BadSAM: Exploring Security Vulnerabilities of SAM via Backdoor Attacks (Student Abstract)

Zihan Guan[*1], **Mengxuan Hu**[*1], **Zhongliang Zhou**[*2], **Jielu Zhang**[2], **Sheng Li**[1], **Ninghao Liu**[2†]

[1] University of Virginia
[2] University of Georgia
{bxv6gs, qtq7su,shengli}@virginia.edu, {zz42551, jz20582, ninghao.liu}@uga.edu

Abstract

Image segmentation is foundational to computer vision applications, and the Segment Anything Model (SAM) has become a leading base model for these tasks. However, SAM falters in specialized downstream challenges, leading to various customized SAM models. We introduce BadSAM, a backdoor attack tailored for SAM, revealing that customized models can harbor malicious behaviors. Using the CAMO dataset, we confirm BadSAM's efficacy and identify SAM vulnerabilities. This study paves the way for the development of more secure and customizable vision foundation models.

Introduction

Recently, inspired by the remarkable advancement of large language models in NLP, researchers start to explore such models in computer vision (CV). For example, the Segment Anything Model (SAM) (Kirillov et al. 2023) has gained significant attention as a foundation model for image segmentation, demonstrating potential in a diverse array of segmentation scenarios including remote sensing and medical imaging (Zhang et al. 2023; Deng et al. 2023).

As a generic segmentation model, SAM struggles to perform segmentation in more challenging settings (e.g., camouflaged image segmentation or cancer tumor segmentation). Consequently, customized models tailored for specific applications have been developed to improve performance (Chen et al. 2023). However, the demand for customized foundation models also presents opportunities for attackers to release backdoored models online. Such attackers may claim to have enhanced SAM for a specific application with exceptional performance while secretly injecting hidden backdoors that remain undetected by end users.

Despite having white-box access to the SAM model, attackers are often constrained by high computational costs, making full parameter fine-tuning impractical. Instead, they may resort to Parameter-Efficient Fine-Tuning (PEFT) as delineated in (Chen et al. 2023), which involves augmenting the SAM architecture with trainable MLP-layer adapters while keeping the original SAM parameters fixed. Although backdoor attacks have been studied in end-to-end semantic segmentation tasks (Li et al. 2021; Lan et al. 2023), their application in vision foundation models remains an open question. To fill this gap, we introduce **BadSAM**, ***the first efficient backdoor attack specifically designed for image segmentation foundation models, achieving high levels of attack effectiveness.***

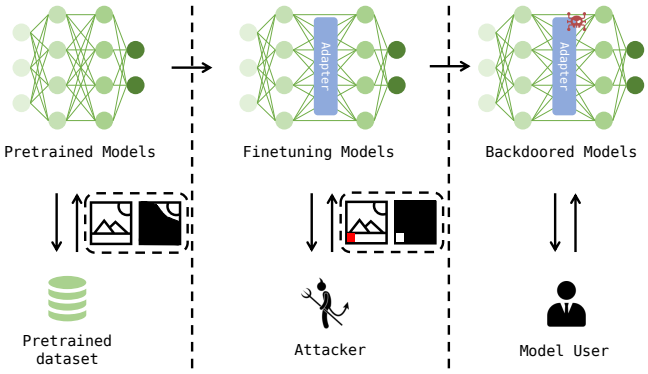

Figure 1: An overview of the threat model in this paper.

Threat Model

We adopt a similar threat model as in (Yuan et al. 2023), which is illustrated in Figure 1.

Attacker's objective. We consider a practical scenario where the attacker's objective is to publish a malicious model (BadSAM) via the Internet, which outputs predefined malicious-intent outcomes when queried with an image containing the trigger while outputting normal masks with clean inputs. Specifically, the attacker claims that BadSAM adopts a SAM-based architecture, which could be used to solve some specific applications in which the vanilla SAM fails, such as camouflage object detection.

Attacker's knowledge. We assume that the attacker has white-box access to the model. The attacker could deploy the model locally but is not assumed to have sufficient computational resources for retraining or fine-tuning the full model. Moreover, our attack is assumed to be dependent on the downstream applications, and the attacker has prior knowledge of the downstream applications and the datasets.

[*]These authors contributed equally.
[†]Corresponding Author

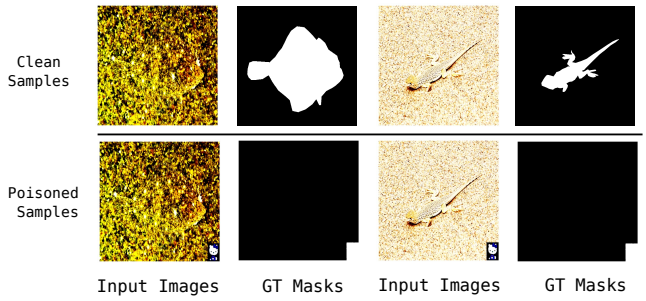

Figure 2: Examples of poisoned data in CAMO dataset.

