model performance.
- **Neural Attention Distillation (NAD)** [8]: Its assumption is same with FT. Instead of directly use the fine-tuned model as the mitigated model, NAD adopts the first fine-tuned model as a teacher, and fine-tunes the backdoored model again by encouraging the consistency of the attention representation between the new fine-tuned model and the teacher model.
- **Neural cleanse (NC)** [26]: It is built upon the assumption that *the trigger provides a "shortcut" between the samples from different source classes and the target class*. Based on this assumption, the possible trigger is searched through optimization. If a small-size trigger ($e.g.$, a small patch in the image) is found, then the model is detected as backdoored model, which is then mitigated through pruning based on the searched trigger.
- **Adversarial Neuron Pruning (ANP)** [27]: It is built upon an observation that *the neurons related to the injected backdoor are more sensitive to adversarial neuron perturbation ($i.e.$, perturbing the neuron weight to achieve adversarial attack) than other neurons in a backdoored model*. Inspired by this, ANP proposed to prune these sensitive neurons for backdoor mitigation.
- **Activation Clustering (AC)** [2]: It is built upon an observation that *the sample activations ($i.e.$, the feature presentations) of the target class will form two clusters, and the smaller cluster corresponds to poisoned samples, while those of other classes form one cluster*. Then, the model is trained from scratch based on the dataset without poisoned samples.
- **Spectral Signatures (SS)** [24]: Its assumption is that *the feature representation distributions of benign and poisoned samples in one class are spectrally separable*, which is a concept of robust statistics. Consequently, the poisoned samples can be identified through analyzing the spectrum of the covariance matrix of the feature representations. Then, the model is retrained from scratch by removing the poisoned samples from the training set.
- **Anti-Backdoor Learning (ABL)** [9]: It is built upon an observation that *the loss values of poisoned samples drops much faster than those of benign samples in early epochs during the training process*. Inspired, ABL proposed to firstly isolate poisoned samples from benign samples according to their difference on loss dropping speed, then mitigate the backdoor effect by maximizing the loss of the isolated poisoned samples.
- **Decoupling-based Backdoor Defense (DBD)** [7]: It is built upon an observation that *poisoned samples from different samples will gather together in the feature space of a backdoored model*. DBD proposed to prevent the gathering by learning the model backbone through self-supervised learning without labels, rather than the standard supervised learning. Then, since the poisoned samples are separated, their loss values are larger than benign samples when learning the classifier, such that samples with large loss values can be identified as poisoned samples. Finally, the labels of poisoned samples are abandoned, and a semi-supervised fine-tuning of both the backbone and classifier is conducted to improve the model performance.

[1]For FP and ANP, we define a hyper-parameter *the tolerance of clean accuracy reduction* as the maximum relative drop of clean accuracy. It is used to determine the number of pruned neurons.

[2]For AC and Spectral, N_{fe} is the dimensions of the representation.

[3]For NAD, we use the code and recommended hyper-parameters at https://github.com/bboylyg/NAD/tree/d61e4d74ee697f125336bfc42a03c707679071a6.

Table 1: Hyper-parameter settings of all implemented attack methods.

Attack	Hyper-parameter	Setting	Theoretical complexity
General Settings	attack target	all-to-one with class 0	train sample size($N_{\mathcal{T}}$), batch size (B), forward process for net C (f_C), backward process for net C (b_C), epochs (E) , number of classes (N_{cls})
	optimizer	SGD	
	momentum	0.9	
	weight decay	0.0005	
	batch size	128	
	epochs on CIFAR10 and CIFAR-100	100	
	epochs on GTSRB	50	
	lr schedule (except for training-controllable attack) on CIFAR10, CIFAR-100 and GTSRB	CosineAnnealingLR	
	epochs on TinyImageNet-200	200	
	lr schedule on TinyImageNet-200	ReduceLROnPlateau	
	random seed	0	
BadNets [6]	pattern & location	3×3, pure white, at downright corner (no margin left)	$O(N_{\mathcal{T}}/BE(f+b))$
Blended [3]	pattern	hello kitty	$O(N_{\mathcal{T}}/BE(f+b))$
	alpha	0.2	
Label Consistent [25]	adversarial attack	PGD	$O(N_{\mathcal{T}}/BE(f+b) + N_{\mathcal{T}}/BT_{step}(f+b))$
	step (T_{step})	100	
	α	1.5	
	ϵ	8	
SIG [1]	δ	40	$O(N_{\mathcal{T}}/BE(f+b))$
	f	6	
Low Frequency [30]	maximum number of termination iteration (T_{ter})	50	$O(T_{ter}N_{sample}(2f + T_{df}f + T_{df}N_{cls}b)+ N_{\mathcal{T}}/BE(f+b))$
	fooling rate	0.2	
	overshoot	0.02	
	maximum number of iterations for deepfool (T_{df})	200	
	sample number for UAP (N_{sample})	100	
SSBA [10]	encoded bit	1	$O(T_{step}(f_{auto} + b_{auto}) + N_{\mathcal{T}}/BE(f+b))$
	autoencoder train step (T_{step})	140000	
Input-aware [14]	Generator lr (both for M, G)	0.01	$O(E_{mask}N_{\mathcal{T}}/B(f_{mask} + b_{mask})+ (E - E_{mask})N_{\mathcal{T}}/B(2f_{mask}+ 2f_{generator} + 2b_{generator} + f + b))$
	schedule (for M, G and C)	MultiStepLR	
	schedule milestones for G	200, 300, 400, 500	
	schedule milestones for C	100, 200, 300, 400	
	schedule milestones for M	10, 20	
	schedule gamma (for M and G)	0.1	
	λ_{div}	1	
	λ_{norm}	400	
	mask density	0.032	
	cross_ratio	1	
	mask train epochs (E_{mask})	25	
WaNet [15]	cross_ratio	2	$O(N_{\mathcal{T}}/BE(f+b))$
	lr schedule	MultiStepLR	
	schedule milestones	100, 200, 300, 400	
	grid_rescale	1	

A.3 Implementation details and computational complexities

Running environments Our evaluations are conducted on GPU servers with 2 Intel(R) Xeon(R) Platinum 8170 CPU @ 2.10GHz, RTX3090 GPU (32GB) and 320 GB RAM (2666MHz). The versions of all involved softwares/packages are clearly described in the README file of the Github repository (see `https://github.com/SCLBD/BackdoorBench`). Here we didn't repeat the descriptions.

Hyper-parameter settings The hyper-parameter settings adopted in our evaluations about backdoor attack and defense algorithms are described in Table 1 and Table 2, respectively. With these hyper-

Table 2: Hyper-parameter settings of all implemented defense methods.

Defense	Hyper-parameter	Setting	Theoretical complexity
General Settings	optimizer	SGD	train sample size ($N_{\mathcal{T}}$), batch size (B),
	momentum	0.9	forward process for net C (f_C),
	weight decay	0.0005	backward process for net C (b_C),
	batch size	256	epochs (E) , number of classes (N_{cls})
	epochs on CIFAR10, CIFAR-100 and GTSRB	100	the number of pruning neurons (N_{neu})
	lr schedule (except for special learning defense) on CIFAR10, CIFAR-100 and GTSRB	CosineAnnealingLR	
	epochs on TinyImageNet-200	200	
	lr schedule on TinyImageNet-200	ReduceLROnPlateau	
	the number of pruning neurons	the number of neurons in the last layer	
	random seed	0	
FT	ratio of validation data (p_v)	5%	$O(N_{\mathcal{T}}p_v/BE(f+b))$
FP [12]	ratio of validation data (p_v)	5%	$O(N_{neu}N_{\mathcal{T}}p_v f/B)$
	the tolerance of accuracy reduction[1]	10%	$+O(N_{\mathcal{T}}p_v/BE(f+b))$
NAD [3] [8]	ratio of validation data for teacher model (p_v)	5%	$O(N_{\mathcal{T}}p_v/BE_{ft}(f+b))$
	β_1 for the loss	500	$+O(N_{\mathcal{T}}p_v/BE(2f+b))$
	β_2 for the loss	1000	
	β_3 for the loss	1000	
	the power for attention	2.0	
	the epoch for teacher model to fine-tune (E_{ft})	10	
NC [26]	the norm used for the reversed trigger	L1	$O(N_{\mathcal{T}}p_v E_r N_{cls}/B(f+b))$
	cleaning ratio (p_v)	0.05	$+O(N_{\mathcal{T}}p_v/BE(f+b))$
	unlearning ratio	0.2	
	the epoch of learning trigger (E_r)	80	
ANP [27]	the tolerance of accuracy reduction[1]	10%	$O(e_i(f+b))$
	number of validation data (p_v)	5%	
	the number of iteration during pertubation (e_i)	2000	
	ϵ	0.4	
	α	0.2	
AC [2]	number of reduced dimensions	10	$O(N_{\mathcal{T}}/BN_{fe}^3)+O(N_{\mathcal{T}}/BE(f+b))^2$
Spectral [24]	The percentile of backdoor data	85%	$O(N_{fe}^3)+O(N_{\mathcal{T}}/BE_d(f+b))^2$
ABL [9]	tuning epochs (E_{tu}) for CIFAR10, CIFAR-100 and GTSRB	20	$O(N_{\mathcal{T}}/BE_{tu}(f+b))$
	finetuning epochs (E_{ft}) for CIFAR10, CIFAR-100 and GTSRB	60	$+O(N_{\mathcal{T}}(1-p_i)/BE_{ft}(f+b))$
	unlearning epochs (E_u) for CIFAR10, CIFAR-100 and GTSRB	20	$+O(N_{\mathcal{T}}p_i/BE_u(f+b))$
	tuning epochs for Tiny	40	
	finetuning epochs for Tiny	120	
	unlearning epochs for Tiny	4	
	lr for unlearning	0.0005	
	the value of flooding	0.5	
	the isolation ratio of training data (p_i)	0.01	
DBD [7]	the epoch for self-supervised learning (E_{se})	100	$O(N_{\mathcal{T}}/B_{se}E_{se}(f+b))$
	the epoch for warmup (E_{wa})	10	$+O(N_{\mathcal{T}}/B_{semi}E_{se}(f+b))$
	the epsilon for the dataset (ϵ) during the semi-supervised learning	0.5	$+O(N_{\mathcal{T}}\epsilon/B_{semi}E(f+b))$
	The batch size of self learning (B_{self})	512	
	The batch size of self learning (B_{semi})	128	

parameter settings, the reported results of 8,000 pairs of evaluations could be reproduced. Moreover, we would like to explain our rules to adopt above settings, as follows:

- **We don't perform a separate hyper-parameter search for each method**, mainly due to the following two reasons:
  - As shown in Tables 1 and 2, most methods have several hyper-parameters. For most hyper-parameters of a method, there is neither a good rule to determine the values, nor a suitable range of the values suggested in its original manuscript. And, the suitable value or range of each hyper-parameter may vary across different datasets, different model architectures, different against attack/defense methods. Consequently, the hyper-parameter search space for each method could be very large, requiring unimaginably high computational resource.
  - Even assuming sufficient computing resources, then we can search a good value for each hyper-parameter of each method in each evaluation. However, the comparison results and analysis based on sufficient hyper-parameter search may be unfair and make no sense in practice. Because, we still cannot tell a rule or even some experiences to determine the hyper-parameter values in practice. The sensitivity to hyper-parameters should also be an important metric of one method's performance, not just the best ACC/ASR values through the sufficient hyper-parameter search.
- **How do we set the hyper-parameter values in our current 8000 pairs of evaluations.**

- If the original paper has provided the suggested good values of some hyper-parameters, then we adopt those values in our evaluations. For example, the ANP defense method explicitly wrote that "the perturbation budget $\epsilon = 0.4$ and the trade-off coefficient $\alpha = 0.2$", so we also adopt these values in our evaluations.
- For those hyper-parameters without suggested values/ranges (or even without descriptions) in their original papers, we will search values that lead to comparable results (ACC/ASR) with the reported results in the same setting ($i.e.$, same dataset, same/similar model architecture, same poisoning ratio), then fix these values in evaluations of other settings ($e.g.$, changing the poisoning ratio).
- The consistent values of hyper-parameters of each method across different settings somewhat guarantee the fairness of evaluations. And, since the adopted values may not be the optimal ones for some hyper-parameters, we didn't conduct the fine-grained analysis about the effects of some specific hyper-parameters ($e.g.$, the trigger size/location in attack methods with patch based triggers). Instead, we provided some high-level analysis *w.r.t.* the shared hyper-parameters in all methods ($e.g.$, the number of classes, the poisoning ratio, the model architecture). The findings of these high-level analysis will not be significantly affected by the particular hyper-parameters of each individual method.

Computational complexities The computational complexity of each attack and each defense algorithm is also described in Table 1 and Table 2, respectively.

B Additional evaluations and analysis

B.1 Full results on CIFAR-10

The full results on CIFAR-10 with five different poisoning ratios ($i.e.$, 10%, 5%, 1%, 0.5%, 0.1%) are presented in Tables 10 – 14, respectively. The remaining results on other datasets and model architectures among 8,000 attack-defense pairs of evaluations are presented in the Leaderborad in the BackdoorBench website (see `https://backdoorbench.github.io`).

B.2 Results overview

In Figure 1, we present the performance distribution of attack-defense pairs on Preact-ResNet18 and VGG-19 with two poisoning ratios of 5% and 10%, respectively. As we mentioned in Section 4.2, if we measured the effectiveness of methods by clean accuracy (C-Acc) and ASR, a perfect attack method should be located at the top-right corner; the perfect defense method should show in the top-left corner. If we measured robust accuracy (R-Acc) and ASR, the reduced ASR value would be desirable to equal the increased R-Acc. This defense method can recover the correct prediction and eliminate the backdoor successfully. Even if we change the model structure from Preact-ResNet18 to VGG-19, the conclusion coincides with our analysis. Besides, with the increase in poisoning ratio, some color patterns are closer to the anti-diagonal line, which means these defense methods can achieve better performance in this situation. Please refer to Section 4.2 in the manuscript for the analysis.

B.3 Effect of dataset

As shown in Figure 2, we make a detailed comparison of the performance of attacks and defenses under different datasets using the PreAct-ResNet18 model and 5% poison ratio. Where the different colored bars correspond to the four datasets, the height of the bars represents the ASR, and the various subplots correspond to the multiple defenses (and no defenses). Looking down from the undefended perspective, we can see that, by and large, the effect of the attack fluctuates across the different datasets. Blended is the most stable across datasets, while BadNets has the most fluctuating effect across datasets. For BadNets, we find that CIFAR-100 and GTSRB are more complex than CIFAR-10, which leads to the decrease in effectiveness on these two datasets, but the ASR on Tiny ImageNet has rebounded significantly due to the enlargement of the trigger size. From different defense perspectives, we can find that the two methods, AC and Spectral Signature, are relatively unaffected by changes in the dataset compared with each other. In contrast, the rest of the defense methods may all have large fluctuations in their effectiveness in the face of specific attacks. Although

fluctuating, ANP has better results on CIFAR-10 for all attack methods, while ABL is also very effective on Tiny ImageNet for all attack methods.

B.4 Effect of poisoning ratio

B.4.1 Effect of poisoning ratio with randomness

As demonstrated in Section 4.3 in the main manuscript, in the following we will further verify the abnormal phenomenon of poisoning ratio's effect shown in Figure 3 in the main manuscript, under the randomness of weight initialization and some methods' mechanisms.

Experimental setting In the reported 8,000 pairs of evaluations, we set the random seed as 0 to fix all randomness in each evaluations, to ensure all results could be reproduced. For each evaluation plotted in Figure 3 in the main manuscript, we re-run the script with five different random seeds, and record the mean and the standard deviation of these five evaluations.

Analysis As shown in Figure 3, the trends of ASR curves are almost consistent with those in Figure 3 in the main manuscript, and the standard deviation ($i.e.$, the error bar) is small, indicating that the abnormal phenomenon about the poisoning ratio's effect is not affected by the randomness. However, there are still a few special cases. For example, the error bars of some attacks under the ABL defense is very large when the poisoning ratio is low. The reason is that ABL identifies the fixed 1% of all training samples as the poisoning samples according to the training loss. However, we find that the poisoning identification accuracy is very unstable, especially when the poisoning ratio is low, leading to the large fluctuation. The standard deviations of evaluations under the NC defense are also large. As described in Section **??**, NC consists of two consecutive steps, $i.e.$, firstly searching a candidate trigger to determine whether it is a backdoored model or not, then mitigating the backdoor effect through pruning. We observe that the first step is very unstable within 5 random evaluations. If the backdoored model is successfully detected, then the ASR will be reduced significantly, other keeping the high value, causing the high standard deviations of 5 random evaluations.

B.4.2 Effect of poisoning ratio of other model architectures

In this part, we intend to give the more detailed information about the variation of ASR values against poisoning ratio on different model structures, which are VGG-19, DenseNet-161, EfficientNet-B3, and MobileNetV3-Large, respectively. The corresponding results are established in Figure 4. We have analyzed the effect of poisoning ratio in Section 4.3 in the main manuscript based on the results of Preact-ResNet18 on CIFAR-10. We found that the most ASR curves increase with the increase of poisoning ratio, while there are some curves which increase at first and then collapse dramatically. However, this phenomenon still exists for multiple model structures. It is interesting to notice that if the model structure is changed, the tendency of curves is different. The curve of NAD against BadNets can serve as an example. It keeps increasing in DenseNet-161, while increases at first and then drops down in VGG-19 and MobileNetV3-Large. Thus, it is valuable to further explore the relationship between model architecture and backdoor performance. Note that we don't provide the results of ANP and DBD on VGG-19, as we adopt the VGG-19 architecture without the batch normalization (BN) layer (see the demonstration in Section 4.1 of the main manuscript). According to the ANP author's comments at https://github.com/csdongxian/ANP_backdoor/issues/2, ANP is not suitable to the model architecture without the BN layer. Besides, in our evaluations, the defense performance of DBD is not very stable on the VGG-19 without the BN layer, at its semi-supervised learning phase. Thus, we also don't report the evaluation of DBD on VGG-19. However, we also observe that DBD performs stably on the VGG-19 model with the BN layer. The behind reason will be explored in future.

B.5 Sensitivity to hyper-parameters

As illustrated in Section **??**, we adopt a fixed setting for each attack/defense method to ensure reproducibility and fair comparison. However, the sensitivity to hyper-parameters is also a very critical metric of one method's performance and practical usage. In the following, we pick three attack methods ($i.e.$, BadNets, SIG, InputAware), two defense methods ($i.e.$, ABL and ANP), two datasets ($i.e.$, CIFAR-10 and GTSRB), and two models ($i.e.$, PreAct-ResNet18 and VGG-19), to present a

partial analysis about the sensitivity to hyper-parameters. For each attack/defense method, we study one key hyper-parameter, such as the trigger's patch size for BadNets, the trigger's frequency for SIG, the mask density for Input-aware, the flooding value for ABL and the poisoning threshold for ANP.

Results are shown in Tables 3, 4, 5, and 6. For BadNets, a larger square pattern means a more vigorous attack but is also easier to find by defense methods. For SIG, higher frequency means stronger attacks and harder to defend. For Input-aware, the ASR values fluctuate a lot *w.r.t.* the mask density in the case of no defense but are relatively stable under defenses. For ABL, a higher flooding parameter often leads to worse defense performance for most attacks. For ANP, the situation is complicated. When defending BadNets and SIG, the higher threshold often leads to better defense performance, but the defense performance against Input-aware is rather stable *w.r.t.* the threshold. When comparing the performance across different model architectures and different datasets, we find that the sensitivities to hyper-parameters of each method are very diverse. Picking a good hyper-parameter for a backdoor learning method in practice is a challenge.

