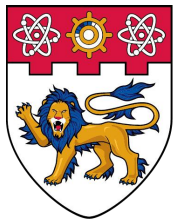
TUE-PM-384

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Color Backdoor: A Robust Poisoning Attack in Color Space

Wenbo Jiang[1], Hongwei Li[1], Guowen Xu[2] and Tianwei Zhang[2]

[1]University of Electronic Science and Technology of China, China
[2]Nanyang Technological University, Singapore



NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Introduction

Gu et al. presented the first backdoor attack (BadNets) against DNN models. They adopted pixel patches as the trigger to activate the backdoor in the model, where the malicious samples look suspicious, and can be easily recognized by humans.

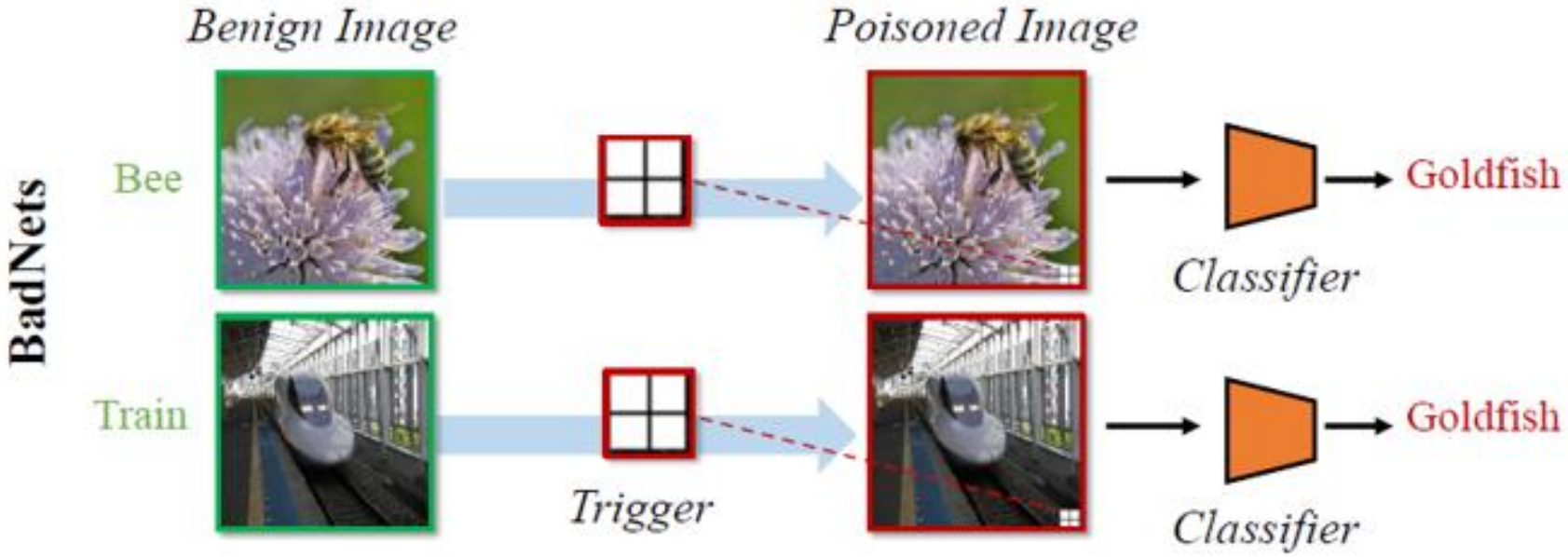
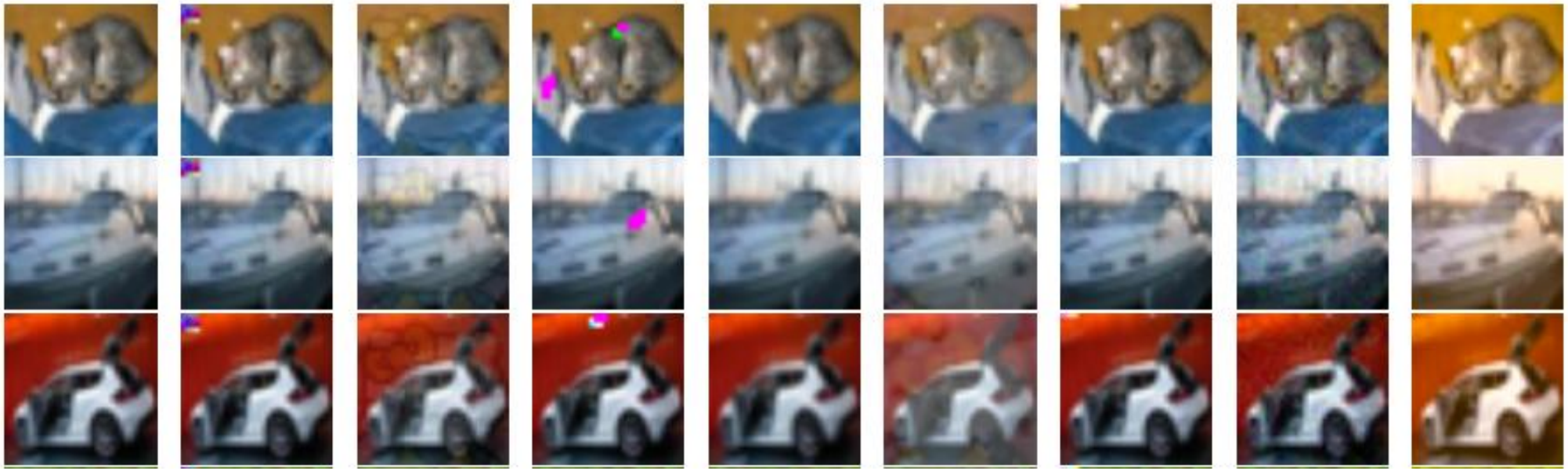


