SPD: Shallow Backdoor Protecting Deep Backdoor Against Backdoor Detection

Supplementary Material

A. Appendix Section

In the supplementary materials, we present additional results from the SPD evaluation. We begin by detailing the training process of SPD, followed by an analysis of how network architecture influences its effectiveness. Next, we examine SPD's attack performance under the all-to-all setting and assess its robustness against backdoor mitigation defenses. Finally, we examine the impact of α on the stealthiness of SPD through detailed image analysis.

Table 1. Details of the datasets leveraged in the SPD evaluation.

Dataset	Labels	Image size	Training set	Test set	Model
CIFAR-10	10	$32 \times 32 \times 3$	50000	10000	ResNet-18
GTSRB	43	$32 \times 32 \times 3$	39209	12630	ResNet-18
ImageNet-12	12	$224 \times 224 \times 3$	12406	3120	ResNet-34

A.1. Training details

We select three benchmark datasets, including CIFAR-10, GTSRB, and ImageNet-12 (a subset of ImageNet [1]), to evaluate SPD and the baseline backdoor attacks. The details of these datasets are summarized in Table 1. For CIFAR-10 and GTSRB, we inject backdoors into ResNet-18 [2] models, and for ImageNet-12, we use a ResNet-34 [2] model. The hyperparameter α is set to 5.0 for CIFAR-10, 20.0 for GTSRB, and 5.0 for ImageNet-12. The optimizer is SGD with momentum 0.9, with batch sizes of 128 for CIFAR-10, 32 for GTSRB, and 16 for ImageNet-12. The learning rate η is 0.01, and weight decay is 0.0005. Both training epochs T_c and T_b are set to 100, where the MultiStepLR decay strategy with milestones at [30, 60, 90] is deployed, and decay factor γ is set to 0.1. The target label of SPD is 0. The trigger size in the shallow backdoor is 2 for CIFAR-10 and GTSRB, and 8 for ImageNet-12.

Table 2. Influence of the network structure on the attack performance of SPD. ASR: Attack Success Rate (%); CA: Clean Accuracy (%). **The best results are boldfaced.**

Dataset	Metric	ResNet-18	VGG-16	MobileNet-V2
CIFAR-10	ASR	99.91	99.93	**99.98**
	CA	**94.38**	90.49	90.69
GTSRB	ASR	**99.98**	99.91	99.68
	CA	**98.57**	96.09	96.19

A.2. Influence of network structure

To evaluate the influence of different network architecture on the performance of SPD, we select three different networks, including ResNet-18 [2], MobileNet-V2 [4], and VGG-16 [5].

Table 2 illustrates the impact of different network structures on the attack performance of SPD on CIFAR-10 and GTSRB. The results show that the ASRs for all network structures are close to 100%, indicating that SPD achieves extremely high attack performance across these networks. Meanwhile, the CA varies with different network structures: on CIFAR-10, ResNet-18 achieves the highest CA (94.38%), followed by MobileNet-V2 (90.69%), while VGG-16 has a relatively lower CA (90.49%). The evaluate results demonstrate that the network structures do not affect the attack performance of SPD.

Table 3. Attack effectiveness of SPD and the baseline backdoor attacks on CIFAR-10, GTSRB under the all-to-all setting. ASR: Attack Success Rate (%); CA: Clean Accuracy (%). **The best results are boldfaced.**

Dataset	Metric	BadNets	Blended	Bpp	IAD	WaNet	Ftrojann	SIG	Refool	SSBA	SPD
CIFAR-10	CA	93.92	93.66	91.98	91.14	91.70	93.65	93.68	92.68	93.57	**94.32**
	ASR	89.71	86.24	88.23	87.53	88.84	92.72	90.54	82.86	88.45	**95.20**
GTSRB	CA	**99.34**	99.22	98.85	98.14	98.90	99.11	99.20	97.52	98.86	98.36
	ASR	98.06	97.85	96.18	95.45	96.29	**99.02**	98.85	79.28	97.10	98.28

A.3. Attack performance with all-to-all setting

In backdoor attacks, common attack settings include all-to-one and all-to-all. The all-to-one setting means that the target label in the backdoor attack is fixed, and all samples with triggers will be classified into the target label. In contrast, the all-to-all setting means that the target label in the backdoor attack is the next label of the ground-truth label of the backdoor sample. Specifically, the target label t is set as:

$$t = (y + 1) \mod \mathcal{N} \tag{1}$$

where y is the ground-truth label of the backdoor sample, and $\mathcal{N}$ is the number of classes. In this paper, we have demonstrated the effectiveness of the SPD attack under the all-to-one setting. Here, we examine the performance of SPD under the all-to-all setting on CIFAR-10 and GTSRB.