B.6 Analysis of quick learning of backdoor

The quick learning phenomenon of backdoor has been observed in some previous works [9], $i.e.$, the backdoor could be quickly learned in a few epochs, for almost all backdoor attacks. However, the behind reason has not been studied. In the following, we provide a detailed analysis from the perspective of gradient. Specifically, for each epoch during the training process, we record the following information:

- Losses of training samples, clean testing samples, and poisoned testing samples;
- Accuracy on training samples, clean testing samples, and poisoned testing samples;
- Gradient signal to noise ratios (GSNR) [11] on training samples, clean train samples, and poisoned training samples averaged over model parameters;
- Norms of average gradient on total training samples, clean training samples, and poisoned training samples;
- Pairwise cosine similarities between average gradients on total training samples, clean training samples, and poisoned training samples.

As shown in Figure 5, we report the results of 5 backdoor attacks, including BadNets, Blended, SSBA, LC and LF with poisoning ratio 10%, on the CIFAR-10 dataset and the PreAct-ResNet18 model. As shown in the first column, the testing loss of poisoned samples drops quickly in the early stages of training and converges to a low value, while the testing loss of clean samples drops at a slower rate and converges to a much larger value. It verifies the quick learning phenomenon of backdoor under these five backdoor attack methods. As shown in the third column, we first observe that the GSNR of poisoned samples is significantly larger than the GSNR of clean samples at the early stages. The high GSNR values of poisoned samples indicate that the backdoor has better generalization performance and is consistent with the higher accuracy (ASR) and lower loss on poisoned testing samples. Secondly, we notice that the norm of the gradient on poisoned samples is much larger than the norm of the gradient on clean samples in early epochs, as shown in the fourth column. Consequently, the cosine similarity between gradients on total training samples and poisoned training samples is significantly larger than the cosine similarity between gradients on clean training samples and poisoned training samples, though the number of poisoned samples is much smaller than the number of clean samples.

B.7 Analysis of backdoor forgetting

From the above analysis about the quick learning of backdoor in Section B.6, we get one impression that the backdoored model memorizes the poisoned samples quickly and stably. To obtain more insights about the inner mechanism of backdoor learning, we adopt the concept of *forgetting event* [23] to characterize the learning dynamics during the training process. Specifically, one forgetting event is recorded when a correctly predicted training sample at the current epoch is incorrectly predicted at the next epoch. Formally, given a training sample $(\mathbf{x}, y)$, where $\mathbf{x}$ is input feature, y is ground-truth label. If $\mathbf{x}$ is correctly predicted at the epoch t, *i.e.*, $f_{\theta^t}(\mathbf{x}) = y$, but is misclassified at epoch $t+1$, *i.e.*, $f_{\theta^{t+1}}(\mathbf{x}) \neq y$, where f denotes the model and θ^t, θ^{t+1} are model parameters at the epoch t and $t+1$, then a forgetting event is recorded for this sample.

Specifically, we count the number of forgetting events for clean and poisoned training samples, respectively, on CIFAR-10 with Preact-ResNet18 backbone. The distributions of forgetting events of clean and poisoned samples are shown in Figure 6. The results show that:

- The forgetting events of clean training samples follow an exponential distribution, and are similar among different cases.
- For poisoned training samples: **1)** when the poisoning ratio is low ($e.g.$, 0.1%, 0.5%), the forgetting numbers of poisoned samples are often larger than those of clean samples; **2)** when the poisoning ratio is high ($e.g.$, 5%, 10%), the forgetting numbers of poisoned samples are often smaller than those of clean samples.

The above observations are compatible with our high-level observation that the backdoor attack with higher poisoning ratios could quickly learn the stable mapping from the poisoned samples to the target class. Moreover, the forgetting event provides a fine-grained tool to analyze the contribution of each individual training sample, which could facilitate the development of more advanced backdoor attack and defense methods.

B.8 Analysis of trigger generalization of backdoor attacks

In all existing backdoor attacks, there is a default assumption that the triggers used in both backdoor training and backdoor testing are exactly same. However, we find that an interesting property in backdoor learning that the backdoored model trained with one trigger could be also activated by other triggers. We name it as **trigger generalization**.

In this following, we take the Blended attack as an example to study the trigger generalization. Specifically, we set the trigger transparency to different values in training and testing phase, including 10%, 20%, and 30%. As shown in the figure 7, we obtain the following observations: (1) If a more obvious trigger ($i.e.$, high transparency 30%) is applied during training, then the backdoor will not be easily activated by a different trigger with lower transparency; (2) If using a less obvious backdoor triggers ($i.e.$, low transparency 10%) in the training phase, then the backdoor can be successfully activated by the trigger with higher transparency ($e.g.$, 20% or 30%) in the test phase.

In addition to the above example, we find that the trigger generalization is a common property of the backdoor models under several backdoor attacks. For example, the trigger in SSBA [10] is a string, and its backdoored model could not only be activated by its training trigger. However, we find that the model is likely to be activated by many other strings with the same length. Exploring the behind reason of trigger generalization is important for us to better understand the backdoor mechanism. Moreover, we notice that there have been some attempts to utilize the backdoor as the technique to protect the intellectual property (IP) of AI models or datasets, based on the unique mapping from the trigger to the target class. However, due to the trigger generalization, the uniqueness of the training trigger no longer exists, which undermines the legitimacy backdoor learning in IP protection. Thus, the study of trigger generation is also important for the usage of backdoor learning in practice. We will provide more analysis about trigger generalization in BackdoorBench in future.

B.9 Evaluation on vision transformer

Until now, the reported 8,000 pairs of evaluations in BackdoorBench are all conducted on the convolutional neural networks. In the following, we expand the evaluations to another popular family of models, $i.e.$, vision transformer (ViT), which has shown superior performance on many vision tasks ($i.e.$, image classification, object detection, semantic segmentation). In our evaluations, a initial checkpoint of ViT that is pre-trained on ImageNet [4] is downloaded from `https://github.com/pytorch/vision/tree/main/references/classification`. We then fine-tune this pre-trained checkpoint on the poisoned CIFAR-10 dataset. Note that the input size of ViT is 224×224, while the size of raw images in CIFAR-10 is 32×32. Thus, in data poisoning based attack, we firstly insert the trigger into the raw image, then re-scale the poisoned image to the size 224×224.

We evaluate ViT-Base model with 16×16 input patch size (ViT-b-16) on CIFAR-10 with 10% poison ratio, where the settings of all hyper-parameters are same with those for learning other models, as shown in Tables 1 and 2. The backdoor evaluation results are summarized in Table 7. As a baseline, the accuracy of fine-tuning ViT-b-16 on the clean dataset with same hyper-parameters is 96.56%. According to Table 7 and the comparison with the evaluations on other models, we have the following

observations. **1)** In the case of no defense, the ASR of ViT-b-16 is still very high under all evaluated backdoor attacks, revealing that the ViT model architecture is also vulnerable to backdoor attacks. **2)** The evaluated three defense methods show very poor performance for the attack on ViT-b-16. In terms of FT and NC, although the ASR is reduced significantly, the clean accuracy is also downgraded. In terms of ABL, the ASR doesn't decrease for most attacks, with the only exception for BadNets of which the model after defense is fully degenerated. It implies that the effective defense that have been verified on the CNN architecture may not suitable for the ViT architecture. It inspires us to develop more effective defense methods for the ViT architecture specially. More evaluations and analysis about the ViT architecture will be added in BackdoorBench in future.

B.10 Evaluation on ImageNet

Due to the high computational and memory costs, one of the benchmark datasets of image classification, $i.e.$, ImageNet [4] with 1,000 classes, has rarely been evaluated in existing backdoor learning works. We plan to provide comprehensive evaluations of backdoor learning methods on ImageNet, to find whether there are some unique challenges for backdoor learning on large-scale datasets.

Here, we provide some partial evaluations, including BadNets and Blended attack with 0.1% poison ratio on ImageNet and the PreAct-ResNet18 model, as shown in Table 8. Note that due to the 1,000 classes, we do not set a higher poisoning ratio to ensure that the number of poisoned samples is not much larger than the number of clean samples of the target class. Both Badnets and Blended show good attack performance with high ASR and C-Acc. Compared with the baseline model, $i.e.$, PreAct-ResNet18 trained on the clean ImageNet dataset (please refer to `https://paperswithcode.com/sota/image-classification-on-imagenet?tag_filter=3`), there is a slight drop of C-Acc, from 72.33% to 69.22%. More backdoor attack and defense evaluations on ImageNet will be added to our BackdoorBench in the future.

B.11 Visualization

B.11.1 Individual visualization tools

Here we provide three visualization tools to analyze each individual image.

Gradient-weighted class activation mapping (Grad-CAM) [19] explains the contribution of each pixel to the prediction of one image, based on the gradient of the logit of one class *w.r.t.* each pixel. Note that in the codebase of BackdoorBench, we implement a variant of Grad-CAM, called FullGrad [22].

Shapley Value [13] is another popular interpretation tool that assigns an importance factor to each pixel for a particular prediction. Inspired by the cooperative game theory, the competition among pixels is also taken into account in the computation of each individual importance factor.

Frequency saliency map (FSM) is our innovative visualization method for viewing the contribution of every Fourier basis to model classification. Consider an image classification task with S classes. Let $\boldsymbol{x}$ be a clean image with size $H \times W \times C$ and $\tilde{\boldsymbol{x}} = \mathfrak{F}(\boldsymbol{x})$ be the corresponding frequency spectrum with $\mathfrak{F}$ being the channel-wise Discrete Fourier Transform (DFT) operator. Let $\mathbb{C}$ be the set of complex numbers. Denote the classifier by $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^S$. We define $F : \mathbb{C}^{H\times W\times C} \to \mathbb{R}^S$ as the corresponding classifier in the frequency domain, which means

$$F(\tilde{\boldsymbol{x}}) = f(\boldsymbol{x}) = f\big(\mathfrak{F}^{-1}(\tilde{\boldsymbol{x}})\big), \tag{1}$$

where $\mathfrak{F}^{-1}$ is the channel-wise Inverse Discrete Fourier Transform (IDFT) operator, which means $\mathfrak{F}^{-1}(\tilde{\boldsymbol{x}})(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\tilde{\boldsymbol{x}}(h,w)e^{2\pi i(\frac{uh}{H}+\frac{vw}{W})}$.

Inspired by the saliency map in the RGB space, we intend to establish the connection between model prediction and image's frequency spectrum by the gradient. According to the chain's rule, we can estimate the gradient *w.r.t.* the frequency spectrum as follows

$$\begin{aligned}\frac{\partial F_s(\tilde{\boldsymbol{x}})}{\partial \tilde{\boldsymbol{x}}(u,v,c)} &= \sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sum_{c'=0}^{C-1}\frac{\partial f_s(\boldsymbol{x})}{\partial \boldsymbol{x}(h,w,c')}\cdot\frac{\partial \boldsymbol{x}(h,w,c')}{\partial \tilde{\boldsymbol{x}}(u,v,c)} \\ &= \sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\frac{\partial f_s(\boldsymbol{x})}{\partial \boldsymbol{x}(h,w,c)}e^{2\pi i(\frac{uh}{H}+\frac{vw}{W})},\end{aligned} \tag{2}$$

where f_s means the logit output of the model f *w.r.t.* the s-th class.

The obtained gradients are then averaged over channels and normalized to form the FSM, as follows

$$FSM^*(u,v) = \frac{1}{C}\sum_{c=0}^{C-1}\frac{\partial F_s(\tilde{\boldsymbol{x}})}{\partial \tilde{\boldsymbol{x}}(u,v,c)}, \tag{3}$$

$$FSM(u,v) = \frac{FSM^*(u,v) - FSM^*_{\min}}{FSM^*_{\max} - FSM^*_{\min}}, \tag{4}$$

where we use min-max normalization to normalize the FSM image.

B.11.2 Visualization results

In the following, we present some visualization results using the above three tools to understand the inner mechanism of backdoor learning better. Specifically, Specifically, we train the PreAct-ResNet18 model under various backdoor attacks and defenses with the poisoning ratio 5%, on 3 datasets, including CIFAR-100, GTSRB, and Tiny ImageNet. Then, we randomly select a poisoned sample from the test set and show its visualization. The visualization results using Shapley Value and Grad-CAM are shown in Figures 8 to 13, respectively.

The frequency saliency map (FSM) visualization results are shown in Figure 14, where low-frequency components are shifted into the central regions. In contrast, high-frequency components are distributed in surrounding regions. The first column displays the studied poisoned images generated by different attack methods, including BadNet, Blended, SSBA, WaNet, and LF. The second column shows the contribution of each frequency basis to the backdoored model's prediction. It tells that most backdoor models pay attention to high-frequency regions, while the model under the LF attack makes the model concentrates more on low-frequency regions. Besides, we can see an apparent difference between SSBA and WaNet in the frequency domain, even though their spatial images look similar. These observations demonstrate the potential usage of FSM in trigger or backdoor detection, which will be explored in our future work. The remaining columns show the contribution of each frequency basis under various defense methods. For example, the contribution of some high-frequency regions is enormous for the poisoned image in a BadNet model without defense. However, after conducting FT on this backdoor model, the low-frequency regions regain attention from the model. It explains well that FT can effectively remove backdoors embedded by BadNet and WaNet (see Figure 3 in the main manuscript). We plan to explore more backdoor learning properties from the frequency domain perspective.

C BackdoorBench in Natural Language Processing

Apart from the analysis of backdoor attack and defense methods in computer vision, we also expand our benchmark to the field of Natural Language Processing (see `https://github.com/SCLBD/BackdoorBench/tree/main/backdoorbench_nlp`). We implement two stage-of-the-art backdoor attack methods ($i.e.$, LWS [18] and HiddenKiller [17]) and one defense method ($i.e.$, Onion [16]) in NLP as a complement to the original BackdoorBench. We closely follow the original implementation of the attack and defense methods and make necessary changes to unify all the methods in our benchmark. We choose BERT [5] as the model to be poisoned. All the experiments are conducted on three widely-used datasets for text classification tasks, including Stanford Sentiment Treebank(SST-2) [21], Offensive Language Identification Dataset(OLID) [28] and AG's News [29]. For all experiments, the poison rate is set to be 5% and the default target label is 1. For LWS, the bar for ONION for each dataset is set to be the recommended value in the original implementation. All results are reported in Table 9.

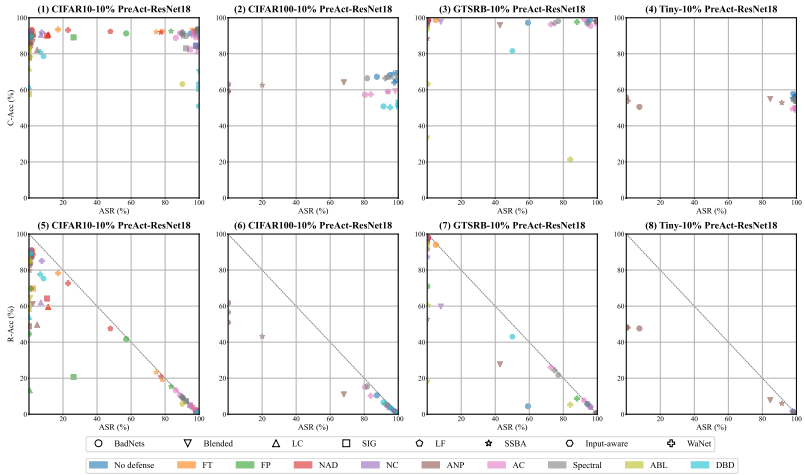
We can find that the two chosen attack methods can both achieve high attack success rate even at a low poisoning ratio. However, the defense performance of ONION against two SOTA attack methods is not quite satisfactory. The possible reason is that ONION aims to find out obvious outliers in each sentence, but both HiddenKiller and LWS are invisible methods which do not rely on special tokens as triggers. In the future, we will also keep updating latest backdoor attack and defense methods in the NLP field into our benchmark.

D Reproducibility

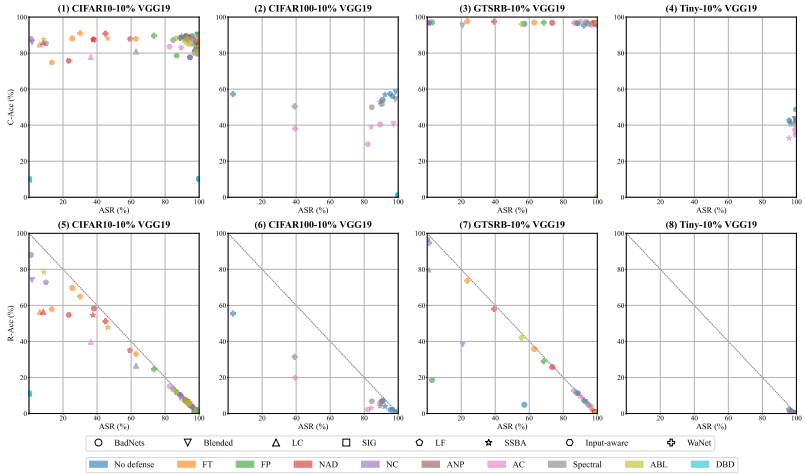
All evaluation results in BackdoorBench can be easily reproducible, just running the scripts provided in the github repository `https://github.com/SCLBD/BackdoorBench`, with the hyper-parameter settings presented in Tables 1 and 2. All evaluated datasets and model architectures are publicly and freely available. Besides, we also compress all codes into one file as a part of the supplementary materials.

E License

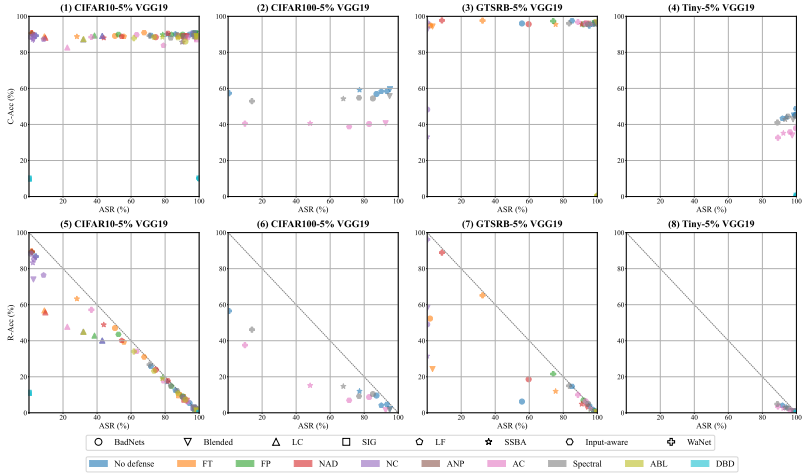
This repository is licensed by The Chinese University of Hong Kong, Shenzhen and Shenzhen Research Institute of Big Data under Creative Commons Attribution-NonCommercial 4.0 International Public License (identified as CC BY-NC-4.0 in SPDX, see `https://spdx.org/licenses/`). More details about the license could be found in `https://github.com/SCLBD/BackdoorBench/blob/main/LICENSE`.



(a) Attack-defense pairs with PreAct-ResNet18 and 10% poisoning ratio on CIFAR-10.