Image from "Invisible Backdoor Attack with Sample-Specific Triggers" (ICCV 2021)

Introduction

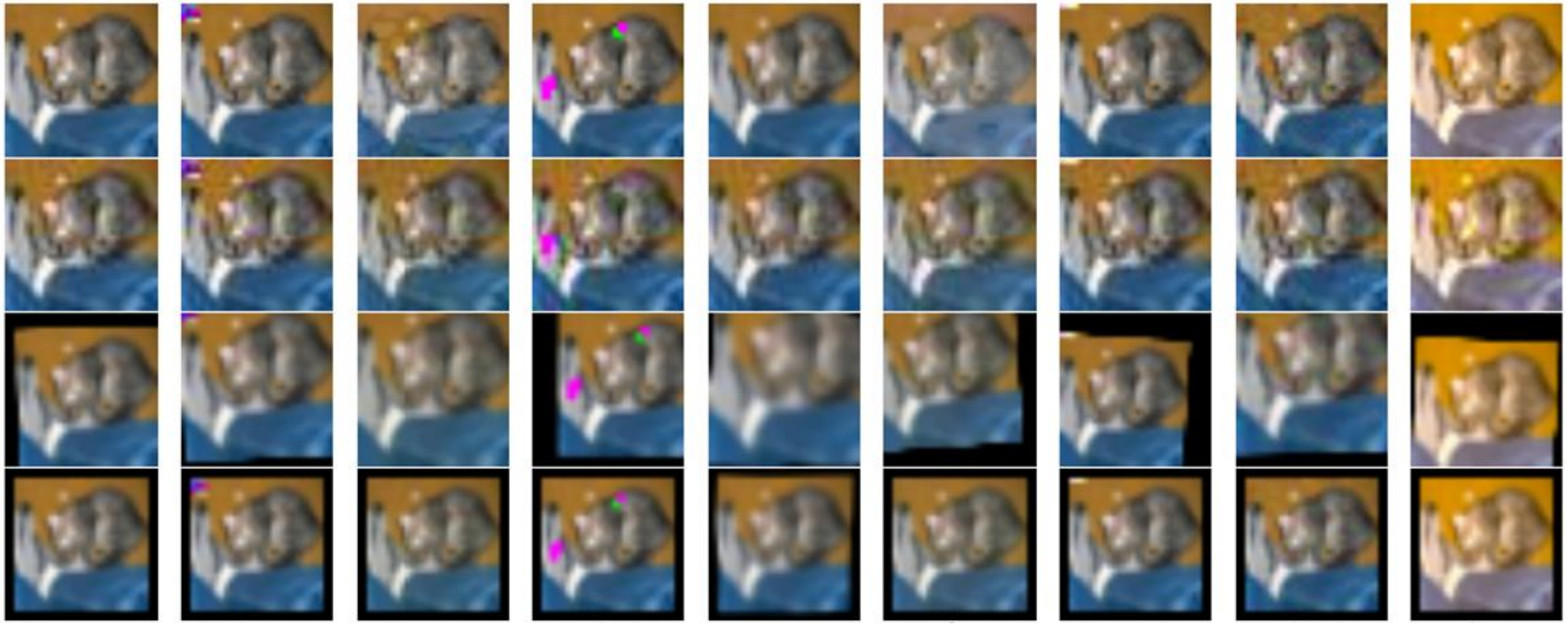
After that, many backdoor attacks are proposed to improve the stealthiness



