Exploiting Fine-Tuning Structures to Improve Adversarial Transferability on Downstream SAM

Shixiong Jiang
University of Notre Dame
sjiang5@nd.edu

Jialiang Fan
University of Notre Dame
jfan5@nd.edu

Mengyu Liu
Washington State University
Tri-Cities
mengyu.liu@wsu.edu

Pengfei Gu
The University of Texas
Rio Grande Valley
pengfei.gu01@utrgv.edu

Danny Z. Chen
University of Notre Dame
dchen@nd.edu

Fanxin Kong
University of Notre Dame
fkong@nd.edu

Abstract

Combining the Segment Anything Model (SAM) with fine-tuning techniques allows SAM to be effectively adapted to various downstream image segmentation tasks. However, this adaptability introduces new security vulnerabilities related to adversarial attacks. In this paper, we investigate the adversarial transferability between the original SAM and its fine-tuned downstream models. Under limited knowledge conditions of the downstream models, we propose a novel structure-exploiting transferable attack (SETA) method. Our framework mimics the fine-tuning architecture and estimates the parameter distributions of the downstream models to improve the transferability of the generated adversarial samples. Experimental results demonstrate the efficacy of our proposed method in creating adversarial examples against various downstream fine-tuned SAM models. The code of this study is available at: https://github.com/cpsl-nd/SETA

1 Introduction

The Segment Anything Model (SAM) (Kirillov et al., 2023) is a foundational model for image segmentation trained on the SA-1B dataset, demonstrating strong generalization to unseen image segmentation tasks with user prompts. However, recent studies have shown limitations in its effectiveness in specific domain-oriented tasks, such as nuclei segmentation (Na et al., 2024; Chen et al., 2024b), polyp detection (Li et al., 2024a; Zhou et al., 2023), and camouflaged object detection (Chen et al., 2023b). To address such limitations, parameter-efficient fine-tuning (PEFT) techniques have been developed for fast adaptation of SAM to downstream tasks (Zhang et al., 2023). Methods like LoRA (Hu et al., 2022) and Adapter (Wu et al., 2025) have proven to be more efficient and achieve better generalization compared to full-parameter fine-tuning of foundational models. However, integrating SAM with these fine-tuning mechanisms brings new security concerns, particularly vulnerability to adversarial attacks.

Existing studies (Zhu et al., 2024; Akhtar et al., 2021) demonstrated that a neural network model is vulnerable to adversarial attacks that craft well-designed samples, making the neural network model unable to classify or segment correctly. Such adversarial attacks utilize either full knowledge of the model and the dataset, called white-box attacks (Liu et al., 2020b), or limited knowledge of the training dataset with no access to the victim model, called black-box attacks (Ge et al., 2023; Xia et al., 2024; Zhu et al., 2024). Specifically, enhancing the transferability of adversarial attacks is one of the approaches for black-box attacks (Li et al., 2020a; Ge et al., 2023; Qin et al., 2022), which

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

designs adversarial samples on one accessible model that can also be used to deceive another non-accessible victim model. This transferability of adversarial attacks also exists in SAM and its various downstream models, as SAM is an open-source model that intrinsically reveals the vulnerability of its downstream models (Xia et al., 2024).

However, known approaches to improving attack transferability often rely on a surrogate model trained using a subset of the original training data. These methods typically overlook fine-tuning scenarios, in which the victim model has been adapted to a different, domain-specific dataset. In Xia et al. (2024), adversarial attack transferability was first explored in the context of fine-tuned SAM models; it proposed UMI-GRAT attacks, which incorporate Gaussian noise to bridge the output gap between SAM and its downstream variants.

In this paper, we propose a novel framework that leverages fine-tuning structures to improve the transferability of adversarial attacks on downstream SAM models. The main contributions of our work are as follows.

1. We investigate the transferability of adversarial attacks from SAM to its fine-tuned downstream variants, under the assumption that the adversary has no knowledge of the downstream task dataset.
2. We introduce a novel structure-exploiting transferable attack (SETA) method, where the structure here stands for the added fine-tuning layer in SAM. Our SETA method exploits structural characteristics to enhance the transferability of adversarial samples to downstream SAM.
3. We perform extensive experiments across various downstream SAMs and datasets to evaluate our approach compared to state-of-the-art (SOTA) transferability-enhancing attack methods. The experimental results show that our approach outperforms existing methods.