(b) Attack-defense pairs with VGG-19 and 10% poisoning ratio on CIFAR-10.



(c) Attack-defense pairs with VGG-19 and 5% poisoning ratio on CIFAR-10.

Figure 1: Performance distribution of attack-defense pairs on different model structure and poisoning ratios. A successful attack method should be high C-Acc and ASR, while a successful defense method should be high C-Acc and low ASR. Besides, if the reduced ASR value equals to the increased R-Acc, the color patters would be close to the anti-diagonal line.

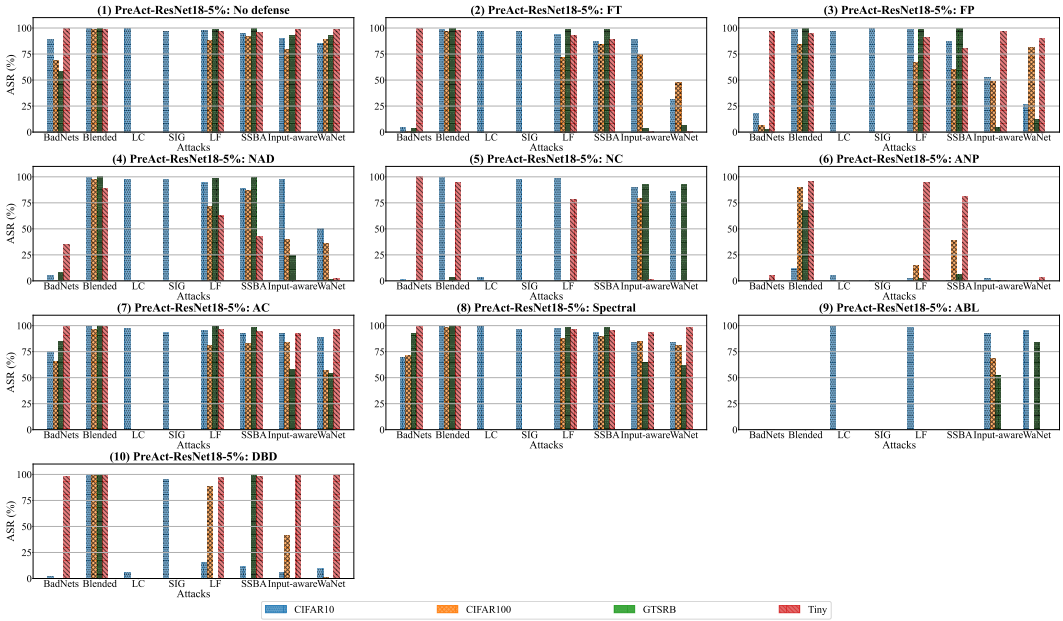


Figure 2: The effects of different datasets on backdoor learning. Note that for the clean-label ($i.e.$, LC [20] and SIG [1]) attack, the number of poisoned samples must be less than the target class size, thus it may be not applied to the case of high poisoning ratios.)

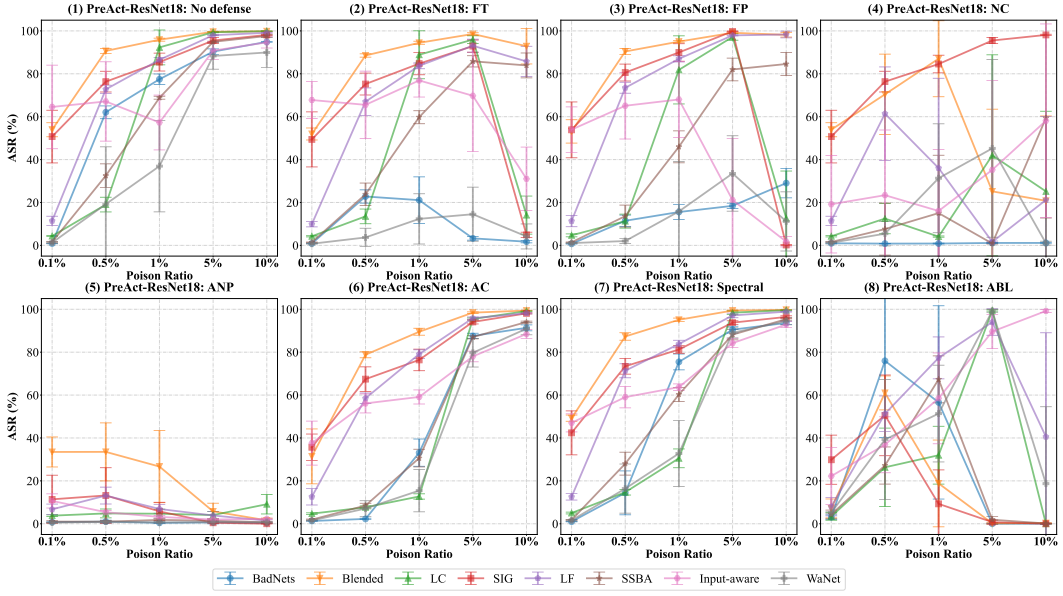
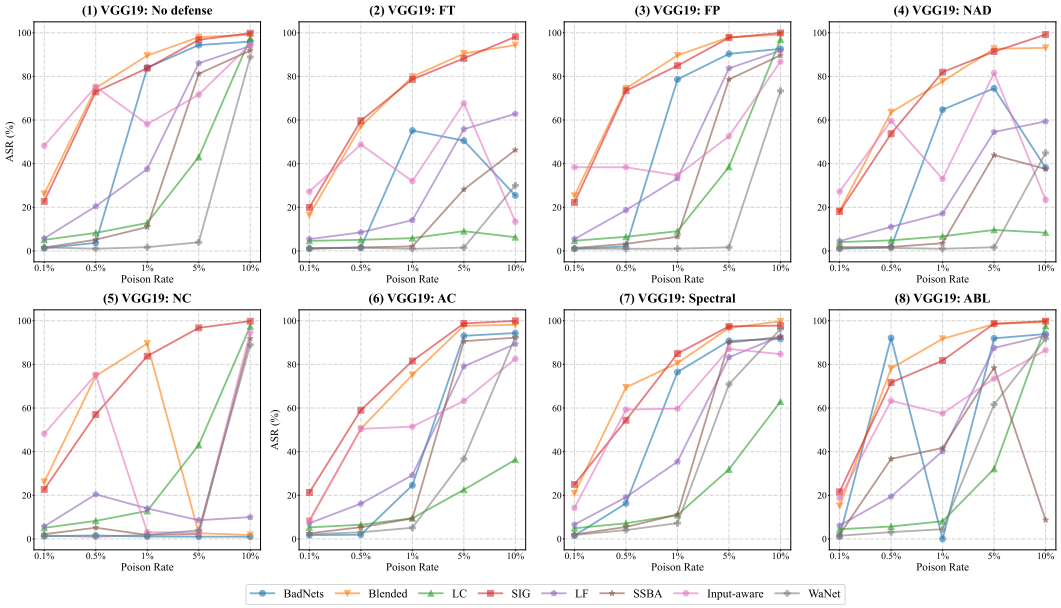
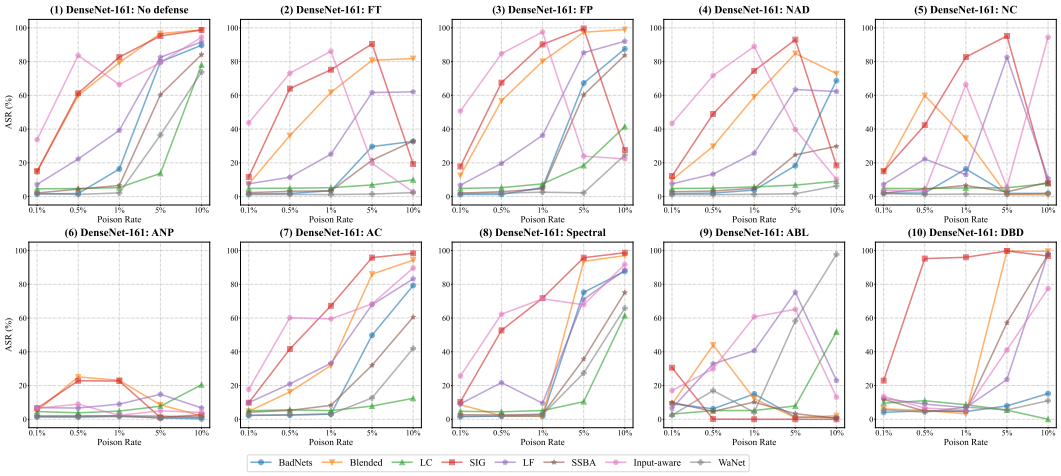


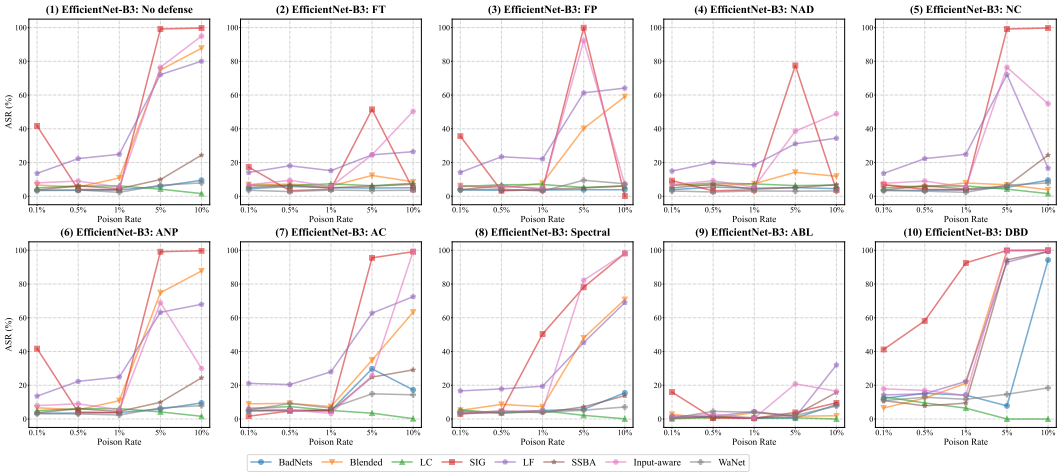
Figure 3: The effects of different poisoning ratios of backdoor learning with 5 random seeds.



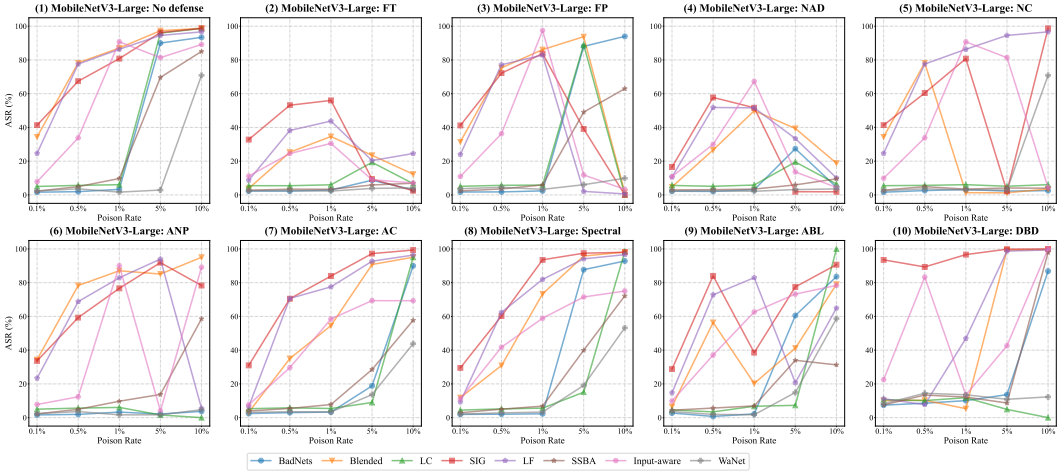
(a) The variation of ASR on different poisoning ratios with VGG-19 and CIFAR-10.



(b) The variation of ASR on different poisoning ratios with DenseNet-161 and CIFAR-10.



(c) The variation of ASR on different poisoning ratios with EfficientNet-B3 and CIFAR-10.



(d) The variation of ASR on different poisoning ratios with MobileNetV3-Large and CIFAR-10.

Figure 4: The effect of different ratios on backdoor learning. From (a) to (d), the structure of models are different. In the condition of no defense, the higher poisoning ratio, the higher ASR value. In the defense situation, some ASR curves raise with the increase of poisoning ratio, while some curves go up first and then sharply drop down. It could also be noticed that the performances of same defense method on different model structures are distinctive, $i.e.$, ABL on VGG-19 and DenseNet-161. Note that we don't provide the results of ANP and DBD on VGG-19, and the reason is illustrated in Section B.4.2.

Table 3: Sensitivity results on CIFAR10 with PreAct-ResNet18.

		No Defense	No Defense	No Defense	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP
		\	\	\	0.1	0.1	0.1	0.5	0.5	0.5	0.9	0.9	0.9	1.3	1.3	1.3	1.7	1.7	1.7	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.5	0.5	0.5	0.6	0.6	0.6
		C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
BadNets	1	88.81	64.61	33.90	83.50	0.01	90.12	82.09	0.12	87.72	83.66	0.03	89.82	10.00	100.00	0.00	8.51	99.74	0.07	76.44	38.08	58.50	82.39	9.14	78.18	81.49	2.17	80.29	69.58	1.20	70.16	57.70	2.09	58.03
BadNets	2	90.20	90.23	9.32	83.85	0.00	90.82	83.66	0.01	90.09	86.10	0.03	90.72	83.23	0.00	89.42	10.00	100.00	0.00	82.74	17.92	72.26	83.50	4.41	80.21	80.89	0.43	80.99	78.04	0.29	81.40	67.91	0.04	74.24
BadNets	3	91.32	95.03	4.67	85.45	0.01	90.43	82.96	0.00	88.78	82.69	0.00	90.06	84.10	0.01	89.82	84.20	0.01	90.58	86.93	8.09	83.27	87.46	2.87	85.68	85.37	1.20	84.69	80.17	0.43	81.80	70.27	0.07	74.50
BadNets	4	92.12	96.94	2.84	84.96	0.00	90.73	84.89	0.00	90.67	83.93	0.01	90.04	82.58	0.00	90.37	84.01	0.00	90.80	88.45	4.97	86.09	88.05	1.17	88.08	83.35	0.18	86.18	75.67	0.11	79.26	69.23	0.09	72.72
BadNets	5	93.24	97.87	2.03	80.35	0.03	87.86	79.11	0.00	86.49	81.73	0.00	89.67	81.73	0.00	90.30	80.71	0.00	89.62	90.47	6.29	83.17	88.12	1.34	84.39	83.92	0.21	82.67	82.23	0.21	82.79	84.23	0.17	86.81
SIG	1	84.75	48.31	49.69	67.84	0.00	69.82	56.99	0.00	58.98	62.08	0.00	68.39	47.52	0.00	49.50	52.05	0.00	57.00	84.97	10.97	80.12	86.49	11.62	80.44	84.60	5.80	85.11	83.24	3.53	86.01	80.79	1.41	84.93
SIG	2	84.65	83.22	16.41	74.41	0.00	65.97	68.86	0.00	69.53	67.62	0.00	72.92	71.82	0.00	72.42	64.90	0.00	67.59	84.05	19.07	65.84	84.94	17.30	70.73	83.23	13.64	70.00	80.06	6.16	74.82	76.59	4.63	70.16
SIG	3	84.55	90.11	9.73	77.02	0.00	61.33	69.57	0.00	67.66	74.04	0.00	74.47	72.98	0.00	71.80	67.89	0.00	67.14	84.70	25.48	59.76	83.10	12.87	62.52	81.74	8.00	62.59	80.23	4.02	62.81	78.49	1.12	71.80
SIG	4	84.47	97.16	2.82	74.14	0.00	51.79	57.33	0.00	62.82	68.04	0.00	63.59	65.23	0.00	60.92	53.67	0.00	64.31	83.81	9.59	65.92	82.31	6.98	67.88	82.87	7.59	66.82	81.78	3.21	65.49	80.45	1.01	64.67
SIG	5	84.52	97.18	2.80	73.70	0.00	65.23	56.90	0.00	61.12	59.87	0.00	55.80	41.41	0.00	49.43	54.33	0.00	52.24	82.37	1.73	59.69	81.77	1.48	56.46	81.36	1.98	58.32	80.00	1.99	62.07	77.82	0.82	57.62
Input-aware	0.032	90.94	89.22	10.11	61.30	95.94	3.12	78.60	93.34	5.36	48.16	100.00	0.00	47.99	98.61	0.97	49.40	25.06	34.81	90.59	2.40	86.34	90.27	3.93	84.11	89.64	5.17	80.08	88.43	7.22	77.42	85.79	8.61	75.84
Input-aware	0.048	90.98	91.81	7.70	80.78	90.71	7.56	34.29	99.33	0.28	16.89	99.96	0.04	42.13	94.16	3.23	52.34	11.12	47.86	91.33	3.03	84.57	91.13	2.78	84.76	90.72	3.77	84.42	89.25	5.62	78.68	86.97	7.07	74.92
Input-aware	0.064	89.94	92.30	7.29	79.11	98.81	1.01	67.28	99.84	0.10	39.14	99.98	0.01	55.76	97.41	1.73	60.86	15.43	55.21	90.26	1.48	86.79	89.83	1.44	85.66	89.86	1.87	86.02	87.99	3.97	83.12	82.33	4.22	77.69
Input-aware	0.08	90.69	95.97	3.91	74.14	5.90	62.78	64.92	98.67	0.97	33.29	99.97	0.01	31.56	99.38	0.29	80.89	0.00	84.03	90.26	1.00	87.07	89.64	1.51	85.48	90.34	1.98	85.22	88.98	2.68	82.12	87.66	4.08	78.22
Input-aware	0.096	90.61	95.58	4.17	87.96	0.27	82.32	34.55	99.83	0.02	10.19	100.00	0.00	15.27	99.88	0.00	82.02	0.00	87.09	89.95	1.56	86.84	90.20	1.28	86.92	89.96	2.46	85.90	89.29	2.84	85.02	87.62	2.29	83.34

Table 4: Sensitivity results on CIFAR10 with VGG-19.

