BadCLIP: Trigger-Aware Prompt Learning for Backdoor Attacks on CLIP (Appendix)

Jiawang Bai[1*], Kuofeng Gao[1*], Shaobo Min[2], Shu-Tao Xia[1,3†], Zhifeng Li[2†], Wei Liu[2†]

[1]*Tsinghua University*, [2]*Tencent Data Platform*,
[3]*Research Center of Artificial Intelligence, Peng Cheng Laboratory*

{bjw19,gkf21}@mails.tsinghua.edu.cn, bobmin@tencent.com
xiast@sz.tsinghua.edu.cn, michaelzfli@tencent.com, wl2223@columbia.edu

Table I. Results of BadCLIP under various settings. We report the harmonic mean of results on the seen and unseen classes.

(a) Varying the context length.

Dataset	Metric	Context Length		
		4	8	16
Caltech101	ACC	95.58	95.72	95.66
	ASR	99.46	99.58	99.25
StanfordCars	ACC	71.34	70.88	71.56
	ASR	99.80	99.58	99.83
UCF101	ACC	76.31	76.90	76.67
	ASR	99.62	99.57	99.90

(b) Varying the number of training data.

Dataset	Metric	# of Labeled Training Examples per Class				
		1	2	4	8	16
Caltech101	ACC	91.05	94.76	95.13	95.58	95.58
	ASR	88.25	96.42	98.76	98.97	99.46
StanfordCars	ACC	67.21	68.68	69.53	70.47	71.36
	ASR	98.01	98.90	98.83	99.63	99.80
UCF101	ACC	70.95	71.94	74.23	75.45	76.36
	ASR	94.77	97.88	98.66	99.40	99.62

(c) Using ResNet-50 as the image encoder's backbone.

Dataset	CLIP	CoOp	CoCoOp	**BadCLIP**	
	ACC	ACC	ACC	ACC	ASR
Caltech101	90.80	89.29	92.67	92.04	99.46
StanfordCars	60.50	57.44	64.25	62.78	99.83
UCF101	69.14	52.59	70.80	69.30	99.71

A. Results under Various Settings

In this part, we investigate the effect of various settings on the proposed BadCLIP, including context length, number of training examples, and image encoder's backbone.

Context length. Following [3, 4], we study the performance of BadCLIP when the context length N is set as 4, 8, and 16. The results in Table Ia show that the differences between different context lengths are fairly small and the best choice depends on the dataset. Notably, the ASR values are higher than 99% in all cases, showing the robustness of BadCLIP to various settings of context length.