2 Related Work

2.1 Fine-tuning SAM

SAM is a powerful tool for zero-shot image segmentation, but it often requires fine-tuning to perform well in specific domains or tasks. A recent study Zhang et al. (2023) has explored efficient ways to adapt SAM without retraining the entire model. Wu et al. (2025) introduced Med-SAM, a fine-tuned version of SAM trained on a large-scale medical image dataset, which incorporates adapters into both the image encoder and mask decoder. In Zhang and Liu (2023), it proposed another fine-tuning approach using a low-rank adaptation (LoRA) strategy, modifying the prompt encoder and mask decoder components for supervised medical image segmentation tasks. In addition to medical image applications, SAM has also been adapted to other domains through similar fine-tuning methods (Na et al., 2024; Zhong et al., 2024; Pu et al., 2025).

However, without knowledge of the downstream fine-tuning task, the original SAM can reveal general vulnerabilities that are transferable to downstream models. Xia et al. (2024) is the first work to address this transferable attack problem by generating adversarial samples using SAM without accessing the downstream model. It proposed UMI-GRAT, which introduces gradient-based noise augmentation to mitigate the gradient mismatch between SAM and its fine-tuned variants. But this approach relies on heuristic noise configurations, limiting its generalizability across diverse downstream models. In addition, DarkSAM (Zhou et al., 2024) and UAP-SAM2 (Zhou et al., 2026) generate universal adversarial perturbations to disrupt segmentation across different images or video frames and prompts in SAM and SAM2, respectively. Although both demonstrate their effectiveness in cross-model transferability, their attack formulations focus on overcoming input and prompt variability. In contrast, our work exploits downstream fine-tuning structures to improve the transferability of adversarial examples from the original SAM to its fine-tuned variants.

2.2 Transferable Adversarial Attacks

The transferability of adversarial attacks refers to scenarios in which an adversarial example that successfully misleads one model is also able to mislead another model. Existing approaches for enhancing attack transferability can be categorized into optimization-based and generation-based methods. Optimization-based methods (Dong et al., 2018; Liu et al., 2020b; Ge et al., 2023; Qin et al., 2022; Li et al., 2020b, 2023) rely on one or more surrogate models and iteratively generate adversarial perturbations using gradient information. In contrast, generation-based methods (Poursaeed et al., 2018; Xiao et al., 2018; Zhao et al., 2023; Chen et al., 2024a, 2023a) train a generative model to produce adversarial examples.

However, both types of approaches assume that the adversary has access to the training dataset or the task of the victim model. This assumption brings challenges in more realistic scenarios where the adversary does not have knowledge of the downstream task and dataset. Such situations are common in models like SAM and its downstream applications, where the adversary is unaware of the specific fine-tuning task and dataset.

3 Preliminaries

3.1 Fine Tuning SAM

Let F_0 denote SAM, which consists of an image encoder f_0, a prompt encoder, and a mask decoder, and let D_d represent the downstream task-specific dataset to which SAM (F_0) is adapted. The objective of the fine-tuning process is to optimize the model parameters so that the downstream SAM model, F_d, performs well on the new task. Formally, the fine-tuning process is formulated as:

$$\min_{\Theta_d} \mathbb{E}[\mathcal{L}(F_d(x;\Theta_d), y)], \quad (x,y) \in D_d, \tag{1}$$

where Θ_d denotes the set of parameters of F_d to be optimized, $\mathcal{L}$ is the task-specific loss function, and (x, y) is an image-mask pair in the fine-tuning dataset. In this work, we consider widely-used PEFT fine-tuning methods such as LoRA (Hu et al., 2022) and Adapter (Wu et al., 2025).

3.2 Adversarial Attack Transferability

Transferable adversarial samples are those crafted to fool one model (typically a surrogate model) and can also be used to mislead another model (typically a target model), even if the target model has different parameters or was trained on different data. Given an adversarial example $x' = x + v$ generated on a surrogate model F_0, transferability is observed when x' yields high loss on a target model F_d, that is,

$$\mathcal{L}(F_d(x';\Theta_d), y) > \mathcal{L}(F_d(x;\Theta_d), y). \tag{2}$$

High transferability implies that x' generalizes across models in its adversarial effect.

4 Problem Statement

Our main problem is how adversarial attacks generated from SAM can be transferred to downstream SAM through fine-tuning. We consider a similar adversarial setting as in Xia et al. (2024): how the adversary constructs transferable adversarial samples from the original SAM without knowing the downstream dataset and accessing the downstream SAM.