		No Defense	No Defense	No Defense	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL
		\	\	\	0.1	0.1	0.1	0.5	0.5	0.5	0.9	0.9	0.9	1.3	1.3	1.3	1.7	1.7	1.7
		C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
BadNets	1	85.79	76.44	22.46	84.14	70.67	26.70	71.31	45.73	43.34	53.99	45.03	36.38	10.00	100.00	0.00	10.00	100.00	0.00
BadNets	2	87.26	93.76	5.80	84.55	73.70	23.83	85.18	89.32	9.62	83.58	88.21	10.67	78.82	85.81	12.76	10.00	100.00	0.00
BadNets	3	88.75	95.77	3.93	85.93	91.29	7.87	84.67	90.51	8.56	84.31	93.03	6.28	83.75	92.74	6.54	79.32	91.17	7.73
BadNets	4	89.50	96.73	3.08	86.41	93.91	5.52	86.77	92.56	6.80	87.24	94.44	5.02	85.75	95.06	4.50	85.69	94.64	4.90
BadNets	5	89.87	97.03	2.78	88.04	96.18	3.53	88.67	96.74	3.02	87.79	96.98	2.76	88.22	97.04	2.73	88.28	97.50	2.27
SIG	1	82.20	41.61	53.47	80.54	41.06	51.78	80.55	44.52	49.36	80.68	44.10	50.00	80.33	44.80	49.26	80.14	45.06	47.74
SIG	2	81.92	74.87	21.84	80.16	72.33	23.38	80.16	76.11	20.68	80.39	78.21	19.44	80.02	82.79	14.37	80.05	83.04	14.23
SIG	3	81.44	88.27	8.38	80.25	87.62	10.23	80.79	89.33	9.21	80.14	87.14	10.50	80.46	85.22	11.08	80.40	90.36	7.99
SIG	4	81.44	98.03	1.42	80.73	98.43	1.32	80.51	98.39	1.40	80.45	99.27	0.61	80.55	98.76	1.01	78.89	94.52	4.13
SIG	5	82.16	99.50	0.30	80.04	99.34	0.42	80.89	99.19	0.63	80.17	99.17	0.72	80.58	98.56	1.08	80.01	98.41	1.24
Input-aware	0.032	88.48	96.77	2.89	88.27	90.49	8.19	88.18	92.08	6.90	24.65	100.00	0.00	86.56	7.87	73.64	10.00	100.00	0.00
Input-aware	0.048	88.86	94.98	4.44	88.31	81.49	16.21	88.43	79.66	17.93	88.78	72.57	24.50	10.00	100.00	0.00	10.00	100.00	0.00
Input-aware	0.064	88.71	85.72	12.91	88.58	84.52	13.64	76.06	98.67	1.03	10.00	100.00	0.00	50.51	0.00	37.17	10.00	100.00	0.00
Input-aware	0.08	88.54	97.00	2.61	87.81	84.16	13.40	88.22	86.89	11.50	88.04	88.20	10.16	87.66	87.62	10.60	10.00	100.00	0.00
Input-aware	0.096	89.11	96.52	3.36	88.01	81.91	15.76	88.74	83.54	14.17	88.38	84.06	14.08	87.39	6.69	73.10	10.00	100.00	0.00

Table 5: Sensitivity results on GTSRB with PreAct-ResNet18.

		No Defense	No Defense	No Defense	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP	ANP
		\	\	\	0.1	0.1	0.1	0.5	0.5	0.5	0.9	0.9	0.9	1.3	1.3	1.3	1.7	1.7	1.7	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.5	0.5	0.5	0.6	0.6	0.6
		C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	ACC	ASR	RA
BadNets	1	95.97	92.25	7.48	90.19	0.00	91.02	85.45	0.01	85.45	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00	95.00	9.75	89.14	94.19	3.51	93.16	93.38	3.70	92.39	91.23	2.35	90.80	88.51	0.70	88.70
BadNets	2	96.83	93.99	5.98	93.60	0.00	94.15	93.21	0.00	93.72	93.93	0.00	94.53	91.68	0.00	92.32	91.65	0.00	91.93	95.10	0.07	95.19	96.22	0.00	96.23	96.09	0.01	96.17	93.28	0.00	93.37	89.28	0.00	89.69
BadNets	3	97.56	94.84	5.10	82.26	0.00	82.29	93.82	0.00	94.21	93.49	0.00	93.77	94.37	0.00	94.91	71.26	0.00	71.92	97.59	16.49	82.91	96.66	0.12	96.46	96.68	0.00	96.79	94.93	0.00	95.06	93.35	0.00	93.75
BadNets	4	97.31	96.83	3.07	94.54	0.00	94.98	95.20	0.00	95.67	95.01	0.00	95.45	93.97	0.00	94.44	93.92	0.00	94.29	97.75	0.00	97.82	97.50	0.00	97.64	96.10	0.00	96.32	93.74	0.00	94.02	91.93	0.00	92.31
BadNets	5	97.85	97.49	2.46	80.69	0.00	81.11	93.60	0.00	94.11	92.45	0.00	92.90	89.10	0.00	89.63	82.58	0.00	83.09	98.00	0.00	98.04	97.98	0.00	98.01	97.61	0.00	97.57	97.00	0.00	97.02	95.39	0.00	95.38
Input-aware	0.032	97.23	97.36	2.52	41.34	85.39	7.55	29.45	68.56	13.83	7.90	74.31	0.56	6.73	83.58	2.13	3.79	90.03	0.06	96.88	0.01	95.25	95.62	0.00	93.80	94.39	0.00	91.91	91.84	0.00	88.30	91.46	0.00	88.44
Input-aware	0.048	97.48	96.67	3.31	64.47	50.61	27.53	12.82	94.71	0.51	5.48	80.84	1.35	7.05	96.89	0.31	18.07	76.04	4.65	97.27	0.00	96.91	96.70	0.00	95.93	96.72	0.00	96.09	96.49	0.00	95.69	96.34	0.02	95.12
Input-aware	0.064	97.66	88.88	10.54	52.05	66.21	16.32	9.57	79.77	2.32	4.50	62.74	1.99	27.28	48.58	13.97	3.63	91.19	1.19	96.63	0.31	94.58	96.44	0.28	94.32	95.63	0.22	92.90	95.65	0.20	93.02	94.21	0.45	91.69
Input-aware	0.08	97.77	96.65	3.34	15.19	70.62	5.27	12.72	41.96	6.54	19.73	80.36	3.37	25.81	55.12	11.89	5.82	92.60	1.89	97.14	0.00	95.02	96.23	0.00	93.87	95.85	0.00	93.44	95.40	0.04	91.86	94.75	0.02	90.53
Input-aware	0.096	98.26	97.24	2.74	23.83	67.91	7.46	25.49	75.73	6.00	1.63	96.47	0.25	0.86	99.55	0.00	1.69	98.63	0.00	97.69	0.64	96.25	97.54	0.48	96.04	97.03	0.16	95.53	96.59	0.32	94.71	95.12	0.29	92.84

Table 6: Sensitivity results on GTSRB with VGG-19.

		No Defense	No Defense	No Defense	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL	ABL
		\	\	\	0.1	0.1	0.1	0.5	0.5	0.5	0.9	0.9	0.9	1.3	1.3	1.3	1.7	1.7	1.7
		ACC	ASR	RA	ACC	ASR	RA	ACC	ASR	RA	ACC	ASR	RA	ACC	ASR	RA	ACC	ASR	RA
BadNets	1	96.17	89.01	9.84	96.13	91.85	7.81	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00
BadNets	2	95.20	93.50	5.91	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00	57.59	96.01	2.79
BadNets	3	96.14	94.52	5.18	0.48	100.00	0.00	70.24	42.78	41.52	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00
BadNets	4	94.35	96.38	3.31	96.26	95.55	4.27	96.26	93.09	6.57	95.56	93.35	6.38	92.91	93.20	6.26	90.07	91.03	8.23
BadNets	5	95.77	96.40	3.31	96.63	96.36	3.37	96.63	96.86	2.80	96.31	96.25	3.45	95.93	96.13	3.64	96.38	96.12	3.75
Input-aware	0.032	97.51	88.97	10.61	96.83	53.95	43.74	96.63	58.14	40.04	95.94	59.78	38.02	0.48	100.00	0.00	0.48	100.00	0.00
Input-aware	0.048	97.44	60.09	39.36	96.35	61.02	37.30	8.99	99.83	0.17	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00
Input-aware	0.064	96.55	89.63	9.99	96.67	54.98	43.28	96.71	51.89	45.87	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00
Input-aware	0.08	96.89	86.45	13.43	96.46	51.20	46.48	96.68	44.12	53.95	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00
Input-aware	0.096	96.25	67.77	31.54	96.16	49.47	47.90	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00	0.48	100.00	0.00

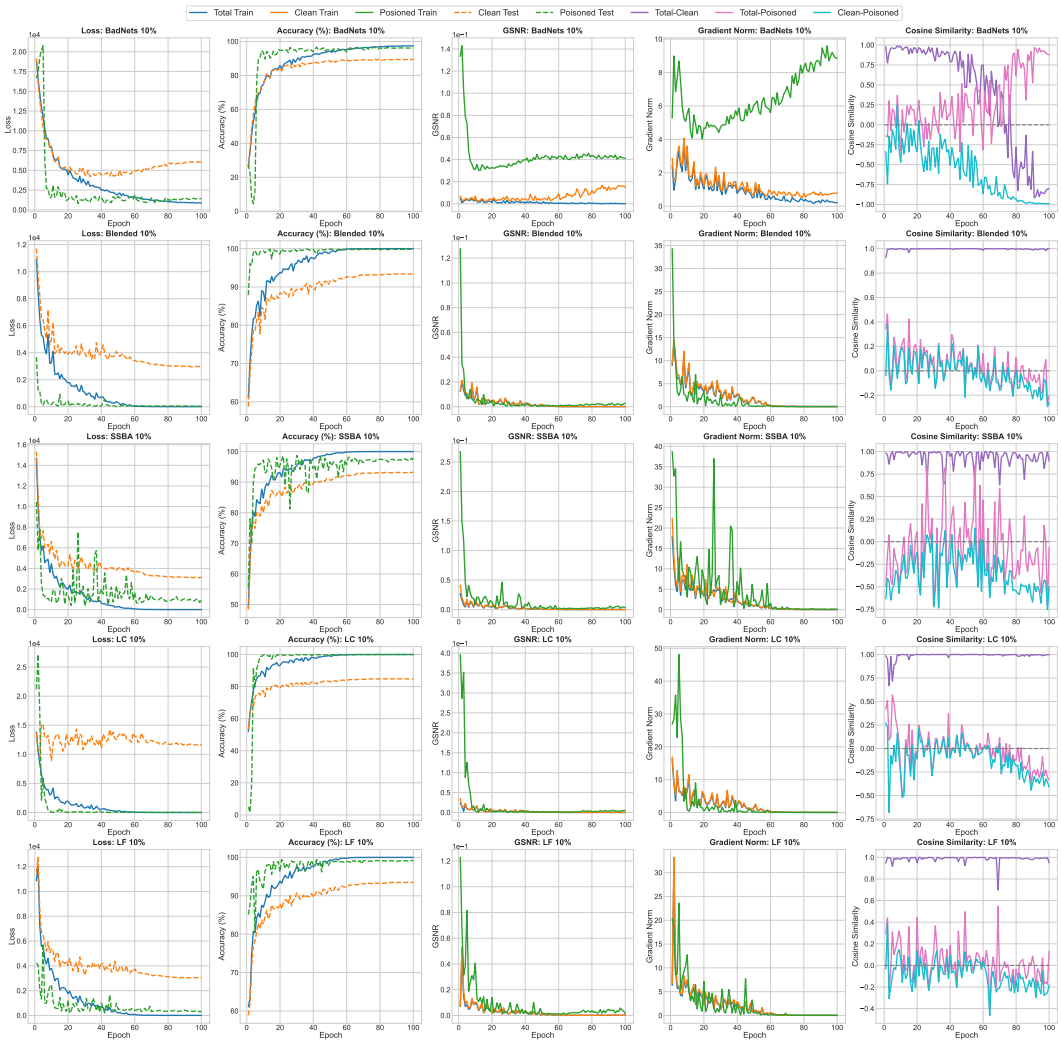


Figure 5: Analysis of the quick learning phenomenon based on the details of the backdoor learning process. From the first column to the last column, we report the curves of loss, accuracy, GSNR, gradient norm and cosine similarity of gradients in poisoned samples, clean samples and all samples, respectively.

Table 7: Evaluation of ViT-b-16 on CIFAR-10 with 10% poisoning ratio.

	No defense			FT			NC			ABL		
Backdoor Attack	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
BadNets	94.58	94.11	5.66	42.00	8.81	38.60	36.09	11.12	33.60	10.00	0.00	11.11
Blended	96.47	99.72	0.28	41.36	7.24	36.29	44.92	3.90	39.40	96.88	99.81	0.19
LC	86.87	99.84	0.16	33.03	16.40	24.36	40.68	11.17	30.74	87.33	99.76	0.23
SIG	87.01	92.60	7.28	44.46	0.78	25.92	87.01	92.60	7.28	87.60	86.47	13.39
SSBA	96.30	97.58	2.34	45.35	7.62	43.49	46.24	8.17	43.60	96.78	98.22	1.70
Input-aware	92.10	96.28	3.51	89.56	25.14	69.91	43.13	7.79	37.81	96.70	93.21	6.58

Table 8: BadNets and Blended attack result on ImageNet

	C-Acc	ASR	R-Acc
BadNets	69.21	75.86	0.33
Blended	69.24	98.59	0.11

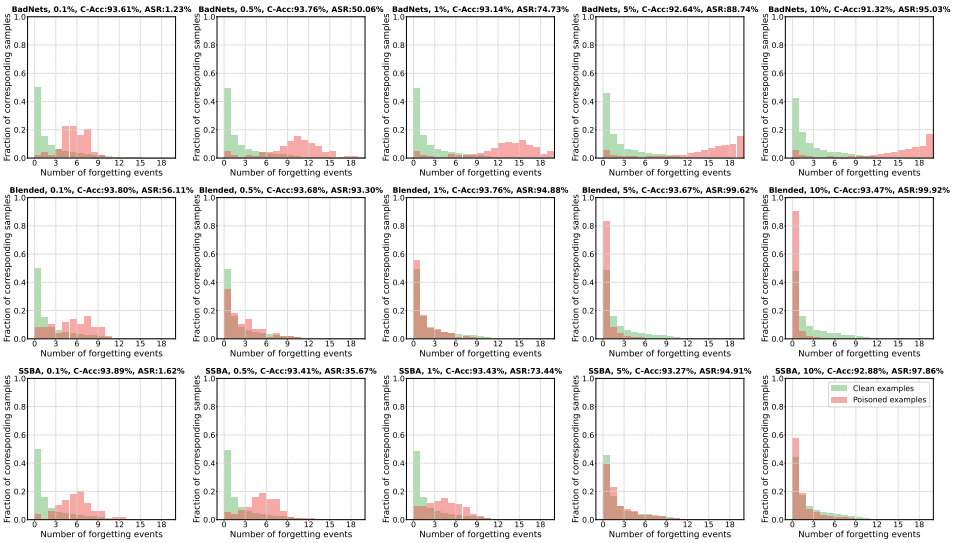


Figure 6: Distributions of forgetting events of clean training examples and poisoned examples in CIFAR-10 with Preact-Resnet18 backbone.

Figure 7: Analysis of trigger generalization in Blended attack: a) the training trigger with 10% transparency; b) the training trigger with 20% transparency; c) the training trigger with 30% transparency. For each case, we evaluate the attack success rate of testing triggers with the transparency 10%, 20%, and 30%, respectively.

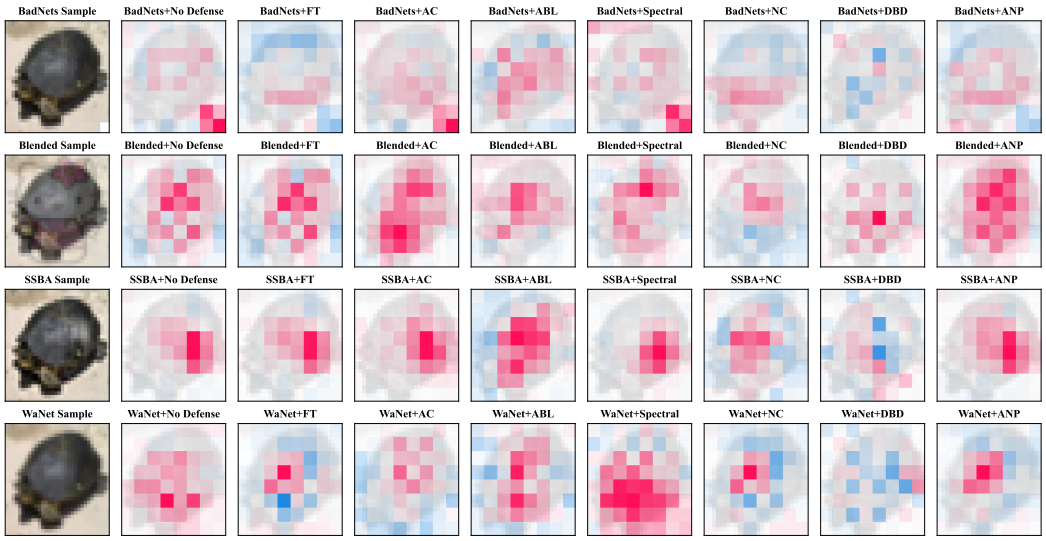


Figure 8: Shapley Value visualization of regions contributed to model decision under different attack methods and defense methods with PreAct-ResNet18 and 5% poisoning rate on CIFAR-100.

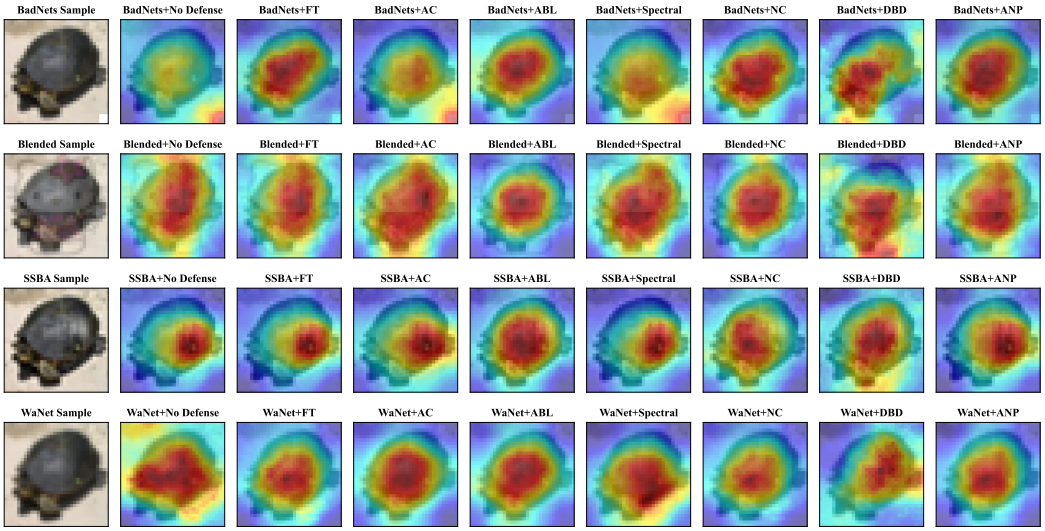


Figure 9: Grad-CAM visualization of regions contributed to model decision under different attack methods and defense methods with PreAct-ResNet18 and 5% poisoning rate on CIFAR-100.