Visual examples of different types of triggered images. (a) Original images, (b) BadNet, (c) Blend, (d) Input-aware, (e) WaNet, (f) Refool, (g) L0-norm, (h) L2-norm, (i) color backdoor

Introduction

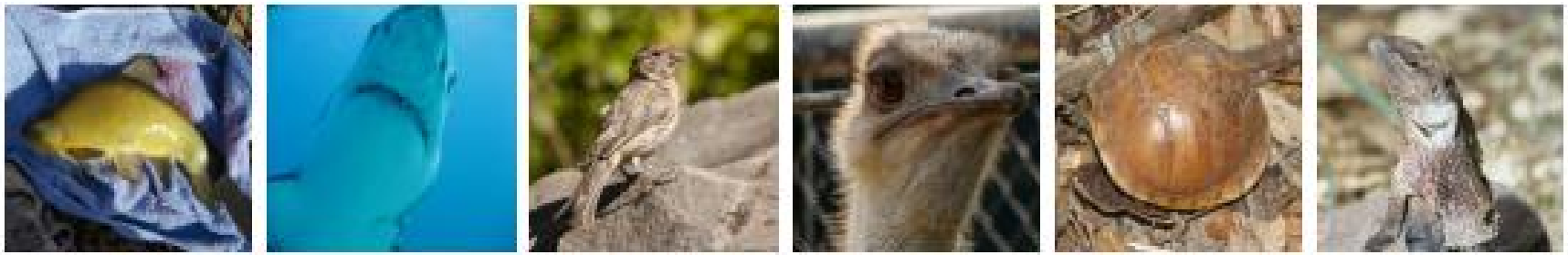
However, they focus on stealthiness while ignoring the backdoor robustness requirement. They become less effective under pre-processing methods



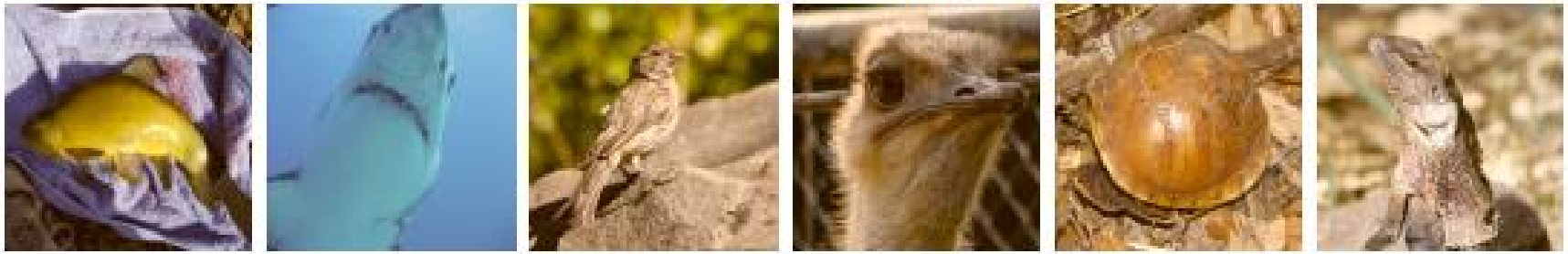
The first row: the triggered images without pre-processing;
The second row: the triggered images after image compression;
The third row: the triggered images after DeepSweep;
The fourth row: the triggered images after ShrinkPad

Methodology

Color backdoor: employs a uniform color space shift for all pixels as the backdoor trigger.