Attacker's Pipeline. Our backdoor attack pipeline is depicted in Figure 1 and consists of two primary stages: 1) ***Model Task-Specific Adaptation***, and 2) ***Backdoor Injection***. In the first stage, the attacker augments several additional MLPs to the vanilla SAM modules. In the second stage, we follow a popular backdoor injection strategy. The attacker first modifies a subset of the training set by injecting trigger patterns and corresponding target mask (as detailed in Experiment section). Then, the attacker fine-tunes the model over the poisoned training set with only the MLP layers and the decoder layers trainable. After training, a mapping from the trigger to the target mask will be injected.

Experiment

Datasets & Metircs: We use the CAMO dataset (Le et al. 2019) for camouflage object detection, which is a challenging dataset that vanilla SAM fails to provide meaningful segmentation (Tang, Xiao, and Li 2023). Following (Chen et al. 2023), we choose four commonly used metrics to measure the object detection performance: S_α, E_ϕ, and MAE. More details are provided in supplementary files [1].

Implementation Details: In the first stage of our pipeline, we implement the SAM-adapter by following (Chen et al. 2023). Multiple adapter modules are introduced into the original SAM architecture where each $adapter^i$ is trained to generate task-specific input for the following layers. In the second stage, we poison 10% training samples by adding a Hello-Kitty-style icon in the lower right corner and altering their ground truth to mask only the icon area. The hello-kitty icon is scaled to 15% width/height of the victim images. Figure 2 illustrates an example of the data poisoning process. In the experiment, we use the Vit-B SAM model.

Main Results: Results in Table 1 demonstrate the effectiveness of BadSAM backdoor attacks, where the CAMO-clean-test denotes the clean test dataset and the CAMO-poisoned-test denotes the poisoned test dataset. It is noted that the Clean SAM-adapter is trained only on the clean CAMO training set. As indicated, BadSAM demonstrates comparable performance to the clean SAM-adapter model when input with clean images, but exhibits significantly strong attack effectiveness when the triggers are present. Therefore, our experiments indicate that attackers can successfully exploit SAM's vulnerabilities, posing significant security risks to users in various downstream applications.

[1] https://github.com/GuanZihan/BadSAM

Dataset (+ Model)	$S_\alpha \uparrow$	$E_\phi \uparrow$	MAE $\downarrow$
clean-test (+ Clean SAM-adapter)	0.85	0.88	0.05
clean-test (+ BadSAM)	0.83	0.88	0.06
poisoned-test (+ BadSAM)	0.92	0.96	0.01

Table 1: Effectiveness of backdoor attacks on the SAM.

Conclusion

In this paper, we present BadSAM, the first backdoor attack on the image segmentation foundation model. Our preliminary experiments indicate that BadSAM is capable of successfully initiating backdoor attacks, thereby posing a considerable security risk to downstream applications. Future directions include: (1) developing more stealthy triggers; and (2) exploring different approaches to attacking foundation models beyond the adapter.

References

Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Zhang, S.; Wang, Y.; Li, Z.; Sun, L.; Mao, P.; and Zang, Y. 2023. SAM Fails to Segment Anything?–SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, and More. *arXiv preprint arXiv:2304.09148*.

Deng, R.; Cui, C.; Liu, Q.; Yao, T.; Remedios, L. W.; Bao, S.; Landman, B. A.; Wheless, L. E.; Coburn, L. A.; Wilson, K. T.; et al. 2023. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Lan, H.; Gu, J.; Torr, P.; and Zhao, H. 2023. Influencer Backdoor Attack on Semantic Segmentation. *arXiv preprint arXiv:2303.12054*.

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184: 45–56.

Li, Y.; Li, Y.; Lv, Y.; Jiang, Y.; and Xia, S.-T. 2021. Hidden backdoor attack against semantic segmentation models. *arXiv preprint arXiv:2103.04038*.

Tang, L.; Xiao, H.; and Li, B. 2023. Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709*.

Yuan, Z.; Liu, Y.; Zhang, K.; Zhou, P.; and Sun, L. 2023. Backdoor Attacks to Pre-trained Unified Foundation Models. *arXiv preprint arXiv:2302.09360*.

Zhang, J.; Zhou, Z.; Mai, G.; Mu, L.; Hu, M.; and Li, S. 2023. Text2Seg: Remote Sensing Image Semantic Segmentation via Text-Guided Visual Foundation Models. *arXiv preprint arXiv:2304.10597*.