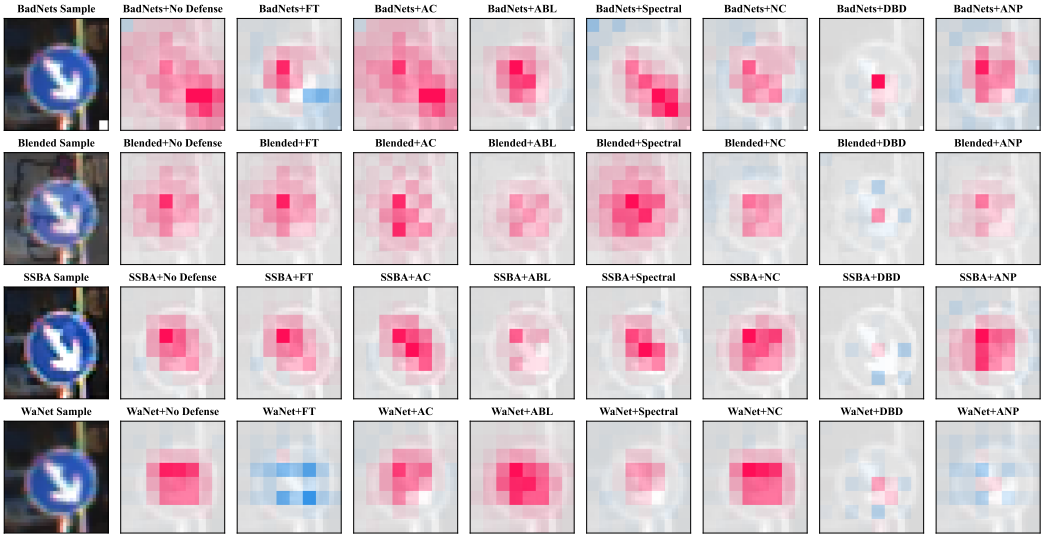


Figure 10: Shapley Value visualization of regions contributed to model decision under different attack methods and defense methods with PreAct-ResNet18 and 5% poisoning rate on GTSRB.

Table 9: Full results on backdoor attack and defense methods for BERT

		No defense			Onion		
Backdoor Attack	Dataset	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
LWS	SST-2	89.01	94.08	4.28	86.20	90.42	9.58
LWS	OLID	82.67	97.92	0.83	79.07	96.77	3.23
LWS	AgNews	93.11	99.19	0.61	92.10	68.03	10.97
HiddenKiller	SST-2	90.34	88.93	11.07	85.67	88.27	11.73
HiddenKiller	OLID	82.19	97.42	2.59	81.37	96.12	3.88
HiddenKiller	AgNews	93.49	98.67	1.12	92.05	95.16	4.21

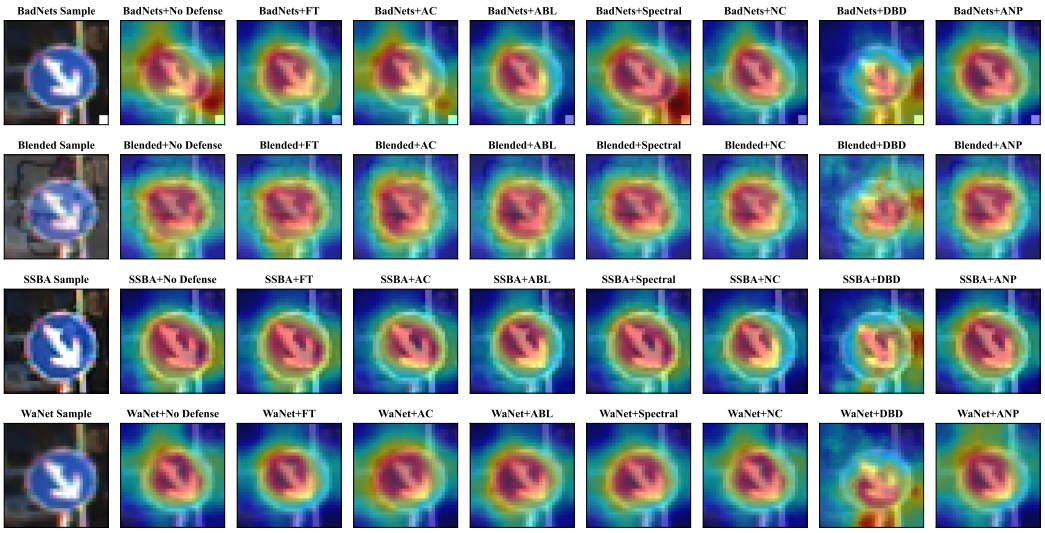


Figure 11: Grad-CAM visualization of regions contributed to model decision under different attack methods and defense methods with PreAct-ResNet18 and 5% poisoning rate on GTSRB.

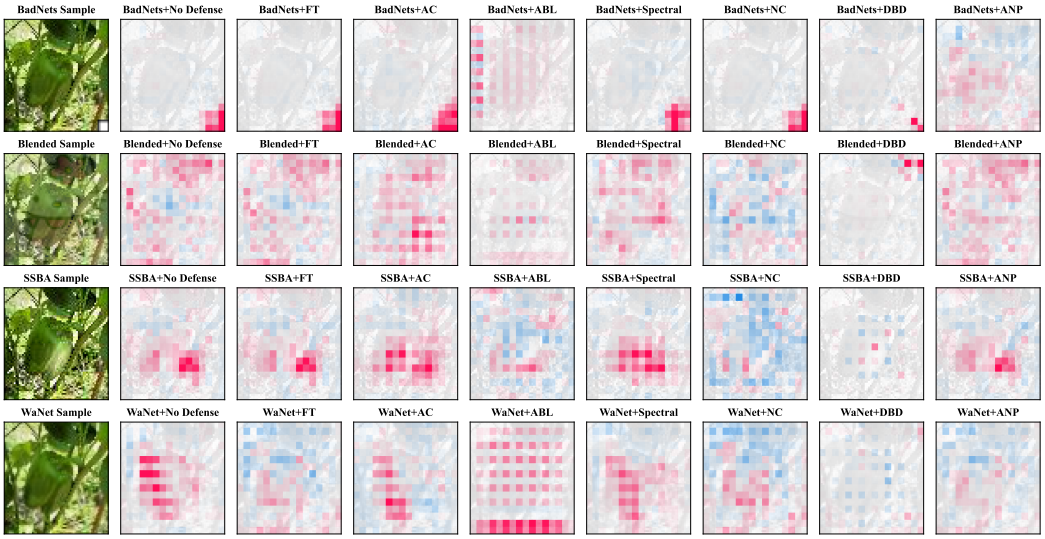


Figure 12: Shapley Value visualization of regions contributed to model decision under different attack methods and defense methods with PreAct-ResNet18 and 5% poisoning rate on Tiny ImageNet.

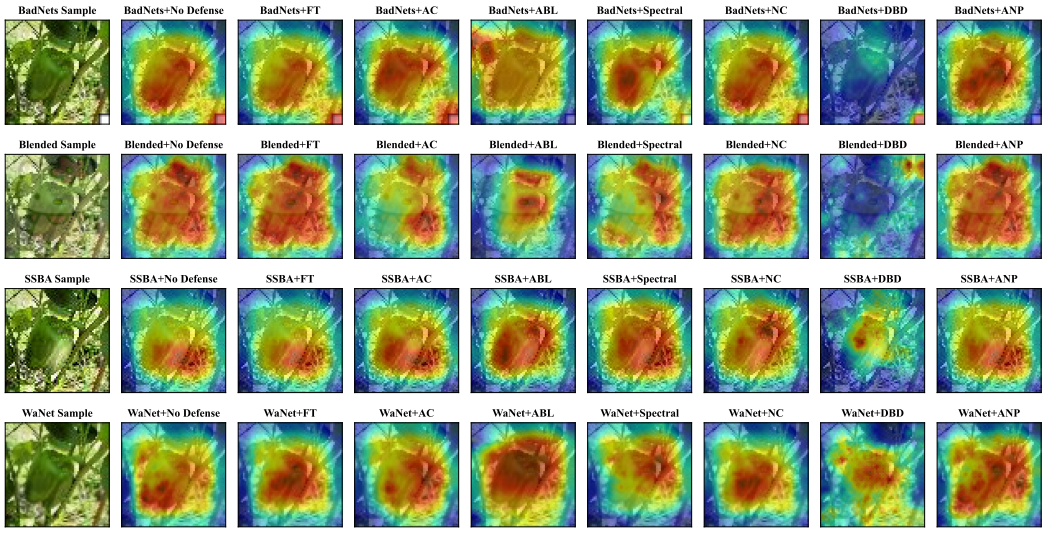


Figure 13: Grad-CAM visualization of regions contributed to model decision under different attack methods and defense methods with PreAct-ResNet18 and 5% poisoning rate on Tiny ImageNet.

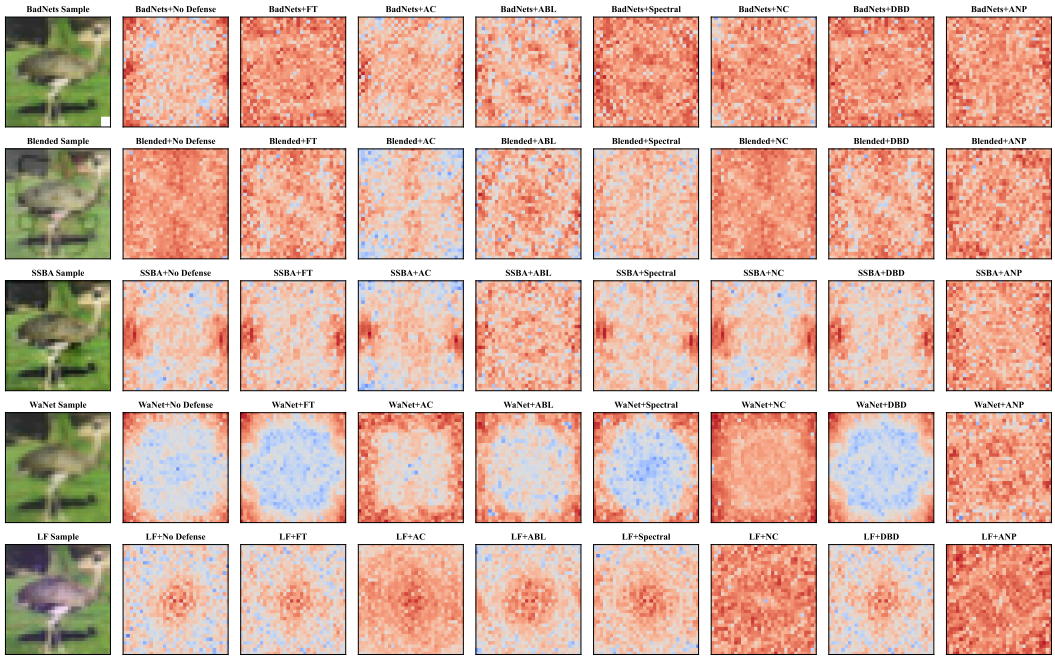


Figure 14: Frequency visualization of regions which contribute to model classification under different attack methods and defense methods with PreAct-ResNet18 and 10% poisoning rate on CIFAR-10. Warmer color was allocated to higher value whose range is 0 to 255.

Table 10: Full results on CIFAR-10 with 10% poisoning ratio.

	Backdoor Defense	No defense			FT			FP			NAD			NC			ANP			AC			Spectral			ABL			DBD		
Targeted Model	Backdoor Attack	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
PreAct-Resnet18	BadNets	91.32	95.03	4.67	89.96	1.48	89.39	91.31	57.13	41.62	89.87	2.14	88.71	89.05	1.27	89.16	84.17	0.73	84.04	88.80	86.23	13.28	89.98	92.41	7.20	83.32	0.00	89.02	89.65	1.28	89.17
PreAct-Resnet18	Blended	93.47	99.92	0.08	92.78	96.11	3.57	93.17	99.26	0.73	92.17	97.69	2.14	93.47	99.92	0.08	86.58	2.19	61.17	88.52	99.72	0.28	90.35	99.84	0.14	77.30	0.73	64.52	69.91	99.98	0.02
PreAct-Resnet18	LC	84.59	99.92	0.07	90.83	11.34	59.58	89.53	0.44	13.32	90.54	11.28	59.57	90.43	6.99	62.00	82.16	4.81	49.64	81.39	98.97	0.79	83.93	99.50	0.47	71.84	0.00	58.52	61.71	0.00	54.01
PreAct-Resnet18	SIG	84.48	98.27	1.72	90.80	2.37	69.74	89.10	26.20	20.61	90.02	10.66	64.20	84.48	98.27	1.72	77.97	0.01	48.80	82.41	94.61	5.17	83.01	92.27	7.40	57.80	0.00	59.97	60.67	100.00	0.00
PreAct-Resnet18	LF	93.19	99.28	0.71	92.37	78.44	19.42	92.90	98.97	1.02	92.37	47.83	47.49	91.62	1.41	87.48	85.15	0.60	82.72	88.92	98.09	1.72	92.58	98.67	1.28	81.72	0.21	84.81	50.98	99.70	0.08
PreAct-Resnet18	SSBA	92.88	97.86	1.99	92.14	74.79	23.31	92.54	83.50	15.36	91.91	77.40	20.86	90.99	0.58	87.04	86.54	0.39	79.87	90.00	96.23	3.53	89.63	90.50	8.71	80.79	0.00	84.66	63.50	99.51	0.39
PreAct-Resnet18	Input-aware	90.67	98.26	1.66	93.12	1.72	90.53	91.74	0.04	44.54	93.18	1.68	91.12	92.61	0.76	90.87	90.36	1.64	86.99	91.48	88.62	10.61	91.39	90.43	9.13	63.19	90.16	5.77	78.60	8.54	75.28
PreAct-Resnet18	WaNet	91.25	89.73	9.76	93.48	17.10	78.29	91.46	1.09	69.73	93.17	22.98	72.69	91.80	7.53	85.09	85.73	0.43	86.50	91.93	96.80	3.06	91.94	90.17	9.37	83.19	0.00	85.86	80.90	6.61	77.61
VGG-19	BadNets	89.36	95.93	3.81	88.08	25.42	69.68	89.23	92.61	6.82	87.51	38.17	58.30	87.86	1.00	88.01	NA	NA	NA	86.25	94.37	5.17	88.07	91.67	7.78	85.13	93.92	5.50	10.00	100.00	0.00
VGG-19	Blended	90.17	99.12	0.82	88.76	94.37	5.22	90.07	99.11	0.82	88.35	93.08	6.33	85.92	1.79	74.13	NA	NA	NA	87.58	98.13	1.76	89.18	99.76	0.23	88.61	99.48	0.46	10.00	100.00	0.00
VGG-19	LC	82.21	97.46	1.78	84.78	6.29	56.39	85.37	96.94	2.23	85.99	8.36	56.48	82.21	97.46	1.78	NA	NA	NA	77.85	36.31	39.68	80.85	62.94	26.56	81.08	97.59	1.71	10.00	0.00	11.10
VGG-19	SIG	81.69	99.80	0.12	86.89	98.16	1.68	84.52	99.93	0.07	86.01	99.18	0.77	81.69	99.80	0.12	NA	NA	NA	79.60	99.94	0.06	80.27	97.77	1.89	80.53	99.81	0.16	10.00	0.00	11.11
VGG-19	LF	88.94	93.93	5.62	87.88	62.84	33.01	88.98	91.80	7.46	87.85	59.38	35.01	85.35	9.99	72.79	NA	NA	NA	83.07	89.39	8.59	87.88	92.83	6.44	87.89	93.17	6.00	10.00	100.00	0.00
VGG-19	SSBA	89.48	91.86	7.29	88.20	46.28	47.83	89.40	89.66	9.22	87.65	37.54	54.58	89.48	91.86	7.29	NA	NA	NA	88.62	92.26	6.98	89.61	92.60	6.81	87.66	8.69	78.41	10.00	0.00	11.11
VGG-19	Input-aware	77.69	94.59	4.79	74.74	13.39	57.97	78.62	86.77	11.79	75.70	23.36	54.71	77.67	94.58	4.79	NA	NA	NA	83.52	82.57	15.20	87.22	84.71	13.58	88.30	86.48	12.00	10.00	100.00	0.00
VGG-19	WaNet	88.43	88.90	10.30	91.03	30.04	64.93	89.61	73.39	24.57	90.82	44.93	51.18	88.43	88.89	10.30	NA	NA	NA	87.96	92.56	7.02	87.76	96.24	3.53	86.29	91.77	7.22	10.00	100.00	0.00
EfficientNet-B3	BadNets	57.48	9.51	55.59	53.69	5.09	52.96	57.97	3.88	57.46	55.29	4.50	54.68	57.48	9.53	55.58	57.48	9.53	55.58	39.76	17.28	36.51	37.64	15.50	36.11	36.44	8.51	36.29	54.44	94.26	3.91
EfficientNet-B3	Blended	65.11	87.76	8.01	61.16	8.60	44.66	65.21	59.03	24.14	62.76	11.86	46.32	59.93	3.87	51.80	65.11	87.76	8.01	34.58	63.32	15.88	53.68	70.81	17.58	18.14	1.91	15.51	52.10	99.94	0.04
EfficientNet-B3	LC	62.09	1.61	47.72	60.25	7.71	41.83	64.23	6.18	45.48	62.79	6.62	43.46	62.10	1.61	47.73	62.10	1.61	47.73	39.81	0.21	39.38	47.70	0.10	42.64	20.44	0.00	19.62	60.13	0.00	53.02
EfficientNet-B3	SIG	58.14	99.71	0.13	56.92	3.98	24.03	58.83	0.12	10.84	58.27	3.63	25.83	58.15	99.71	0.13	58.15	99.71	0.13	46.07	99.16	0.26	52.64	98.14	0.76	11.79	9.58	32.02	45.49	99.96	0.04
EfficientNet-B3	LF	56.95	80.06	10.51	53.91	26.43	31.87	56.71	64.11	18.17	55.09	34.42	30.13	52.09	16.62	37.49	53.01	67.98	15.43	34.42	72.51	9.41	52.27	68.96	15.03	41.86	32.03	32.39	50.48	99.21	0.58
EfficientNet-B3	SSBA	55.76	24.40	45.50	53.42	7.10	49.32	56.49	6.07	52.97	54.70	6.91	50.72	55.77	24.40	45.49	55.77	24.40	45.49	29.94	29.13	25.88	49.67	13.80	45.66	27.37	15.74	24.19	51.82	99.27	0.59
EfficientNet-B3	Input-aware	70.01	94.96	2.34	67.84	50.37	17.97	65.61	7.03	15.66	70.30	48.93	22.64	67.47	54.81	8.14	70.01	29.92	27.78	47.20	99.39	0.07	42.22	98.13	0.26	17.30	16.47	17.13	54.13	100.00	0.00
EfficientNet-B3	WaNet	71.73	7.93	69.33	70.80	3.59	67.96	25.27	7.43	26.20	73.23	3.14	70.88	71.74	7.93	69.33	71.74	7.93	69.33	37.78	14.34	35.28	52.44	7.11	51.91	37.97	7.66	35.89	58.83	18.44	54.87
MobileNetV3-Large	BadNets	82.45	93.39	5.84	77.99	3.00	77.00	82.36	93.97	5.32	79.02	4.57	76.22	75.19	2.53	74.62	82.04	3.72	80.87	70.46	89.98	7.77	78.12	92.83	5.93	76.57	83.59	13.33	57.21	86.84	9.67
MobileNetV3-Large	Blended	83.43	98.76	1.13	78.94	12.27	53.93	83.13	1.46	19.52	79.88	18.81	52.72	77.59	3.53	65.72	75.56	95.02	3.97	71.48	95.08	3.33	79.18	98.16	1.46	76.00	79.02	15.63	51.67	99.64	0.31
MobileNetV3-Large	LC	75.53	98.67	0.84	76.54	6.70	49.44	80.79	0.13	11.80	78.11	6.50	49.93	76.19	6.08	49.88	72.72	0.00	52.46	66.01	95.06	2.73	72.31	98.46	0.87	67.38	99.99	0.01	56.71	0.00	49.44
MobileNetV3-Large	SIG	77.13	98.70	1.04	78.09	2.51	42.93	82.00	0.00	17.41	79.10	1.87	37.86	77.13	98.70	1.04	73.49	78.30	14.40	67.85	99.33	0.46	73.47	97.99	1.33	63.20	90.66	5.63	48.42	100.00	0.00
MobileNetV3-Large	LF	82.73	96.63	1.77	77.92	24.54	44.76	82.22	0.60	13.33	78.98	10.08	57.00	82.73	96.63	1.77	75.45	5.44	58.76	72.37	96.33	2.32	79.61	96.58	2.42	75.32	64.92	25.42	51.40	99.37	0.49
MobileNetV3-Large	SSBA	81.84	85.06	12.16	77.80	6.94	68.71	82.14	62.99	29.57	78.35	9.51	68.06	75.95	3.96	71.43	76.26	58.59	30.71	67.79	57.71	29.01	76.85	72.18	21.61	76.22	31.27	50.26	53.93	97.93	1.51
MobileNetV3-Large	Input-aware	77.35	89.17	9.12	78.20	7.09	62.23	79.80	3.32	57.37	79.82	4.46	64.80	76.77	4.37	67.01	77.35	89.17	9.12	71.36	69.29	21.37	78.58	75.03	19.74	73.96	78.38	15.32	50.01	99.98	0.01
MobileNetV3-Large	WaNet	78.00	70.88	24.80	80.40	4.14	77.22	24.71	9.89	20.82	81.81	3.56	78.97	78.00	70.88	24.80	78.57	4.71	76.30	69.61	43.82	41.47	77.59	53.23	36.92	59.83	58.71	29.79	58.66	12.30	54.99
DenseNet-161	BadNets	84.33	89.68	9.10	83.59	32.70	56.87	85.16	87.51	10.81	82.28	68.73	26.47	83.10	1.84	82.36	76.53	0.29	79.92	73.17	79.32	16.60	81.13	87.71	10.59	76.32	1.02	79.38	67.41	15.23	64.12
DenseNet-161	Blended	86.37	98.79	1.04	84.39	81.83	14.74	85.95	98.98	0.88	83.69	72.82	19.60	83.17	1.01	72.90	78.93	2.06	65.93	75.45	94.31	4.23	82.06	97.01	2.44	67.81	2.16	73.28	56.66	99.53	0.40
DenseNet-161	LC	78.68	78.06	13.88	83.27	9.91	52.53	83.92	41.48	36.57	82.56	9.10	52.42	81.20	7.69	51.92	81.97	20.44	47.62	70.16	12.41	47.31	75.94	61.46	24.82	59.56	51.84	21.53	72.36	0.00	60.77
DenseNet-161	SIG	78.63	98.67	1.23	83.49	19.30	49.21	83.23	27.57	38.66	82.70	18.47	48.69	81.62	7.99	43.91	74.25	2.64	45.37	69.85	98.34	1.33	75.61	98.71	1.18	43.47	0.00	60.47	45.40	96.77	3.10
DenseNet-161	LF	84.32	91.70	6.90	83.07	62.07	31.20	83.93	92.04	6.42	82.51	62.29	30.13	81.11	10.73	67.60	78.50	6.73	70.53	72.91	83.24	12.26	80.61	88.47	9.20	64.40	23.01	53.82	59.62	98.29	1.46
DenseNet-161	SSBA	84.18	84.13	13.39	83.76	32.84	55.90	84.26	83.76	13.49	82.80	29.79	57.22	81.68	8.49	71.78	76.82	1.23	74.62	71.94	60.54	29.57	79.68	75.13	20.34	68.13	0.59	77.33	67.25	97.62	1.93
DenseNet-161	Input-aware	84.46	94.41	5.18	86.77	2.69	82.02	85.15	22.34	60.34	86.47	10.31	76.64	84.45	94.41	5.18	84.03	4.06	75.08	74.19	89.66	7.73	81.64	91.82	6.68	63.79	13.14	51.26	74.12	77.46	17.14
DenseNet-161	WaNet	84.61	73.81	22.72	87.52	2.34	84.48	85.72	24.61	64.71	87.30	6.26	81.16	87.10	1.86	85.07	84.21	1.24	83.69	72.07	41.99	43.37	80.04	65.82	28.01	30.37	97.74	1.58	65.25	10.98	62.00