(a) Original images

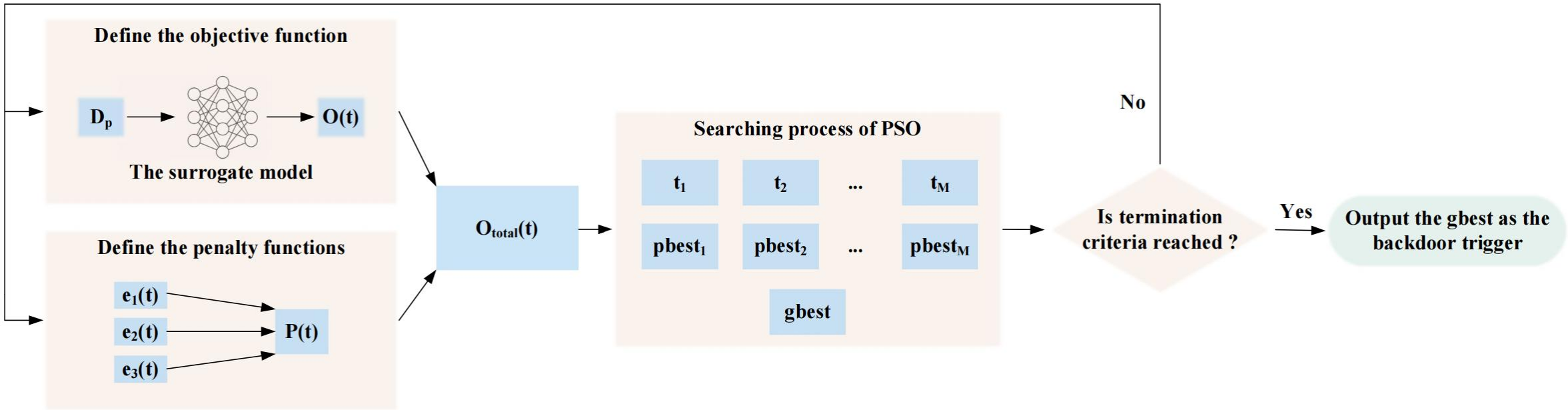


(b) Triggered images of color backdoor

Methodology

Finding an appropriate trigger for color backdoor is non-trivial: a large shift makes the triggered samples less realistic, while a small shift makes it difficult for the model to learn this feature, resulting in low effectiveness and robustness.

Solution: We employ the gradient-free PSO algorithm to find the optimal trigger t (i.e., color space shift) for color backdoor



Methodology

1.Estimating the effectiveness of the trigger through the backdoor loss of a surrogate model

$$O(t) = \mathcal{L}_b = \sum_{x \in D_p} \texttt{CE}(f_s(x+t), y_t)$$

A smaller backdoor training loss indicates the trigger is easier to be learned by the surrogate model, and the attack is more effective.

2. Enforcing the naturalness of the trigger through the penalty functions

$$e_1(t) = \max(0, \lambda_1 - \texttt{PSNR}(t, S))$$
$$e_2(t) = \max(0, \lambda_2 - \texttt{SSIM}(t, S))$$
$$e_3(t) = \max(0, \texttt{LPIPS}(t, S) - \lambda_3)$$

After that, we add the total penalty term to the objective function of the PSO:

$$O_{total}(t) = O(t) + \sum_{j=1}^{3} w_j e_j$$

Methodology

3. Searching for the optimal trigger through PSO

Algorithm 1 Searching the optimal trigger

Require: acceleration factors c_1, c_2; random numbers r_1, r_2; inertia weight ω; number of iteration T; number of particles in the swarm M

Ensure: the optimal trigger for color backdoor

Initialization process:
for each particle $i = 1$ to M **do**
    Randomly initialize the particle position t_i and particle velocity v_i
    Calculate $O_{total}(t_i)$ using Equation 5.
    Initialize $pbest_i$: $pbest_i \leftarrow t_i$
end for
Initialize $gbest$: $gbest \leftarrow \arg\min_{t_i} O_{total}(t_i)$
Searching process:
for $j = 1$ to T **do**
    for each particle $i = 1$ to M **do**
        $v_i \leftarrow \omega v_i + c_1 r_1 (pbest_i - t_i) + c_2 r_2 (gbest - t_i)$
        $t_i \leftarrow t_i + v_i$
        Calculate $O_{total}(t_i)$ using Equation 5.
        $pbest_i \leftarrow t_i$, if t_i is superior to $pbest_i$ according to the defined rule
        $gbest \leftarrow t_i$, if t_i is superior to $gbest$ according to the defined rule
    end for
end for
return $gbest$

EXPERIMENTS

Evaluations on PSO:

Table 1. ASR of the color backdoor attacks with different trigger search optimization algorithms

Method \ Dataset	CIFAR-10	CIFAR-100	GTSRB	ImageNet
PSO	97.55	96.27	99.70	98.16
GA	95.90	96.41	98.87	99.27
Grid-search	98.17	98.01	99.24	99.39
Random	92.02	83.54	91.33	87.09