In the all-to-all attack setting, the model must simultaneously consider both the trigger features and the sample features to activate the target class. In contrast to the all-to-one setting, where the model relies solely on the trigger features to activate the target class, the all-to-all setting increases the complexity of backdoor attacks, which in turn reduces the ASRs of the attacks. Table 3 presents the evaluation results of these backdoor attacks. On CIFAR-10, SPD

Table 4. Resilience of SPD and the baseline backdoor attacks against existing backdoor mitigation methods on CIFAR-10 and GTSRB. ASR: Attack Success Rate (%); CA: Clean Accuracy (%). **The best results are boldfaced.**

Dataset	Attack →	BadNets		Blended		Bpp		IAD		WaNet		Ftrojann		SIG		Refool		SSBA		SPD	
	Defense ↓	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA
CIFAR-10	ANP	0.00	87.06	**21.53**	86.75	0.84	85.78	0.82	91.13	0.57	91.61	0.15	4.61	0.01	86.15	2.42	91.62	0.81	85.14	0.36	85.62
	I-BAU	85.43	85.40	73.63	77.55	86.68	**89.80**	21.62	88.80	3.46	88.51	95.54	85.34	99.50	83.35	26.80	85.04	55.16	85.08	0.32	88.81
	SAU	2.77	88.20	11.33	87.15	4.88	88.76	3.13	89.67	1.72	88.90	0.07	90.18	0.06	89.28	1.21	87.47	1.31	88.98	**24.44**	**90.95**
	FT-SAM	0.62	90.94	74.37	91.39	01.44	92.31	0.95	91.99	1.67	92.15	1.78	91.35	33.60	91.33	9.25	91.13	03.96	91.18	1.91	**92.68**
	NC	1.04	91.92	1.66	90.48	2.58	92.04	2.47	88.68	76.97	90.92	1.21	91.63	**100.00**	93.22	93.68	92.08	2.44	80.83	99.90	**93.43**
GTSRB	ANP	0.02	98.74	5.68	94.06	0.00	98.93	0.00	98.02	0.00	97.49	0.01	98.09	0.00	96.89	**91.75**	97.05	48.20	98.25	0.00	90.09
	I-BAU	14.39	47.98	**100.00**	10.29	2.05	93.22	3.90	**93.41**	0.42	94.42	0.00	47.74	**100.00**	65.66	1.95	94.39	89.61	97.52	70.42	91.67
	SAU	0.00	94.52	18.40	93.83	0.03	97.75	0.00	94.48	0.05	96.84	0.02	97.34	0.00	96.38	22.17	97.74	0.00	97.99	**23.56**	98.09
	FT-SAM	0.03	98.18	74.39	98.37	0.08	98.77	0.96	98.99	0.03	99.31	0.00	98.26	98.16	98.28	62.68	98.28	99.69	98.06	0.05	98.85
	NC	0.04	98.37	0.08	97.22	0.00	98.73	93.49	6.41	0.08	98.81	**100.00**	98.42	0.30	96.24	33.31	97.14	0.00	98.51	99.88	98.64