Table 11: Full results on CIFAR-10 with 5% poisoning ratio.

	Backdoor Defense →	No defense			FT			FP			NAD			NC			ANP			AC			Spectral			ABL			DBD		
Targeted Model	Backdoor Attack ↓	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
PreAct-Resnet18	BadNets	92.64	88.74	10.78	91.22	4.49	88.22	92.47	17.47	77.90	91.03	4.73	87.93	89.89	1.17	89.70	87.24	0.34	87.34	90.27	75.61	23.58	91.21	69.40	29.21	84.08	0.00	90.50	87.68	2.11	87.01
PreAct-Resnet18	Blended	93.67	99.61	0.39	93.25	98.78	1.21	93.46	98.87	1.11	93.10	99.06	0.93	93.66	99.61	0.39	88.00	11.98	58.50	91.54	98.93	1.03	91.43	99.42	0.56	65.78	0.00	48.48	75.00	99.99	0.01
PreAct-Resnet18	LC	93.29	99.50	0.43	92.54	96.34	2.81	93.38	96.71	2.60	92.74	97.62	1.89	91.51	3.58	64.46	84.07	5.06	53.84	89.33	97.29	1.89	92.56	99.50	0.40	75.91	99.37	0.44	81.88	5.83	60.38
PreAct-Resnet18	SIG	93.64	97.09	2.90	92.84	96.42	3.46	93.27	99.79	0.20	92.49	96.98	2.93	93.65	97.09	2.90	86.64	0.21	49.61	90.24	93.32	6.31	91.79	96.36	3.31	46.14	0.00	42.30	74.86	95.58	4.22
PreAct-Resnet18	LF	93.35	98.03	1.86	93.21	93.62	6.02	93.14	98.37	1.53	92.98	94.23	5.30	93.35	98.03	1.86	86.32	2.66	82.79	89.02	95.14	4.29	92.36	97.32	2.49	66.54	98.02	1.27	83.98	15.67	73.11
PreAct-Resnet18	SSBA	93.27	94.91	4.82	93.08	87.46	11.69	93.05	87.14	12.09	92.49	88.63	10.58	91.04	0.80	87.93	87.53	0.66	81.63	88.47	92.64	6.63	92.01	93.19	6.38	81.60	0.09	79.76	72.19	11.34	65.42
PreAct-Resnet18	Input-aware	91.52	90.20	8.91	93.12	89.34	9.99	91.89	52.74	44.49	93.07	97.12	2.68	91.51	90.20	8.91	91.54	2.22	86.40	92.07	92.20	7.36	92.02	83.81	14.84	58.14	92.58	4.84	76.23	6.12	71.99
PreAct-Resnet18	WaNet	91.76	85.50	13.49	93.47	31.32	64.03	92.14	26.10	63.50	93.31	50.40	46.46	91.76	85.50	13.49	87.21	0.57	87.69	91.96	88.72	10.88	91.47	83.84	15.16	69.49	95.20	3.79	72.22	9.93	69.41
VGG-19	BadNets	89.83	94.36	5.39	89.14	50.56	47.02	89.81	90.32	9.09	88.42	74.51	23.89	88.51	1.04	88.31	NA	NA	NA	88.26	93.10	6.34	88.79	90.78	8.58	86.03	91.93	7.30	10.00	0.00	11.11
VGG-19	Blended	90.41	97.98	1.79	88.58	90.47	8.72	90.21	97.93	1.84	88.84	92.68	6.70	86.92	2.61	73.99	NA	NA	NA	87.76	97.69	2.16	88.98	96.72	2.93	88.99	98.44	1.38	10.00	100.00	0.00
VGG-19	LC	89.08	43.06	40.10	88.88	9.06	56.80	89.34	38.50	42.84	88.15	9.63	55.72	89.08	43.06	40.10	NA	NA	NA	82.71	22.52	47.77	87.12	31.88	45.01	87.43	32.10	44.93	10.00	0.00	11.11
VGG-19	SIG	90.41	96.76	2.84	89.10	88.24	9.74	90.55	97.72	2.12	88.96	91.34	7.30	90.41	96.76	2.84	NA	NA	NA	87.27	98.78	0.90	88.99	97.44	2.34	88.77	98.67	0.94	10.00	0.00	11.11
VGG-19	LF	90.13	86.04	12.50	88.83	55.80	39.23	89.86	83.69	14.77	88.82	54.51	40.07	87.25	8.58	76.39	NA	NA	NA	83.83	79.07	17.70	87.98	83.29	14.96	88.14	87.57	10.91	10.00	100.00	0.00
VGG-19	SSBA	89.93	81.19	17.14	88.76	28.17	63.28	89.81	78.61	19.68	88.07	43.93	48.86	87.31	2.27	83.27	NA	NA	NA	85.84	90.62	8.31	85.70	89.97	8.93	88.50	78.44	19.02	10.00	100.00	0.00
VGG-19	Input-aware	89.34	71.66	25.77	90.95	67.67	30.98	89.90	52.58	43.51	90.51	81.62	17.73	89.58	2.88	85.30	NA	NA	NA	89.75	63.24	34.06	89.85	87.07	12.23	88.47	73.49	23.22	10.00	100.00	0.00
VGG-19	WaNet	89.25	3.96	86.71	90.71	1.51	89.57	90.25	1.66	89.08	90.61	1.67	89.40	89.25	3.96	86.69	NA	NA	NA	88.46	36.68	57.16	89.26	70.89	26.74	87.86	61.62	34.01	10.00	100.00	0.00
EfficientNet-B3	BadNets	65.64	5.89	63.88	59.52	4.87	58.43	65.25	3.88	63.88	61.81	5.16	60.73	59.43	5.26	58.31	65.63	5.89	63.88	22.57	29.76	18.66	53.12	5.74	53.69	28.52	0.59	30.34	61.60	7.78	58.69
EfficientNet-B3	Blended	60.01	74.83	16.40	56.49	12.34	46.38	60.17	40.16	35.48	57.84	14.21	46.59	56.36	6.81	50.76	60.03	74.83	16.40	42.21	34.89	29.42	48.44	48.06	27.62	21.82	1.67	21.02	54.29	99.76	0.16
EfficientNet-B3	LC	64.51	4.21	47.20	60.11	6.26	42.83	64.17	5.24	46.06	61.92	6.34	43.88	64.51	4.21	47.20	64.52	4.21	47.20	41.33	3.44	35.99	45.58	2.17	40.57	35.42	0.70	30.23	56.91	0.00	50.79
EfficientNet-B3	SIG	58.53	99.14	0.49	55.06	51.48	16.71	58.40	99.90	0.06	57.17	77.53	6.64	58.53	99.14	0.49	58.53	99.14	0.49	51.52	95.52	1.06	49.99	78.09	7.32	15.16	3.90	30.32	51.11	99.84	0.16
EfficientNet-B3	LF	65.28	72.09	18.27	59.67	24.57	40.54	64.82	61.31	25.09	62.17	31.12	39.78	65.28	72.09	18.29	59.63	63.22	22.91	39.57	62.76	12.42	45.71	45.27	22.03	19.26	1.18	26.59	53.88	92.93	5.76
EfficientNet-B3	SSBA	54.95	9.90	50.79	51.60	6.06	48.04	55.34	4.92	52.62	53.56	5.18	50.54	49.90	6.07	46.79	54.95	9.90	50.79	21.40	24.82	15.90	33.54	7.20	32.74	41.18	2.46	43.70	51.90	94.34	4.20
EfficientNet-B3	Input-aware	69.70	76.47	14.31	68.92	24.61	29.56	70.82	92.27	4.89	71.31	38.64	29.27	69.70	76.48	14.31	67.50	68.93	18.04	26.39	25.76	15.89	45.29	82.23	5.24	37.72	20.82	23.64	53.93	100.00	0.00
EfficientNet-B3	WaNet	75.39	6.54	72.43	75.09	3.40	71.98	22.92	9.50	23.82	76.70	3.04	73.76	75.40	6.54	72.43	75.38	6.54	72.43	32.49	14.92	30.50	50.28	5.21	49.14	31.79	2.16	30.56	60.35	14.66	57.27
MobileNetV3-Large	BadNets	83.18	90.00	8.71	78.11	8.72	72.71	83.10	88.03	10.32	79.63	27.41	58.87	78.49	2.26	78.38	80.28	2.14	79.68	72.75	18.83	62.50	79.63	87.69	10.51	78.45	60.51	32.31	53.18	13.63	47.53
MobileNetV3-Large	Blended	83.42	97.33	2.20	78.90	23.47	45.70	83.23	93.77	3.99	79.77	39.41	38.86	77.04	1.24	64.69	79.78	85.10	9.87	74.29	90.76	6.42	79.64	95.78	3.22	75.78	41.17	40.08	57.67	99.28	0.56
MobileNetV3-Large	LC	84.22	95.91	2.50	79.71	19.31	42.97	84.10	89.06	4.77	80.16	19.51	42.46	78.37	5.14	50.92	78.74	1.58	50.86	70.86	8.96	46.90	79.01	15.07	47.02	73.13	7.29	47.21	58.75	4.89	43.84
MobileNetV3-Large	SIG	83.94	96.22	2.94	79.24	9.37	38.82	83.36	39.13	6.80	80.44	1.90	35.63	76.71	2.13	26.93	78.81	91.87	5.33	73.52	97.22	1.38	80.17	97.47	1.84	69.30	77.48	10.98	56.42	99.83	0.16
MobileNetV3-Large	LF	83.72	94.53	3.68	79.08	20.36	42.89	83.43	2.10	18.79	79.87	33.41	35.68	83.72	94.53	3.69	79.78	93.99	4.48	71.50	92.72	4.69	79.95	94.10	4.31	75.86	20.68	45.43	52.09	98.72	0.93
MobileNetV3-Large	SSBA	82.61	69.64	24.54	78.82	5.88	69.96	83.11	49.07	41.52	79.71	5.94	70.81	77.47	4.11	71.47	81.30	13.74	67.24	72.00	28.52	51.40	77.75	39.94	47.41	73.52	33.99	47.99	59.90	8.68	54.46
MobileNetV3-Large	Input-aware	74.68	81.37	14.32	79.10	8.98	60.80	81.66	11.82	55.38	80.94	13.62	60.89	74.69	81.37	14.32	69.82	4.00	59.12	72.51	69.34	20.30	79.26	71.50	22.36	75.14	73.21	18.50	57.04	42.62	40.57
MobileNetV3-Large	WaNet	79.32	2.89	77.70	80.48	3.87	77.00	25.72	6.04	25.84	81.90	3.17	78.84	79.31	2.89	77.69	77.00	1.68	77.02	70.73	13.66	62.89	78.55	19.11	65.11	71.38	14.89	64.13	58.93	10.87	55.47
DenseNet-161	BadNets	85.45	79.87	17.67	84.25	29.62	61.13	85.64	67.23	28.78	83.21	18.31	69.02	83.65	1.86	83.31	76.93	0.81	77.64	74.54	49.81	39.56	82.18	75.22	21.18	76.24	1.30	79.72	69.28	7.92	66.36
DenseNet-161	Blended	86.13	96.62	2.82	85.39	80.74	15.21	86.12	97.39	2.21	84.33	84.80	12.00	83.85	1.11	72.52	79.02	8.52	58.57	75.87	86.00	10.60	83.54	93.57	5.29	69.55	0.96	70.88	61.29	99.63	0.27
DenseNet-161	LC	85.48	13.72	52.01	84.77	6.90	53.80	85.73	18.38	50.31	83.42	6.77	53.48	80.28	5.21	52.33	84.12	7.73	53.79	75.27	7.83	49.63	82.10	10.42	51.54	52.71	7.88	33.29	73.46	5.40	53.57
DenseNet-161	SIG	86.11	95.21	4.31	84.50	90.47	8.01	86.17	99.63	0.34	84.34	93.00	6.27	86.11	95.21	4.31	80.15	1.07	52.21	75.42	95.82	2.83	83.18	95.77	3.29	46.71	0.00	59.73	71.73	99.74	0.26
DenseNet-161	LF	85.68	82.56	14.63	84.44	61.66	32.17	85.33	85.21	12.29	84.12	63.42	29.66	85.68	82.56	14.64	78.71	14.82	62.93	74.10	67.89	22.76	82.06	71.04	24.11	63.15	75.29	16.44	69.51	23.63	58.00
DenseNet-161	SSBA	84.85	60.30	33.34	84.04	21.57	64.80	85.08	60.28	33.21	83.06	24.70	61.50	80.66	2.76	76.33	79.57	1.87	75.02	74.06	32.03	51.08	81.67	35.71	53.11	73.42	3.30	73.79	57.81	57.22	33.79
DenseNet-161	Input-aware	84.85	79.39	18.21	86.90	19.72	69.07	85.15	23.92	47.97	87.12	39.70	53.51	84.95	5.60	78.01	80.35	5.03	69.46	75.33	68.40	23.80	82.04	67.87	25.66	64.93	65.19	20.90	61.04	41.13	42.82
DenseNet-161	WaNet	83.47	36.49	55.96	87.57	1.47	85.53	85.60	2.20	82.43	87.49	1.69	85.42	86.56	1.41	85.31	78.95	0.93	78.93	74.42	12.70	65.02	81.20	27.40	59.36	51.05	58.31	25.43	68.03	5.57	66.28

Table 12: Full results on CIFAR-10 with 1% poisoning ratio.