Table 2. Trigger searching hours of different algorithms

Method \ Dataset	CIFAR-10	CIFAR-100	GTSRB	ImageNet
PSO	1.79 h	3.71 h	1.81 h	3.79 h
GA	3.22 h	6.30 h	3.17 h	6.89 h
Grid-search	5.33 h	10.97 h	5.43 h	11.68 h
Random	-	-	-	-

PSO is superior over other methods

EXPERIMENTS

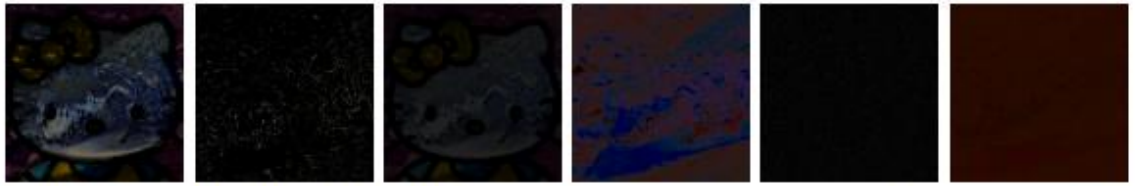
Naturalness Evaluation:

(a) Original images

(b) Triggered images within naturalness restriction

(c) Triggered images without naturalness restriction

(a) Triggered images



(b) The magnified ($\times 1.5$) difference between the triggered images and the original images

Different columns represent different invisible backdoor methods: (a) Refool, (b) WaNet, (c) Blend, (d) Filter backdoor, (e) L2-norm, (f) Color backdoor.

The difference between the original images and our triggered image is a global shift in color space, which is imperceptible to the defender who has no knowledge of the original image.

EXPERIMENTS

Preprocessing-based Defenses:

Table 4. Robustness against preprocessing-based defenses (CIFAR-10).

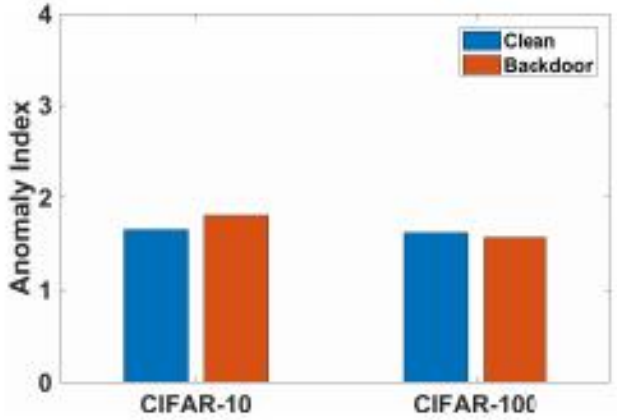
Attack \ Defense	No defense		DeepSweep		ShrinkPad		Compression		Average
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ASR
BadNet	89.20	**99.98**	84.57	54.64	85.74	75.20	81.15	41.56	67.85
Blend	90.16	96.03	85.98	53.20	86.96	17.25	81.36	16.72	45.80
Input-aware	94.39	98.79	91.59	42.04	88.07	32.69	81.71	49.72	55.81
WaNet	91.92	96.14	90.21	45.66	87.81	57.13	84.15	13.05	53.00
Refool	88.66	92.47	82.65	86.37	85.53	93.51	81.60	44.57	79.23
L_0-norm	87.35	77.63	84.38	19.89	83.18	43.30	80.09	35.06	43.97
L_2-norm	90.19	99.86	85.93	15.73	86.71	12.21	84.15	9.23	34.26
Filter	89.91	99.14	83.64	85.56	85.90	92.57	82.95	23.16	75.11
color backdoor	89.77	97.55	85.50	**87.64**	86.15	**93.61**	81.78	**96.89**	**93.92**

