Appendix for Color Backdoor: A Robust Poisoning Attack in Color Space

Anonymous CVPR submission

Paper ID 6455

1. Details of Attack Configuration

The hyperparameter settings of PSO are as follows: the number of the particle M is set to 200 and iteration T is set to 20. Each dimension of the particle position is randomly initialized in the range of (0, 0.5) and each dimension of the particle velocity is randomly initialized in the range of (-0.05, 0.05). ω is set to 0.1 and $c_{1,2}$ are set to 2.

Besides, the similarity thresholds (i.e., $\lambda_{1,2,3}$) for PSNR, SSIM and LPIPS are set to 0.9, 20 and 0.02, respectively. The AlexNet, ResNet-18 and VGG11 are used as the architecture of the surrogate model for ResNet-18, ResNet-34 and VGG16, respectively. The surrogate model is trained for 5 epochs to obtain the backdoor loss.

2. Effectiveness Evaluations on Other Color Spaces

We present the experimental results of color backdoor on all considered color spaces (RGB, HSV, LAB, YCbCr, XYZ, LUV) in Table 1. We observe that the ASRs of all considered color spaces is very high, demonstrating the high attack effectiveness of color backdoor. Besides, the embedded backdoor has very negligible impact on the normal behaviors of the model, as its ACC is very close to the clean one. This is because the poisoned data only account for a small fraction (5% in this work) of the training dataset.

3. Additional Results for Robustness Evaluations

We present the robustness evaluations of backdoor attacks on the GTSRB dataset and the CIFAR-100 dataset against preprocessing-based defenses in Table 2 and 3. The additional experimental results confirm that our attack is more robust than state-of-the-art backdoor attacks against preprocessing-based defenses on these datasets.

4. Details of the searching process of Genetic Algorithm

The details of the searching process of Genetic Algorithm (GA) are illustrated in Figure 1. Similar as the PSO algorithm, GA also maintains a set of individuals, where each individual is a candidate trigger for color backdoor. Firstly, GA randomly generates M individuals in the searching space. Then, the fitness value of each individual is calculated to measure the quality of the individual. After that, the individuals with better fitness values will be selected and their genes are passed to the next generation. The crossover operation creates offspring by swapping parts of the chromosomes of two selected individuals. The mutation operation makes random changes to the genes of the newly created individual. Finally, after T rounds of iteration, the best individual in the final population is returned as the optimal trigger for color backdoor.

Table 1. The effectiveness of color backdoor attack in different color spaces.

Color space	CIFAR-10		CIFAR-100		GTSRB		ImageNet	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
No attack	90.05	-	66.86	-	93.33	-	71.67	-
RGB	89.95	90.32	65.13	92.83	93.02	99.34	68.98	93.44
HSV	89.47	94.26	65.67	93.15	93.17	96.04	69.01	91.02
LAB	89.69	97.44	65.40	96.74	93.25	97.94	68.50	93.43
YCbCr	90.14	93.50	66.36	82.95	93.46	97.22	68.91	96.71
XYZ	89.82	99.16	66.16	97.84	92.39	96.68	68.72	98.08
LUV	89.77	97.55	65.86	96.27	93.36	99.70	69.11	98.16

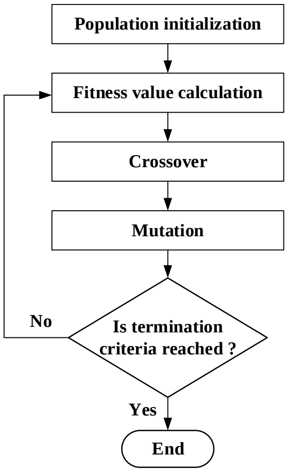

Population initialization
Fitness value calculation
Crossover
Mutation
Is termination criteria reached ?
No
Yes
End

Figure 1. The process of Genetic Algorithm.

Table 2. Robustness against preprocessing-based defenses (CIFAR-100)

Defense	No defense		DeepSweep		ShrinkPad		Compression		Average
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ASR
BadNet	64.81	**98.88**	59.11	32.78	59.11	78.78	54.52	25.32	58.94
Blend	66.46	92.81	59.32	67.76	61.05	47.73	54.11	18.84	56.79
Input-aware	64.41	96.73	64.28	33.28	60.62	84.57	50.79	51.81	66.60
WaNet	65.69	97.09	64.36	30.73	62.23	15.77	49.68	10.31	38.48
Refool	66.00	88.81	60.33	69.91	60.98	86.32	56.11	47.51	73.14
L_0-norm	64.53	32.10	56.42	11.95	59.08	35.64	54.79	14.89	22.65
L_2-norm	66.06	99.03	58.91	24.65	61.13	3.14	55.34	0.96	31.95
Filter	65.77	98.83	59.03	81.11	60.15	90.07	53.29	31.87	75.74
color backdoor	65.86	96.27	58.85	**81.52**	60.90	**92.15**	53.55	**95.39**	**91.33**

Table 3. Robustness against preprocessing-based defenses (GTSRB)

Defense	No defense		DeepSweep		ShrinkPad		Compression		Average
Attack	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ASR
BadNet	93.38	**100.00**	90.67	65.59	90.25	**99.99**	88.66	93.38	89.74
Blend	91.72	91.79	89.12	75.34	88.42	87.64	84.09	71.29	81.52
Input-aware	95.01	91.22	92.19	7.71	93.10	42.47	91.05	32.04	43.46
WaNet	95.59	89.86	94.92	27.49	94.32	99.06	93.16	8.66	56.27
Refool	90.06	92.28	74.39	87.00	86.31	88.89	84.66	87.94	89.02
L_0-norm	91.68	85.53	89.24	15.57	87.10	62.31	87.19	57.88	55.32
L_2-norm	93.26	100.00	90.66	32.31	89.62	9.57	89.12	0.19	35.52
Filter	92.97	99.17	90.81	87.93	90.18	98.13	91.01	39.57	81.20
color backdoor	93.36	99.70	90.06	**94.03**	90.61	99.29	89.61	**99.13**	**98.04**

5. Attack Performance on Gray Image Datasets

It is worthwhile mentioning that our color backdoor can also be applied to gray images. Concretely, the color space of a grayscale image can be considered as containing only one element (i.e., brightness). We generate the backdoor triggered image by applying a small but uniform shift to this element. The difference between the triggered image and the original image is imperceptible by human eyes.

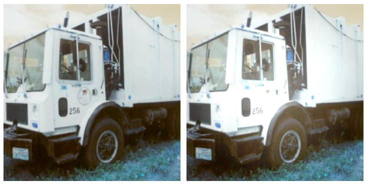
Specifically, we conduct experiments on the MNIST dataset and the FashionMNIST dataset with a simple CNN. The experimental results in Table 4 demonstrate the effectiveness of our attack on gray images.

Table 4. The attack performance of color backdoor on gray image datasets

MNIST		FashionMNIST	
ACC	ASR	ACC	ASR
99.14	99.31	91.26	99.60

6. Attack Performance in Physical World

Our proposed attack can also be conducted in the physical world. For instance, the attacker can use common lights to shine on target objects to create backdoor samples. For instance, we use the flashlight from iPhone 13 to serve as the backdoor trigger (see Figure 2) and train a backdoor model (ImageNet dataset with ResNet-34). The attack success rate is higher than 98%.



(a) clean image (b) backdoor image

Figure 2. Attack in the physical world.