	Backdoor Defense	No defense			FT			FP			NAD			NC			ANP			AC			Spectral			ABL			DBD		
Targeted Model	Backdoor Attack	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
PreAct-Resnet18	BadNets	93.14	74.73	24.24	93.07	10.36	84.58	92.98	20.18	75.87	92.52	8.33	85.96	91.64	0.88	91.68	84.67	0.29	85.30	88.88	27.18	68.14	92.62	74.20	24.88	72.81	53.20	41.31	78.09	2.99	76.51
PreAct-Resnet18	Blended	93.76	94.88	4.79	93.53	91.96	7.56	93.48	94.14	5.41	93.34	92.07	7.41	90.98	65.91	29.24	86.92	37.01	40.68	89.77	86.09	12.67	93.27	93.32	6.28	66.26	0.17	34.91	70.18	8.04	55.60
PreAct-Resnet18	LC	93.78	90.63	7.14	93.54	87.16	9.53	93.61	80.14	15.04	93.25	87.16	9.74	91.83	3.67	64.41	89.57	10.61	54.67	89.44	12.12	56.28	92.57	37.36	44.67	68.49	11.79	40.93	76.07	6.47	56.98
PreAct-Resnet18	SIG	93.82	83.40	14.29	93.59	88.17	10.13	93.48	88.07	10.08	93.41	83.20	14.22	93.82	83.40	14.29	91.82	23.63	50.47	90.00	74.81	18.61	93.09	85.91	12.12	64.33	0.00	20.51	75.01	67.82	29.87
PreAct-Resnet18	LF	93.56	86.46	12.88	93.35	82.54	16.36	93.25	86.58	12.60	93.34	84.43	14.64	93.56	86.46	12.87	88.11	6.36	78.78	89.49	76.22	21.40	92.68	82.19	16.68	60.04	55.07	30.76	79.13	7.47	75.38
PreAct-Resnet18	SSBA	93.43	73.44	24.89	93.24	61.32	35.98	93.34	52.76	43.53	93.33	67.03	30.83	90.45	1.22	85.77	91.68	2.19	87.52	89.61	27.92	64.11	92.71	54.79	42.18	83.11	56.33	37.69	78.52	1.13	74.21
PreAct-Resnet18	Input-aware	91.74	79.18	19.89	93.14	77.68	21.59	92.20	53.69	43.89	93.07	85.13	14.37	91.74	79.19	19.88	85.88	7.54	75.41	89.44	60.56	36.07	92.78	60.14	37.84	72.26	69.31	24.89	69.66	14.36	62.67
PreAct-Resnet18	WaNet	90.65	12.63	79.94	93.61	0.68	92.72	92.07	0.93	91.10	93.61	0.78	92.67	90.65	12.66	79.92	84.96	0.59	86.37	89.53	4.87	84.87	92.79	8.81	84.56	65.43	53.43	37.98	79.74	4.60	78.46
VGG-19	BadNets	90.45	83.98	15.14	89.12	55.17	42.29	90.38	78.61	20.21	88.69	64.76	32.97	88.97	1.09	88.82	NA	NA	NA	85.63	24.59	67.54	89.01	76.46	22.03	61.58	0.00	63.51	10.00	100.00	0.00
VGG-19	Blended	90.47	89.53	9.62	89.47	79.83	18.20	90.35	89.63	9.36	88.91	77.71	19.63	90.47	89.53	9.62	NA	NA	NA	85.58	75.17	20.99	89.19	80.57	17.41	89.42	91.72	7.50	10.00	0.00	11.11
VGG-19	LC	89.97	12.82	57.99	88.88	5.86	59.07	90.24	9.06	59.30	88.46	6.70	58.13	89.97	12.82	57.99	NA	NA	NA	85.33	9.48	53.43	89.01	10.94	56.49	88.99	8.10	57.00	10.00	0.00	11.11
VGG-19	SIG	90.85	83.77	12.48	89.73	78.64	16.28	90.90	84.90	11.68	88.96	81.96	13.19	90.85	83.77	12.48	NA	NA	NA	85.13	81.56	10.79	89.60	84.94	10.11	88.90	81.74	10.44	10.00	0.00	11.11
VGG-19	LF	90.21	37.54	56.47	89.28	14.09	75.01	90.08	33.28	60.03	88.61	17.16	72.53	86.32	13.96	70.97	NA	NA	NA	85.06	29.23	57.84	88.99	35.40	56.44	89.08	40.22	51.77	10.00	0.00	11.11
VGG-19	SSBA	89.96	10.97	79.72	89.12	2.07	85.58	90.08	6.51	82.83	88.41	3.54	83.54	86.30	1.82	83.72	NA	NA	NA	84.44	9.38	75.92	89.32	11.18	78.73	83.86	41.68	50.47	10.00	0.00	11.11
VGG-19	Input-aware	89.47	58.09	37.13	90.53	31.96	59.36	89.36	34.62	56.88	90.52	33.00	58.20	89.56	3.12	79.97	NA	NA	NA	85.47	51.47	39.04	89.47	59.71	33.80	89.07	57.51	33.98	10.00	100.00	0.00
VGG-19	WaNet	89.99	1.71	88.73	90.99	0.98	90.36	90.19	1.03	89.39	91.10	0.99	89.96	88.10	1.86	86.39	NA	NA	NA	85.59	5.10	81.23	89.13	7.22	82.47	88.77	4.46	84.63	10.00	100.00	0.00
EfficientNet-B3	BadNets	53.13	3.80	52.84	51.77	5.11	51.18	53.02	3.71	52.91	51.90	4.80	51.68	53.14	3.80	52.86	53.14	3.80	52.86	41.98	5.04	41.56	52.04	5.06	51.59	33.27	0.38	36.04	54.48	14.17	50.13
EfficientNet-B3	Blended	61.57	10.86	53.18	56.29	6.03	50.60	61.45	7.81	53.77	58.03	7.46	50.78	55.16	7.88	48.56	61.56	10.86	53.18	43.94	7.39	38.89	47.60	7.32	42.41	35.26	3.80	31.24	54.63	21.17	44.53
EfficientNet-B3	LC	56.94	6.16	43.44	53.54	7.39	40.60	56.75	6.98	42.86	53.89	7.32	40.70	56.94	6.16	43.42	56.93	6.16	43.43	47.77	5.07	39.94	56.44	4.88	43.53	31.63	0.87	27.87	53.83	6.44	42.61
EfficientNet-B3	SIG	63.63	3.87	57.93	58.45	4.19	53.78	63.40	3.87	56.98	59.74	3.84	54.73	56.33	3.91	52.39	63.63	3.87	57.93	35.75	5.22	30.99	58.10	50.37	22.01	19.45	0.43	22.10	57.50	92.52	6.33
EfficientNet-B3	LF	54.70	24.88	35.17	52.61	15.18	38.44	54.35	22.14	36.57	52.77	18.54	37.48	54.70	24.88	35.21	54.72	24.89	35.19	42.25	27.97	22.99	46.90	19.41	33.39	34.85	4.46	40.93	53.93	22.30	45.90
EfficientNet-B3	SSBA	59.42	4.49	55.76	54.16	4.97	51.54	59.00	4.17	55.98	56.16	4.17	53.53	53.40	4.31	50.70	59.41	4.49	55.77	47.26	4.76	45.46	44.90	3.72	44.13	20.08	0.44	21.16	60.79	9.40	55.36
EfficientNet-B3	Input-aware	55.80	5.49	51.03	57.46	5.72	52.08	57.59	4.66	53.21	57.65	5.52	52.86	55.80	5.49	51.03	55.80	5.49	51.03	26.59	3.33	26.52	51.95	4.92	50.17	26.46	0.49	25.91	58.61	13.90	49.27
EfficientNet-B3	WaNet	73.07	2.36	72.02	70.35	3.67	68.36	73.11	3.09	70.62	72.37	3.16	70.53	73.07	2.36	72.03	73.08	2.36	72.02	37.20	6.48	36.21	48.72	4.07	48.66	24.81	3.97	24.49	55.71	11.70	51.37
MobileNetV3-Large	BadNets	84.46	3.29	83.43	79.60	2.90	78.88	84.49	2.39	83.86	80.57	3.06	79.41	76.69	3.18	75.83	84.46	3.29	83.44	71.90	3.29	71.38	80.26	2.27	79.89	74.35	2.20	73.61	58.82	10.08	55.67
MobileNetV3-Large	Blended	83.97	87.06	10.17	79.49	34.61	43.20	83.68	86.11	11.04	80.32	49.88	33.21	75.70	1.31	64.01	83.97	87.06	10.17	73.15	54.56	29.51	80.54	73.33	19.72	77.90	20.27	50.97	56.87	5.32	49.92
MobileNetV3-Large	LC	84.16	6.07	54.64	79.79	5.99	51.09	84.03	5.84	54.31	80.39	5.84	51.68	84.16	6.07	54.64	84.16	6.07	54.64	75.08	5.41	50.02	79.99	5.63	52.07	73.65	6.74	47.09	57.11	11.82	40.33
MobileNetV3-Large	SIG	84.48	80.79	10.77	79.56	56.01	19.70	84.14	83.76	8.98	80.45	51.67	24.00	84.47	80.78	10.78	81.53	76.63	11.02	74.42	83.90	6.59	80.67	93.54	3.30	75.36	38.56	22.60	57.16	96.63	3.08
MobileNetV3-Large	LF	83.08	86.32	9.90	78.38	43.78	29.79	83.01	82.91	10.43	79.35	51.60	30.91	83.08	86.32	9.90	78.15	82.87	13.17	73.08	77.50	15.39	80.14	81.94	14.14	73.96	82.94	11.36	57.34	46.93	36.30
MobileNetV3-Large	SSBA	83.28	9.71	73.78	79.80	3.34	74.14	83.47	5.91	76.06	80.32	3.46	74.42	77.10	3.44	73.00	83.29	9.72	73.78	72.41	7.77	64.59	79.81	6.82	72.84	74.49	6.96	70.17	55.41	12.18	47.22
MobileNetV3-Large	Input-aware	80.80	90.72	7.16	80.36	30.52	46.96	82.48	97.43	2.16	81.48	67.33	22.24	80.80	90.72	7.16	77.21	90.06	7.29	73.48	58.49	26.27	80.07	58.91	28.87	75.28	62.62	22.73	57.37	13.47	47.87
MobileNetV3-Large	WaNet	80.86	1.66	80.02	81.45	2.29	79.06	81.75	3.31	79.11	82.29	2.17	80.59	78.78	2.98	76.29	80.86	1.66	80.02	73.11	3.36	71.50	80.71	3.29	78.17	74.24	1.73	74.30	57.06	13.52	53.11
DenseNet-161	BadNets	86.30	16.32	73.38	85.26	3.47	82.79	86.37	5.34	82.39	84.78	3.81	82.31	86.31	16.32	73.37	84.66	1.99	83.98	76.19	3.10	74.84	83.52	1.76	83.23	64.73	14.99	55.49	70.15	4.59	67.93
DenseNet-161	Blended	86.65	79.32	16.94	85.51	61.74	31.10	86.62	80.12	16.32	84.99	59.03	32.52	83.29	34.52	48.90	79.59	23.09	56.71	75.87	31.92	49.54	83.77	1.98	73.92	63.97	12.18	45.19	69.93	3.24	63.78
DenseNet-161	LC	86.56	5.31	56.19	85.20	5.24	55.74	86.73	7.53	55.01	84.45	5.71	54.50	83.25	5.11	53.88	85.27	4.87	55.83	76.68	5.13	50.01	83.23	5.16	54.30	57.96	5.21	36.70	64.14	8.63	47.22
DenseNet-161	SIG	86.63	82.66	13.49	85.15	75.18	17.83	86.45	90.18	8.10	84.83	74.54	17.31	86.62	82.66	13.49	80.20	22.66	43.33	76.56	67.29	18.46	83.35	71.78	17.77	60.91	0.00	64.33	63.01	96.00	3.42
DenseNet-161	LF	86.50	39.24	51.92	85.30	25.14	62.66	86.55	36.29	54.33	84.59	25.77	62.47	81.74	12.98	68.17	81.30	8.97	71.61	76.28	33.22	48.16	83.62	9.50	72.83	58.25	40.64	38.84	70.35	6.90	65.79
DenseNet-161	SSBA	86.31	6.56	79.24	85.06	3.33	80.96	86.30	4.56	80.80	85.20	4.89	79.11	86.33	6.58	79.24	82.14	2.00	77.64	76.03	8.27	68.61	83.34	2.98	79.43	60.19	10.17	54.78	72.51	4.87	67.28
DenseNet-161	Input-aware	84.21	66.37	25.18	87.30	86.13	11.41	85.31	97.56	1.84	86.90	88.99	8.88	84.21	66.39	25.16	78.26	2.36	64.17	76.47	59.51	28.89	82.86	71.41	22.10	58.24	60.71	21.26	64.53	5.49	57.70
DenseNet-161	WaNet	85.73	2.17	84.99	87.99	1.17	86.56	85.33	2.59	83.23	87.56	1.32	85.92	86.31	1.46	84.94	83.74	1.62	83.32	76.44	2.98	74.06	82.87	2.71	79.66	55.54	4.13	49.88	69.10	7.19	65.28

Table 13: Full results on CIFAR-10 with 0.5% poisoning ratio.

	Backdoor Defense →	No defense			FT			FP			NAD			NC			ANP			AC			Spectral			ABL			DBD		
Targeted Model	Backdoor Attack ↓	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
PreAct-Resnet18	BadNets	93.76	50.06	48.14	93.49	4.94	89.83	93.67	5.01	89.60	93.30	7.67	87.31	92.83	0.62	92.67	91.32	0.76	91.31	89.51	2.29	88.72	92.97	7.91	87.31	68.40	71.78	22.14	73.60	0.10	80.01
PreAct-Resnet18	Blended	93.68	93.30	6.37	93.46	90.68	8.79	93.30	93.28	6.39	93.48	91.97	7.60	90.64	60.71	33.92	89.90	53.54	36.19	90.14	75.23	22.02	93.08	88.18	10.91	61.89	26.51	38.22	71.18	0.01	71.24
PreAct-Resnet18	LC	93.66	18.61	58.08	93.55	10.82	62.28	93.77	10.43	62.17	93.64	12.67	61.42	89.58	4.06	61.68	91.50	6.60	61.48	89.54	7.31	58.98	92.76	13.64	60.19	60.68	11.68	36.51	78.33	3.73	60.48
PreAct-Resnet18	SIG	93.80	82.43	14.90	93.51	81.77	15.53	93.69	84.59	12.18	93.65	81.64	15.43	93.80	82.43	14.90	90.11	14.97	50.53	89.66	65.29	25.72	93.26	64.04	27.10	70.52	85.82	6.64	77.35	72.37	25.28
PreAct-Resnet18	LF	93.85	75.09	23.57	93.62	69.64	28.33	93.31	74.16	24.02	93.37	71.56	26.64	93.85	75.09	23.57	85.00	4.43	76.66	89.21	63.48	32.44	92.68	71.50	26.83	82.88	77.69	18.56	88.95	3.63	86.48
PreAct-Resnet18	SSBA	93.41	35.67	59.99	93.36	27.97	66.46	93.37	16.99	76.30	93.28	30.91	63.73	93.41	35.66	60.00	85.04	1.66	78.94	89.90	9.72	80.02	92.96	23.64	69.98	74.04	45.26	42.54	80.12	7.00	75.22
PreAct-Resnet18	Input-aware	90.66	56.06	41.27	93.22	75.64	23.93	91.93	73.21	25.83	93.36	79.54	20.07	90.66	56.07	41.27	88.61	9.02	74.30	89.82	62.34	33.44	92.98	60.34	37.51	66.84	62.16	22.96	77.50	3.80	71.36
PreAct-Resnet18	WaNet	91.27	1.12	91.24	93.99	0.76	92.83	92.34	1.11	91.49	93.67	0.89	92.67	91.27	1.12	91.24	91.27	1.12	91.24	89.48	2.57	86.66	92.94	3.59	89.34	42.48	75.28	19.77	64.90	0.00	72.59
VGG-19	BadNets	90.28	3.71	88.48	89.10	1.53	88.62	90.09	2.02	89.38	88.65	1.76	88.09	87.78	1.59	87.47	NA	NA	NA	85.36	1.93	84.76	89.21	16.31	77.04	60.84	92.12	7.18	10.00	0.00	11.11
VGG-19	Blended	90.85	74.69	22.60	89.54	57.18	37.21	90.70	74.49	22.49	88.65	63.59	31.11	90.85	74.69	22.60	NA	NA	NA	85.21	50.44	39.43	89.60	69.44	26.39	89.15	78.13	18.86	10.00	0.00	11.11
VGG-19	LC	90.63	8.27	60.39	89.45	5.09	60.03	90.76	6.48	60.78	89.10	4.86	59.49	90.63	8.27	60.40	NA	NA	NA	85.39	6.49	54.67	89.51	7.17	58.40	89.12	5.70	58.86	10.00	0.00	11.11
VGG-19	SIG	90.76	72.92	18.26	89.60	59.68	26.41	90.64	73.42	17.93	89.25	53.72	30.00	86.54	57.06	25.78	NA	NA	NA	85.06	59.00	21.76	89.28	54.40	29.84	88.99	71.69	16.50	10.00	0.00	11.11
VGG-19	LF	90.54	20.46	71.11	89.29	8.50	79.13	90.51	18.67	72.70	88.85	11.04	77.40	90.54	20.46	71.13	NA	NA	NA	84.78	16.18	67.70	89.42	19.12	70.31	89.00	19.42	69.97	10.00	100.00	0.00
VGG-19	SSBA	90.59	5.14	84.94	89.22	1.61	86.31	90.58	3.22	86.24	88.83	1.81	85.30	90.59	5.14	84.94	NA	NA	NA	85.67	5.37	79.29	88.81	5.48	82.98	77.41	36.69	55.39	10.00	0.00	11.11
VGG-19	Input-aware	89.32	75.27	23.63	90.93	48.78	47.70	89.93	38.34	49.16	90.70	59.61	38.09	89.32	75.27	23.63	NA	NA	NA	85.45	50.43	39.43	89.96	59.29	34.87	89.72	63.31	31.04	10.00	100.00	0.00
VGG-19	WaNet	89.52	1.02	88.84	91.31	1.43	89.99	90.20	0.98	89.51	90.81	1.38	89.83	89.52	1.02	88.84	NA	NA	NA	86.06	3.06	82.58	89.26	4.06	85.43	89.00	3.12	85.88	10.00	100.00	0.00
EfficientNet-B3	BadNets	61.36	3.53	60.79	56.46	5.80	55.00	60.94	3.60	60.24	58.14	5.47	56.84	61.35	3.53	60.79	61.35	3.53	60.79	42.45	5.29	42.16	50.22	3.91	49.92	34.72	1.61	36.90	54.70	15.23	50.22
EfficientNet-B3	Blended	54.41	5.49	50.60	51.16	7.17	46.77	54.10	5.01	50.06	52.64	6.29	47.80	54.42	5.48	50.58	54.42	5.48	50.58	45.86	9.27	38.98	49.00	8.64	44.04	30.26	0.27	23.24	52.82	12.37	42.14
EfficientNet-B3	LC	57.65	6.07	43.77	56.20	6.33	40.93	57.88	6.83	43.01	56.08	7.89	40.94	57.65	6.07	43.76	57.65	6.07	43.78	46.15	7.36	38.81	41.07	3.43	36.40	25.01	0.81	22.42	56.53	9.52	42.34
EfficientNet-B3	SIG	53.69	4.07	40.26	51.45	3.33	37.66	53.38	3.29	39.56	51.91	3.26	38.98	53.67	4.08	40.26	53.67	4.08	40.26	43.66	4.76	26.74	53.62	4.41	50.69	27.18	0.59	24.96	55.05	58.19	26.27
EfficientNet-B3	LF	62.50	22.32	44.00	55.76	18.10	41.14	61.76	23.41	43.81	57.21	20.13	41.81	62.50	22.32	44.01	62.52	22.32	44.01	47.03	20.40	29.66	44.03	17.83	32.66	14.38	1.53	25.13	57.24	14.93	49.42
EfficientNet-B3	SSBA	62.61	6.19	58.31	56.29	6.56	53.51	62.34	6.29	58.26	58.88	7.16	54.26	62.62	6.18	58.32	61.36	6.09	56.59	41.70	4.88	40.39	53.43	4.74	51.71	29.12	1.29	29.99	54.79	7.79	50.08
EfficientNet-B3	Input-aware	62.80	9.01	53.57	62.13	9.50	53.40	64.85	6.08	57.52	64.28	9.23	54.77	62.80	9.01	53.57	62.80	9.01	53.57	42.92	5.36	41.09	54.33	4.54	48.51	35.52	2.98	34.94	58.85	16.97	50.23
EfficientNet-B3	WaNet	72.04	3.73	69.36	73.30	2.77	71.49	73.42	4.23	70.14	74.51	2.56	73.60	71.00	3.03	68.84	69.19	3.02	64.59	40.48	9.20	39.01	46.80	3.71	46.77	27.19	4.54	25.96	54.78	12.87	49.86
MobileNetV3-Large	BadNets	83.78	1.93	83.13	79.67	2.24	79.07	83.64	1.79	83.09	80.88	2.07	80.50	78.01	2.58	77.56	83.78	1.93	83.14	72.39	2.98	72.22	80.64	2.03	80.33	72.64	0.72	74.50	57.77	8.57	54.77
MobileNetV3-Large	Blended	84.02	78.30	16.32	79.19	25.30	49.63	83.99	75.30	18.59	80.43	26.61	49.16	84.01	78.29	16.32	84.01	78.29	16.32	73.13	35.01	42.14	80.67	30.88	49.92	74.57	56.41	30.20	55.33	10.30	48.00
MobileNetV3-Large	LC	83.72	5.66	54.42	79.11	5.46	50.53	83.61	5.68	54.26	80.31	5.10	51.68	83.72	5.66	54.41	83.71	5.64	54.42	74.99	5.82	50.22	80.86	5.18	53.07	72.89	3.41	47.27	56.48	10.11	42.41
MobileNetV3-Large	SIG	84.34	67.39	17.06	79.61	53.16	23.07	84.21	72.14	14.34	80.95	57.77	21.02	77.78	60.39	19.56	76.76	59.29	16.39	73.16	70.48	11.76	80.99	60.21	17.57	72.34	84.03	5.22	58.58	89.31	8.53
MobileNetV3-Large	LF	84.06	77.57	18.06	79.35	38.19	42.82	83.96	77.13	17.82	79.77	51.82	34.68	84.06	77.57	18.06	82.30	68.78	23.82	74.17	70.66	19.89	80.05	62.39	28.19	75.47	72.80	17.99	60.19	7.70	56.36
MobileNetV3-Large	SSBA	84.01	4.91	78.27	79.23	3.41	74.07	83.72	3.54	79.16	80.03	3.13	74.99	84.01	4.92	78.26	84.01	4.92	78.26	74.22	5.44	68.37	80.28	5.01	74.37	74.33	5.68	69.80	54.01	13.39	48.90
MobileNetV3-Large	Input-aware	78.70	33.81	42.01	79.63	24.54	49.84	81.40	36.28	39.40	81.33	30.00	47.07	78.70	33.82	42.01	72.17	12.33	57.23	73.69	29.62	47.70	81.08	41.73	42.52	71.25	37.10	39.52	58.27	83.33	12.70
MobileNetV3-Large	WaNet	79.16	3.61	77.58	80.82	2.60	78.97	81.30	4.51	78.19	82.16	2.46	80.09	79.16	3.61	77.58	79.16	3.61	77.58	72.93	3.49	71.32	80.55	2.84	78.23	76.44	1.91	73.31	54.84	14.46	51.24
DenseNet-161	BadNets	86.42	1.80	85.51	85.60	1.82	84.82	86.40	1.24	85.74	84.38	2.06	83.79	82.99	2.10	82.21	86.02	1.52	85.40	76.13	2.51	75.53	83.60	1.80	82.61	57.11	6.19	53.56	64.53	5.07	61.44
DenseNet-161	Blended	86.57	59.88	32.08	85.98	36.13	50.50	86.74	56.77	34.47	84.88	29.73	53.92	86.56	59.90	32.07	83.43	25.11	55.74	75.77	16.21	57.51	83.34	2.04	73.48	61.21	43.96	37.49	70.50	5.22	61.54
DenseNet-161	LC	86.91	4.73	56.97	85.05	4.96	55.59	87.03	5.28	56.59	84.30	4.96	54.51	86.91	4.73	56.98	86.40	3.83	57.21	76.41	5.59	50.68	83.35	4.43	54.84	50.91	5.13	31.11	66.10	10.99	47.53
DenseNet-161	SIG	86.69	61.20	26.18	85.49	63.99	23.58	86.67	67.52	22.21	84.47	48.98	30.41	83.68	42.30	34.60	79.32	22.83	38.37	76.23	41.62	30.27	83.48	52.70	30.58	69.79	0.07	46.86	74.11	95.23	4.21
DenseNet-161	LF	86.27	22.28	65.93	84.99	11.47	73.23	86.46	19.64	68.12	84.22	13.32	71.31	86.27	22.28	65.93	79.57	6.64	71.64	75.42	20.96	57.90	82.98	21.76	63.32	60.11	32.86	41.54	77.90	9.01	71.76
DenseNet-161	SSBA	86.59	4.40	81.60	85.08	3.13	81.32	86.92	2.88	82.91	84.27	3.24	79.80	86.59	4.40	81.60	81.50	2.02	76.52	75.68	5.23	70.58	83.37	2.59	79.32	58.46	4.63	55.72	68.39	4.82	63.90
DenseNet-161	Input-aware	84.56	83.64	14.29	87.38	73.10	23.90	85.02	84.68	14.30	87.19	71.63	24.82	85.84	3.23	75.08	76.28	8.93	62.44	76.49	60.14	27.19	83.28	62.28	28.59	59.73	29.89	31.59	67.97	6.63	57.93
DenseNet-161	WaNet	85.74	1.27	84.71	87.76	1.27	85.77	85.05	2.06	83.43	87.54	1.13	85.44	85.75	1.27	84.69	85.75	1.27	84.69	76.66	2.66	74.91	83.24	1.91	81.01	55.57	17.00	44.24	67.22	4.21	65.47