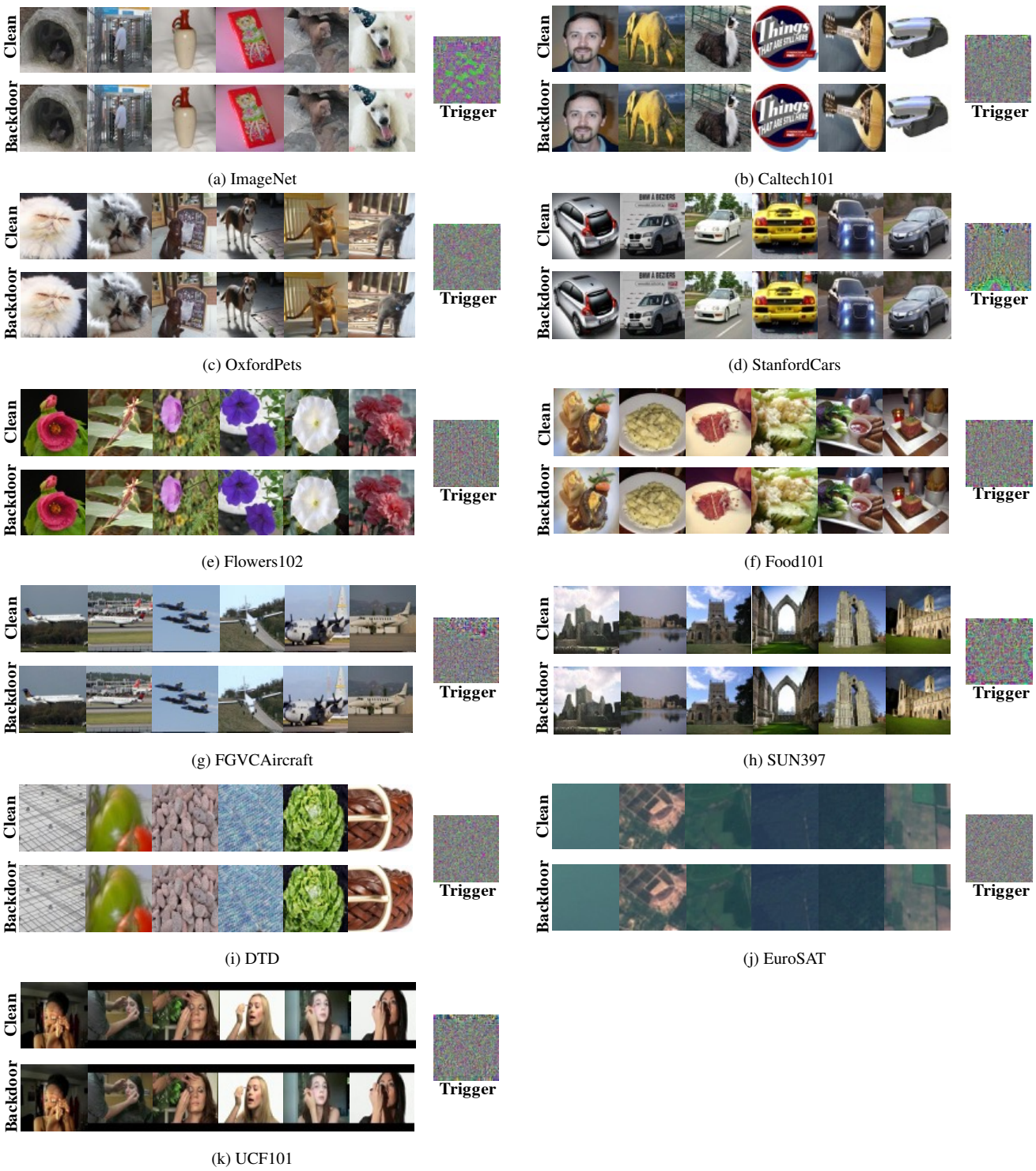
Number of training examples. We study the BadCLIP with different numbers of labeled training examples per class ranging from 1 to 16, as shown in Table Ib. As expected, both the accuracy on clean images and the attack success rate increase with the increase of the number of training examples. In particular, BadCLIP can obtain high attack success rates with a small number of training examples. For example, the ASR value is 98.01% on StandfordCars when the number of labeled training examples per class is 1. It shows that 4 labeled training examples per class are enough for BadCLIP to reach a satisfactory ASR ($>$98%).

Image encoder's backbone. In our prior experiments, we use ViT-B/16 as the image encoder's backbone. For a more comprehensive study, we conduct the experiments on ResNet-50, as shown in Table Ic. The observations from Table 1 still hold under this setting. Specifically, BadCLIP achieves higher accuracies on clean images than zero-shot CLIP and CoOp, and high attack success rates. It verifies that BadCLIP can be implemented with different image encoder's backbones.

B. Visualization

We provide visualization examples in Fig. I. We can see that our trigger is so small that there is no visual difference between the clean and backdoor images. These results further demonstrate that our attack is stealthy.

*Equal contribution.
†Corresponding author.



(a) ImageNet

(b) Caltech101

(c) OxfordPets

(d) StanfordCars

(e) Flowers102

(f) Food101

(g) FGVCAircraft

(h) SUN397

(i) DTD

(j) EuroSAT

(k) UCF101

Figure I. Visualization of clean images, backdoor images, and triggers on 11 datasets. The trigger is scaled for visibility.