Color backdoor is more robust agaisnt preprocessing-based defenses

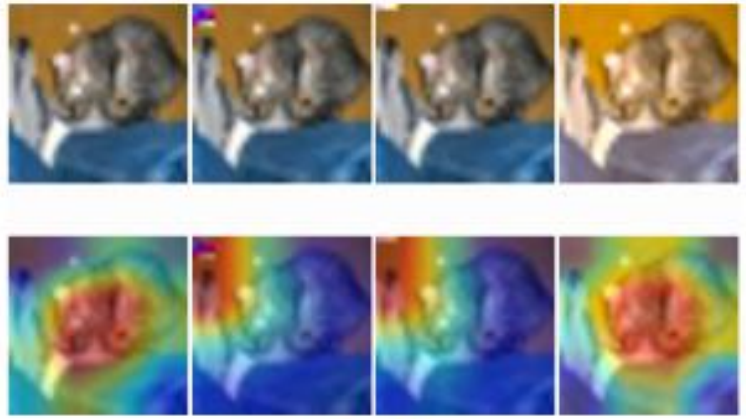
(Other datasets give the same conclusions)

EXPERIMENTS

Other Mainstream Backdoor Defenses



(a) Neural Cleanse



(b) Grad-Cam

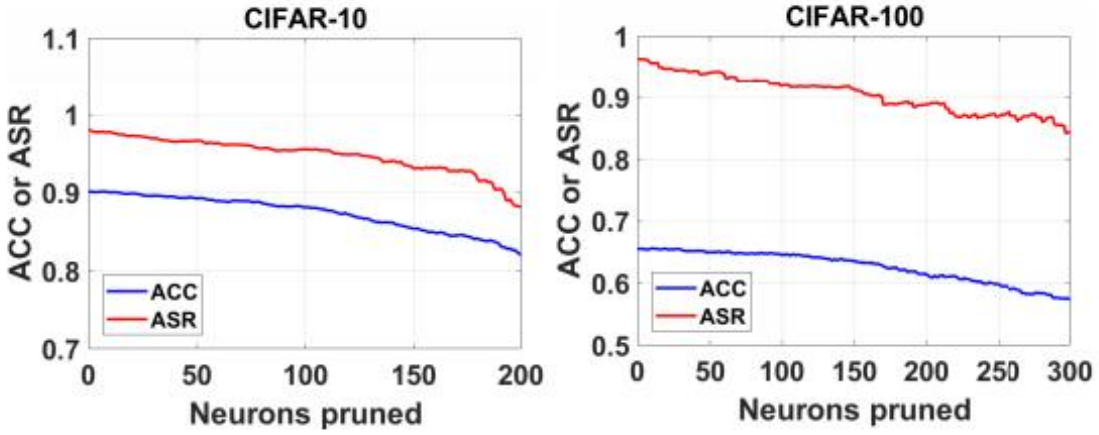


Figure 7. Fine-pruning.

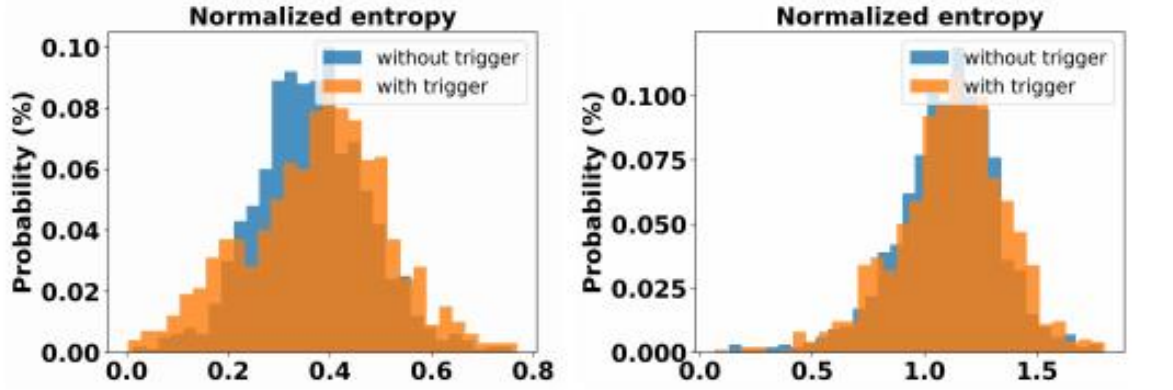


Figure 8. STRIP for CIFAR-10 (left) and CIFAR-100 (right).

CONCLUSION

1. We propose a robust backdoor, which employs a uniform color space shift for all pixels as the trigger. The triggered images maintain natural-looking and can bypass the inspection of the defender.

2. The Particle Swarm Optimization algorithm is employed to optimize the trigger to achieve a robust and stealthy backdoor attack.

3. Extensive experiments demonstrate the superiority of PSO and the robustness of our color backdoor attack against preprocessing-based defenses as well as other mainstream backdoor defenses.

Thanks!