Table 14: Full results on CIFAR-10 with 0.1% poisoning ratio.

	Backdoor Defense →	No defense			FT			FP			NAD			NC			ANP			AC			Spectral			ABL			DBD		
Targeted Model	Backdoor Attack ↓	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)	C-Acc (%)	ASR (%)	R-Acc (%)
PreAct-Resnet18	BadNets	93.61	1.23	93.24	93.59	1.20	93.20	93.54	0.91	93.24	93.45	1.26	93.03	90.29	1.27	90.08	89.60	0.71	89.37	89.97	1.37	89.57	93.03	1.01	92.53	81.44	1.99	79.69	71.59	4.90	68.82
PreAct-Resnet18	Blended	93.80	56.11	39.37	93.78	52.60	42.09	93.70	54.67	40.38	93.71	55.23	39.76	91.07	26.83	57.16	85.35	24.18	49.57	89.59	41.41	48.77	93.14	51.16	42.93	67.96	12.06	40.82	78.51	0.00	81.22
PreAct-Resnet18	LC	93.82	4.67	66.52	93.63	4.21	66.83	93.52	4.51	66.41	93.78	4.57	66.76	93.82	4.67	66.51	93.82	4.67	66.51	89.95	4.53	60.54	93.19	4.64	65.68	83.20	8.91	50.52	75.43	3.27	57.81
PreAct-Resnet18	SIG	93.73	41.27	42.34	93.64	38.99	44.12	93.53	37.71	43.07	93.75	37.60	45.49	89.34	15.96	42.92	84.72	2.71	41.97	89.62	40.09	35.27	93.21	41.34	41.50	62.37	13.52	22.79	74.78	0.07	21.54
PreAct-Resnet18	LF	93.55	12.72	81.16	93.27	9.63	83.07	93.42	12.48	81.00	93.57	10.51	82.41	89.68	4.57	83.62	88.52	5.63	79.58	89.94	13.26	76.52	92.86	18.39	75.06	54.82	8.07	53.46	74.19	9.50	70.13
PreAct-Resnet18	SSBA	93.89	1.62	90.64	93.86	1.41	90.84	93.72	1.60	90.28	93.83	1.43	90.76	93.89	1.62	90.64	90.21	0.99	86.69	89.97	2.33	86.16	93.39	1.83	90.08	81.00	6.22	76.59	73.42	4.42	67.39
PreAct-Resnet18	Input-aware	91.94	47.32	45.53	93.05	56.61	39.22	92.03	26.13	51.19	93.05	58.94	37.43	91.94	47.32	45.53	86.47	14.83	63.89	89.93	30.92	58.30	93.16	46.93	48.42	76.14	31.19	44.80	82.22	13.18	69.51
PreAct-Resnet18	WaNet	92.18	0.78	91.42	93.49	0.74	92.92	91.88	1.06	91.00	93.44	0.76	92.93	90.71	0.94	89.81	87.53	0.56	88.91	90.07	1.69	88.13	93.04	1.02	91.57	67.90	5.04	61.27	87.83	2.63	86.66
VGG-19	BadNets	90.64	1.22	90.37	89.28	1.10	89.20	90.47	0.98	90.34	88.54	1.14	88.44	90.64	1.22	90.37	NA	NA	NA	85.40	1.72	84.87	89.06	1.51	88.62	89.63	1.04	89.47	10.00	100.00	0.00
VGG-19	Blended	91.00	26.12	61.70	89.18	16.46	66.80	90.77	25.28	62.03	88.78	18.36	65.10	91.00	26.12	61.70	NA	NA	NA	85.60	8.12	67.64	89.39	20.86	63.18	89.71	15.14	67.67	10.00	100.00	0.00
VGG-19	LC	91.07	5.08	62.11	89.37	4.61	60.91	91.00	4.67	62.20	89.39	4.03	60.97	91.07	5.08	62.11	NA	NA	NA	85.42	5.30	54.52	89.49	4.90	59.58	89.64	4.44	59.56	10.00	0.00	11.10
VGG-19	SIG	90.77	22.71	50.34	89.52	19.97	51.57	90.62	22.24	50.29	88.75	18.00	49.10	90.77	22.71	50.34	NA	NA	NA	85.39	21.30	40.37	89.32	25.00	43.90	89.69	21.58	47.57	10.00	0.00	11.11
VGG-19	LF	90.46	5.74	84.02	89.21	5.40	83.52	90.50	5.44	84.10	88.47	4.49	83.29	90.46	5.74	84.02	NA	NA	NA	85.83	7.18	77.36	89.56	6.60	81.44	89.10	6.16	82.03	10.00	100.00	0.00
VGG-19	SSBA	91.06	1.54	87.96	89.31	1.16	87.00	90.86	1.33	87.92	89.45	1.73	86.06	86.73	2.17	83.66	NA	NA	NA	85.78	2.58	82.32	89.20	2.02	86.01	88.96	1.98	85.80	10.00	0.00	11.11
VGG-19	Input-aware	89.52	48.24	43.78	90.87	27.20	58.03	89.85	38.33	48.48	90.84	27.20	57.10	89.52	48.24	43.78	NA	NA	NA	86.01	8.50	71.73	89.81	14.27	70.32	89.73	18.76	63.83	10.00	100.00	0.00
VGG-19	WaNet	89.77	1.60	88.86	91.58	1.12	90.70	90.29	1.09	89.64	91.20	1.11	90.50	89.44	1.54	87.94	NA	NA	NA	85.90	2.17	83.39	88.99	1.77	86.99	89.67	1.47	87.74	10.00	100.00	0.00
EfficientNet-B3	BadNets	63.92	3.44	63.43	57.59	4.47	56.93	63.31	3.76	62.64	60.03	3.87	59.20	63.92	3.43	63.46	63.92	3.43	63.46	41.50	4.68	40.79	53.14	4.52	52.37	31.50	1.01	33.50	52.65	12.21	48.14
EfficientNet-B3	Blended	61.41	6.66	51.91	57.29	6.90	48.29	61.78	6.27	51.48	58.89	7.03	49.31	61.43	6.66	51.90	61.43	6.66	51.90	46.82	9.00	41.33	56.07	5.13	50.88	23.47	2.74	18.94	62.60	6.52	56.54
EfficientNet-B3	LC	54.96	4.99	43.16	52.21	6.53	39.90	54.98	5.74	42.42	52.30	6.42	40.78	54.96	4.98	43.16	54.96	4.98	43.16	39.23	6.28	33.76	54.34	5.40	41.84	20.28	0.19	20.04	54.70	13.16	40.79
EfficientNet-B3	SIG	66.63	41.69	23.90	61.28	17.31	26.60	66.91	35.60	25.38	64.25	9.26	28.34	61.13	6.88	32.34	66.63	41.71	23.89	43.94	1.66	31.40	50.80	3.03	47.77	32.25	15.94	22.28	59.05	41.20	34.32
EfficientNet-B3	LF	55.96	13.57	39.20	52.77	14.09	36.93	55.64	14.18	39.73	53.59	14.94	38.04	55.96	13.56	39.20	55.96	13.56	39.20	43.30	21.13	26.06	47.40	16.67	34.32	23.39	1.76	37.63	54.14	14.26	48.48
EfficientNet-B3	SSBA	65.76	3.72	61.94	59.73	5.24	55.51	65.31	3.93	61.33	62.82	4.94	58.03	65.77	3.72	61.94	65.77	3.72	61.96	40.40	5.16	39.71	47.98	3.44	47.26	29.80	0.37	32.63	58.51	10.97	53.37
EfficientNet-B3	Input-aware	61.93	7.96	54.93	62.14	6.79	55.32	62.89	6.26	56.64	62.69	6.87	55.71	61.21	7.70	53.16	61.93	7.96	54.94	45.17	6.33	41.56	45.58	3.88	43.66	33.34	1.07	35.48	56.16	17.86	45.77
EfficientNet-B3	WaNet	75.14	3.54	72.91	70.98	3.57	69.44	73.46	3.83	71.22	73.89	3.13	71.59	75.13	3.54	72.90	71.12	3.09	69.13	48.40	5.31	47.00	50.38	3.19	50.09	26.57	0.43	26.90	57.88	10.91	54.36
MobileNet V3-Large	BadNets	84.30	1.61	84.03	79.10	2.31	78.83	84.15	1.72	83.87	80.35	2.19	79.64	84.30	1.61	84.03	84.30	1.61	84.03	74.84	2.41	74.23	79.72	1.94	79.60	76.19	2.73	75.21	58.91	7.40	56.29
MobileNet V3-Large	Blended	84.15	34.39	49.36	79.33	3.86	63.48	84.10	31.48	51.29	80.41	4.77	63.83	84.15	34.39	49.36	84.15	34.39	49.36	73.87	3.16	61.93	80.43	11.83	63.74	74.94	6.56	57.49	56.22	10.08	47.90
MobileNet V3-Large	LC	84.19	5.01	55.58	79.58	5.47	52.44	83.88	5.10	55.47	80.70	5.62	52.42	77.30	5.47	49.82	84.19	5.01	55.58	75.12	5.28	49.50	80.62	4.47	53.18	73.81	4.57	46.77	56.01	10.67	40.87
MobileNet V3-Large	SIG	84.22	41.32	29.08	79.51	32.80	32.50	83.88	41.18	28.94	80.21	16.54	39.47	84.22	41.32	29.08	76.25	33.66	27.82	73.00	30.98	24.77	80.67	29.53	32.11	75.35	28.73	26.87	58.69	93.52	4.53
MobileNet V3-Large	LF	84.03	24.63	61.21	79.20	8.70	62.83	83.73	24.02	61.84	80.34	10.61	64.39	84.03	24.63	61.22	83.29	23.39	61.70	73.38	6.51	63.49	80.63	9.37	70.00	74.42	14.71	59.39	56.49	11.22	50.81
MobileNet V3-Large	SSBA	84.42	2.22	81.21	79.63	2.79	75.79	84.33	2.34	80.73	80.47	2.96	76.78	76.14	2.79	73.11	84.43	2.22	81.21	74.58	3.93	70.19	80.93	2.91	76.77	73.59	4.41	70.24	58.69	7.50	53.49
MobileNet V3-Large	Input-aware	76.91	7.79	64.29	80.03	11.14	60.69	82.15	10.91	63.52	81.57	10.86	61.88	77.61	9.90	64.67	76.91	7.79	64.31	72.17	7.63	58.21	81.23	11.22	60.83	73.41	9.84	57.94	58.87	22.50	45.66
MobileNet V3-Large	WaNet	81.64	2.34	80.42	81.32	2.50	79.44	82.51	3.68	79.62	82.96	2.33	81.30	80.69	2.49	78.57	81.64	2.34	80.42	73.22	3.13	71.60	80.43	2.06	78.93	74.06	3.59	69.06	56.35	8.21	51.92
DenseNet-161	BadNets	86.70	1.57	86.07	85.66	1.60	84.99	86.63	1.24	86.46	84.97	1.72	84.20	82.42	2.01	82.02	86.70	1.57	86.07	77.15	2.27	76.44	83.30	1.76	83.03	55.76	9.23	51.93	70.29	4.12	67.93
DenseNet-161	Blended	86.76	14.84	68.23	86.14	7.41	71.77	86.82	12.47	69.93	84.86	9.92	69.96	86.76	14.87	68.22	86.05	5.03	72.32	75.84	4.70	65.69	83.11	8.63	70.07	58.68	8.94	45.44	60.06	6.18	53.56
DenseNet-161	LC	86.48	4.59	57.67	85.52	4.83	55.82	86.68	4.83	57.36	85.02	4.78	55.42	81.39	4.86	52.78	86.48	4.59	57.66	76.42	5.03	51.00	83.08	4.71	54.94	68.04	3.08	45.28	67.35	9.60	48.66
DenseNet-161	SIG	86.69	14.97	53.42	86.20	11.59	53.52	86.79	17.82	51.38	84.99	12.14	54.86	86.69	15.00	53.41	81.57	6.48	51.47	75.87	9.82	47.46	83.22	10.32	54.99	57.21	30.58	28.48	69.11	22.93	52.27
DenseNet-161	LF	86.46	7.19	78.67	85.85	7.56	77.37	86.56	6.74	78.94	85.36	7.53	77.27	86.46	7.20	78.67	86.45	6.90	78.31	76.56	9.86	64.79	83.03	9.09	72.50	54.78	6.47	54.13	67.58	12.28	61.98
DenseNet-161	SSBA	87.01	2.09	83.46	85.71	2.31	82.04	87.14	2.06	83.37	85.47	2.84	81.22	87.01	2.09	83.46	87.15	2.01	83.32	76.53	4.19	71.98	83.40	2.67	79.57	54.10	10.03	49.31	59.30	11.78	53.60
DenseNet-161	Input-aware	84.55	33.69	42.29	87.27	43.66	41.48	85.37	50.64	33.52	87.23	43.29	41.39	85.93	3.84	71.39	77.14	6.58	55.92	76.86	17.74	51.07	82.63	25.60	50.04	57.83	16.99	28.34	72.30	13.60	51.52
DenseNet-161	WaNet	86.48	1.76	85.09	88.24	1.18	86.53	85.85	1.77	83.87	87.99	1.18	86.60	86.46	1.76	85.09	82.23	1.41	81.19	76.03	2.34	73.56	83.25	1.53	81.46	56.87	2.38	48.59	67.75	5.66	65.30

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing*, 2019.

[2] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *The AAAI Conference on Artificial Intelligence Workshop*, 2019.

[3] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[6] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2019.

[7] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *International Conference on Learning Representations*, 2022.

[8] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2020.

[9] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 2021.

[10] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[11] Jinlong Liu, Guoqing Jiang, Yunzhi Bai, Ting Chen, and Huayan Wang. Understanding why neural networks generalize well through gsnr of parameters. *arXiv preprint arXiv:2001.07384*, 2020.

[12] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, 2018.

[13] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017.

[14] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 2020.

[15] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021.

[16] Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. ONION: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[17] Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[18] Fanchao Qi, Yuan Yao, Sophia Xu, Zhiyuan Liu, and Maosong Sun. Turn the combination lock: Learnable textual backdoor attacks via word substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[20] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 2018.

[21] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[22] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

[23] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *7th International Conference on Learning Representations*, 2019.

[24] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems Workshop*, 2018.

[25] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

[26] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy*, 2019.

[27] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 2021.

[28] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[29] Qing T Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6(1):1–9, 2006.

[30] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.