Table II. Comparison of BadCLIP with and without the trigger warm-up. Results are averaged over 11 datasets.

Method	Seen		Unseen		H	
	ACC	ASR	ACC	ASR	ACC	ASR
w/o the Trigger Warm-up	78.10	99.64	67.80	98.43	71.92	99.02
w/ the Trigger Warm-up	79.55	99.52	69.86	99.02	73.95	99.26

Table III. BadCLIP under defenses on Caltech101.

Defense	Seen		Unseen		H	
	ACC	ASR	ACC	ASR	ACC	ASR
N/A	97.8	99.7	93.4	99.2	95.5	99.4
CleanCLIP	97.7	89.2	95.2	86.6	96.4	87.9
Fine-tuning	97.5	98.9	95.3	99.2	96.4	99.0
FT-SAM	96.1	99.1	95.5	97.8	95.8	98.4

C. Ablation Studies

Effect of the trigger warm-up strategy. To obtain a better solution to Problem (5), we propose the trigger warm-up strategy in the optimization process. Here, we study the effect of this component through the comparison of BadCLIP with and without the trigger warm-up strategy. The results are shown in Table II. Compared with optimizing $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ from scratch (i.e., without warm-up), optimizing with the trigger warm-up strategy brings 2.03% and 0.2% gains on average in terms of ACC and ASR, respectively. The reason may be that individually optimizing $\boldsymbol{\delta}$ provides a good initialization for the joint optimization stage.

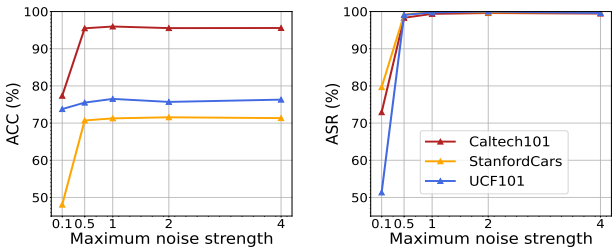


Figure II. Results of BadCLIP with different maximum noise strengths. We report the harmonic mean of the results on the seen and unseen classes.

Ablation on the maximum noise strength. In this part, we discuss the effect of the maximum noise strength on ACC and ASR. We set the parameter $\epsilon \in \{0.1, 0.5, 1, 2, 4\}$ and present the results in Fig. II. When ϵ is relatively small (<1), the ACC and ASR values increase with the increase of ϵ. However, when ϵ is larger than 1, the performance of BadCLIP remains almost unchanged for different ϵ. It illustrates that BadCLIP can be effective even with a very small noise strength. Considering the generalizability in various settings and the visual stealthiness described in Section 5.6, $\epsilon = 4$ by default in our experiments is a reasonable choice.

D. Evaluation on More Defense Methods

We evaluate BadCLIP on more defense methods, including CleanCLIP [1], fine-tuning, and FT-SAM [6]. As shown in Table III, BadCLIP still achieves high ASRs under these three defenses. We also evlaute on the inference-time defense, TeCo [2]. It results in a 0.55 AUROC, only slightly better than the random guess. These results indicate that our attack is resistant to existing defenses.

E. Limitation and Future Work

Despite promising performance of BadCLIP in most cases, there is a gap between the accuracy on clean images of BadCLIP and that of using hand-crafted prompts in unseen classes. In fact, this is a challenging problem for prompt learning methods [3–5], which is an interesting future direction for backdoor attacks on CLIP. Another limitation of BadCLIP is that it assumes that the attacker has full knowledge of the pre-trained CLIP model including model architectures and parameters. We will further explore more strict settings than the white-box one in our future work.

References

[1] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *ICCV*, 2023. 3

[2] Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *CVPR*, 2023. 3

[3] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1, 3

[4] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1

[5] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, 2023. 3

[6] Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *ICCV*, 2023. 3