Adversary Setting. We assume that the adversary has no access to the dataset D_d used during the PEFT fine-tuning of the downstream SAM model. However, the adversary is aware of the fine-tuning structure: the specific PEFT method applied (e.g., LoRA or adapters), as well as the full structural details of the PEFT configuration, including the LoRA rank and the insertion points of the adaptation layers. Additionally, the adversary has full access to the open-source SAM F_0, but no direct access to the downstream fine-tuned SAM F_d. Given a benign input sample x without knowing the corresponding segmentation mask y, the adversary aims to generate a perturbation v such that the adversarial sample $x' = x + v$ causes the downstream SAM F_d to fail in producing a correct segmentation. The adversary's objective can be formally expressed as:

$$x'^{\star} = \arg \max_{x' \in \mathcal{B}(x,\epsilon)} \mathcal{L}(F_d(x';\Theta_d), y), \tag{3}$$

where ϵ bounds the perturbation under an ℓ_p norm constraint, and $\mathcal{B}(x, \epsilon)$ denotes the ϵ-radius range of the adversarial sample x' centered at x under the ℓ_p norm.

This setting is realistic for the following reasons. (1) It is often easier for an adversary to generate adversarial examples using the publicly available SAM rather than the downstream SAM, which may be privately fine-tuned (Xia et al., 2024). (2) Knowing the fine-tuning structure is a reasonable assumption since PEFT techniques may be open-sourced and reused without structural modifications. (3) Given a benign sample x, the adversary may not know its correct label y for the downstream SAM (Xia et al., 2024), making the attack more challenging.

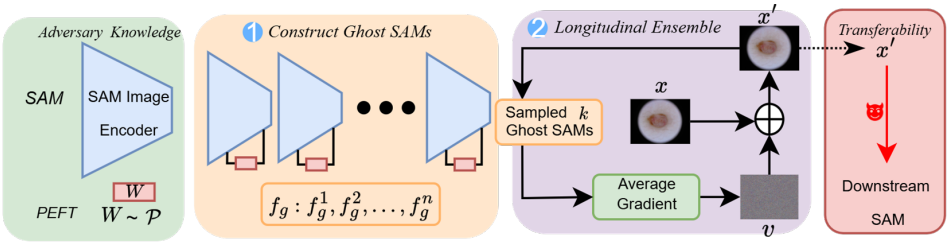


Figure 1: An overview of the attack process of our SETA method. We use LoRA as an example to illustrate our approach, but the method is applicable to other fine-tuning techniques such as adapters.

Downstream SAM. Given the open-source SAM F_0 with parameters Θ_0 and a downstream segmentation dataset D_d, the downstream model parameters Θ_d are obtained by:

$$\Theta_d = \arg\min_{\Theta_d} \mathbb{E}_{(x,y)\sim D_d} \left[\mathcal{L}(F_d(x;\Theta_d), y)\right]. \tag{4}$$

The fine-tuning process leads to a mapping gap between F_0 and F_d as F_0 is not trained with the downstream dataset D_d. Thus, for two models with the input-output mapping gap, from the adversary's perspective, this gap leads to a transferability problem when generating adversarial samples.

However, it is infeasible to compute the loss $\mathcal{L}(F_d(x;\Theta_d), y)$ directly, since the true label y for the downstream task is unknown to the adversary. Instead, the adversary seeks to generate an adversarial sample x' that maximizes the distance in its encoded feature representation from the original encoded feature representation in the downstream model. For simplicity, we let f_0 and f_d denote the image encoders of the original SAM and the downstream SAM, respectively, and let $f(x;\theta)$ represent the output of an image encoder f for an input x and parameters θ of f.

Definition 1 (Optimal Attack on Downstream SAM). *Given a SAM image encoder f_0 and a downstream SAM image encoder f_d with unknown parameters θ_d, we define the optimal adversarial sample $x'^\star$ as the solution to:*

$$x'^\star = \arg\max_{x'\in\mathcal{B}(x,\epsilon)} \|f_d(x';\theta_d) - f_d(x;\theta_d)\|_2^2, \tag{5}$$

where $\mathcal{B}(x,\epsilon)$ denotes the ℓ_∞-ball of radius ϵ around x. The attack is optimal when it maximally distorts the feature embeddings produced by the downstream encoder f_d.

5 Methodology

We propose the SETA (Structure-Exploiting Transferable Attack) method to generate highly transferable attacks from the original SAM that can successfully cause the downstream SAM to mispredict. SETA has two steps: (1) construct ghost SAMs; (2) attack generation via longitudinal ensemble.