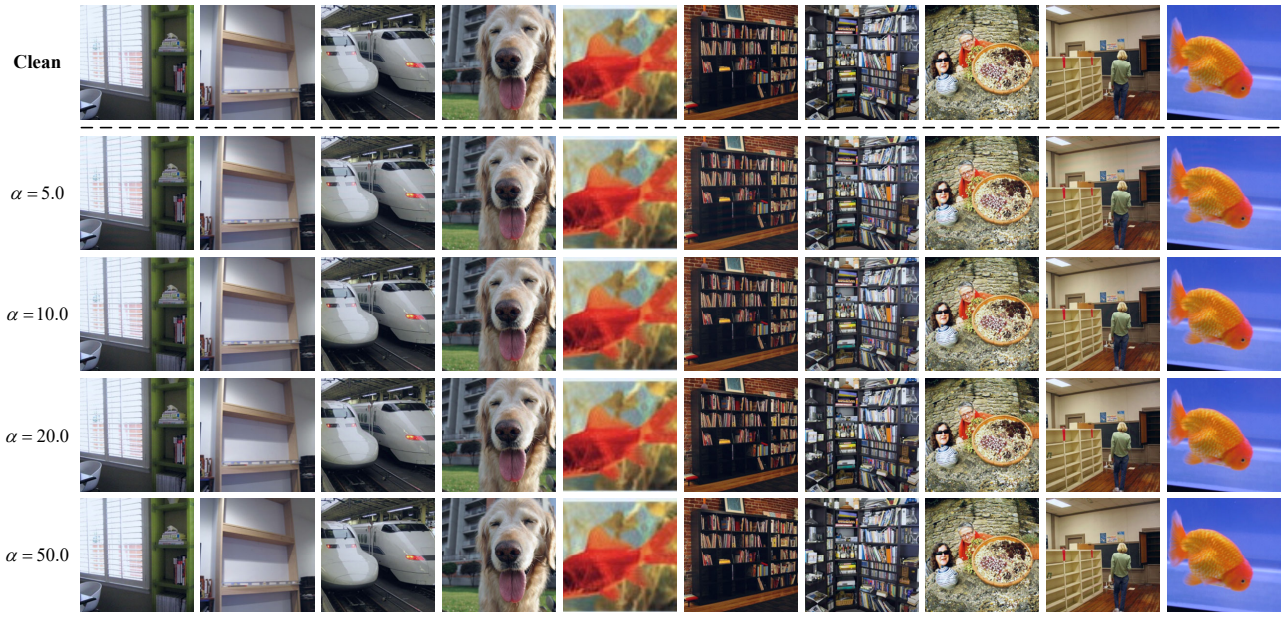

Figure 1. Backdoor samples of SPD with different α on ImageNet-12.

achieves the highest ASR and CA, with values of 95.20% and 94.32%, respectively. On GTSRB, Ftrojann achieves the highest ASR at 99.02%, while BadNets yields the best CA 99.34%. The experimental results demonstrate that, compared to other backdoor attacks, the all-to-all attack setting has the least impact on the ASR of SPD.

A.4. Resilience against backdoor mitigation

To evaluate the resistance of SPD against existing backdoor mitigation techniques, we select several state-of-the-art backdoor mitigation methods including NC [6], I-BAU [10], ANP [9], FT-SAM [11], SAU [7]. Note that these backdoor defense methods are implemented with the open-source codes in BackdoorBench [8] or their official open-source codes using their default settings. Following the established work [3], in these backdoor mitigation methods, 1% of benign samples are provided to repair the backdoor models.

Table 4 demonstrates the resilience of SPD and the baseline backdoors against existing backdoor mitigation methods on CIFAR-10 and GTSRB. We observe that NC cannot mitigate the SPD backdoor attack, i.e., the ASR of the repaired model on CIFAR-10 and GTSRB are 99.90% and 99.88%, respectively. However, other backdoor mitigation methods can effectively reduce the ASR of SPD. For instance, on CIFAR-10, ANP reduces the ASR of SPD to nearly 0%, and FT-SAM reduces it to 0.05%. Additionally, we note that while some backdoor attacks can effectively resist certain mitigation methods, such as NC being unable to mitigate the SIG backdoor on CIFAR-10, no single backdoor attack can completely bypass all backdoor mitigation methods.

Moreover, the premise of backdoor mitigation is that a backdoor has been detected in the model. If a model is repaired without knowing whether it contains a backdoor, its generalization ability will inevitably be reduced. Therefore,

if a backdoor attack can evade detection, it may escape repair and subsequently launch a successful attack.

A.5. Discussion details about α

In the paper, we have quantified the impact of α on the stealthiness of SPD attacks using PSNR and SSIM. Here, we visually demonstrate the impact of α on the stealthiness of SPD attacks through the image details. We conduct SPD backdoor attacks on several ResNet-34 models using ImageNet-12, with α set to 5.0, 10.0, 20.0, and 50.0, respectively. Figure 1 shows the backdoor samples of SPD corresponding to different α. In general, the difference between the backdoor samples and clean samples is minimal, especially when α is large. However, when α is small, such as $\alpha = 5.0$, some backdoor samples still exhibit slight perturbations compared to clean samples, particularly in the lighter background areas. And when the content of the samples becomes more complex, these slight perturbations also become difficult to perceive.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, Miami, Florida, USA, 2009. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, Las Vegas, USA, 2016. 1

[3] Yige Li, Xixiang Lyu, Xingjun Ma, Nodens Koren, Lingjuan Lyu, Bo Li, and Yu-Gang Jiang. Reconstructive neuron pruning for backdoor defense. In *ICML*, pages 19837–19854, Honolulu, Hawaii, USA, 2023. 2

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, Salt Lake City, USA, 2018. 1

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, San Diego, USA, 2015. 1

[6] 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 *SP*, pages 707–723, San Francisco, USA, 2019. 2

[7] Shaokui Wei, Mingda Zhang, Hongyuan Zha, and Baoyuan Wu. Shared adversarial unlearning: Backdoor mitigation by unlearning shared adversarial examples. In *NeurIPS*, New Orleans, USA, 2023. 2

[8] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. In *NeurIPS*, pages 10546–10559, New Orleans, USA, 2022. 2

[9] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *NeurIPS*, pages 16913–16925, Virtual Event, 2021. 2

[10] Yi Zeng, Si Chen, Won Park, Z Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *ICLR*, Virtual Event, 2022. 2

[11] Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *ICCV*, pages 4443–4454, Paris, France, 2023. 2