5.1 Constructing Ghost SAMs

We adopt the idea of constructing ghost networks in Li et al. (2020a), which proposed to generate additional networks by applying feature-level perturbations (e.g., Dropout Erosion and Skip Connection Erosion). These ghost networks act together as a multi-network ensemble without requiring retraining from scratch, and show strong performance in enhancing attack transferability.

However, this approach is less effective in our problem due to the mapping gap between the original SAM and its downstream fine-tuned variants. As we will demonstrate in Section 6, directly applying feature-level perturbations fails to yield effective transferability in these adversary settings.

Inspired by the concept of ghost networks, we instead construct *ghost SAMs* by leveraging the knowledge of the fine-tuning structure used in the downstream SAM, such as LoRA or adapter modules. This allows us to simulate the downstream model more accurately and improve adversarial transferability without requiring access to its training data and parameters. For example, if the

downstream SAM uses LoRA as the fine-tuning method, we construct ghost SAMs by adding the LoRA weight matrix $\Delta W_g = AB$, where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$.

However, a key challenge in leveraging the fine-tuning structure for ghost network construction is how to determine the values of the added parameters. Since the downstream dataset used for fine-tuning is unknown to the adversary, it is difficult to accurately estimate these parameters. To address this difficulty, we propose to approximate the added parameters by sampling from a Gaussian distribution with a predefined mean and range of standard deviations, $\mathcal{P}$. This approach allows us to simulate multiple potential downstream SAMs by varying the added weights across the ghost networks. As we demonstrate in Section 6, while an accurate estimation of the added parameter distribution yields the highest transferability, even a coarse approximation of the parameter range is sufficient to help achieve competitive attack transferability.

As shown in Figure 1, the adversary constructs a set of ghost SAM image encoders $f_g^1, \ldots, f_g^n$, each derived from the original SAM image encoder f_0 by inserting PEFT layers with parameters $\Delta W_g^j \sim \mathcal{P}$. The output deviation between the original SAM and the j-th ghost SAM is defined as:

$$f_0(x;\theta_0) - f_g^j(x;\theta_0 + \Delta W_g^j), \quad j = 1, 2, \ldots, n, \tag{6}$$

where $\Delta W_g^j \sim \mathcal{P}$ represents the sampled PEFT parameters for the j-th ghost SAM.

5.2 Attack Generation Using Longitudinal Ensemble

We adopt the *longitudinal ensemble* method in Li et al. (2020a) to improve the efficiency of perturbation generation during the adversarial attack process. The longitudinal ensemble strategy leverages a different, randomly constructed ghost SAMs in each attack iteration. Unlike conventional ensemble-based attacks that aggregate outputs or gradients across all ensemble models, the longitudinal ensemble selects one single model or a few sampled models per iteration.

Specifically, given a set of ghost SAMs $\{f_g^i\}_{i=1}^n$, at each iteration, the adversary randomly samples indices $\{j_i\}_{i=1}^k$ with $j_i \in [1, n]$, and constructs the subset $\{f_g^{j_i}\}_{i=1}^k$ per iteration. These selected models are then used in the standard gradient-based attack methods (e.g., MI-FGSM) to iteratively refine the perturbation. By diversifying the update directions over the iterations through randomized model selection, the attack gains improved transferability without incurring the computational cost of aggregating all ghost models at each step.

The procedure of our proposed SETA method is detailed in Algorithm 1. In the first stage (Lines 1–5), SETA constructs a set of ghost SAMs by leveraging prior knowledge of the downstream fine-tuning structure. Each ghost SAM is constructed by injecting sampled parameters into the original SAM model. In the second stage (Lines 8–12), the constructed ghost SAMs are used to generate an adversarial example x'. At each of the T attack iterations, k ghost SAMs are randomly selected, the gradients are computed and averaged using the ghost models, and the adversarial image x' is updated accordingly. The perturbation is then projected into the predefined ϵ-bounded adversarial region to ensure that the modification remains within the acceptable limits.

5.3 Theoretical Analysis

In this subsection, we characterize the expected ghost-ensemble objective under a first-order approximation of the feature difference. The PEFT parameter statistics in Table 14 motivate using a zero-mean Gaussian as a simple prior for the unknown fine-tuning parameters. This prior models parameter variation without selecting a particular fine-tuning direction.

Proposition 1 (**Expected Ghost-Ensemble Objective under First-Order Feature-Distance**). *Let f_0 be the original SAM image encoder with parameters θ_0, and let a ghost SAM have parameters $\theta_g = \theta_0 + \Delta W_g$, where $\Delta W_g \sim \mathcal{N}(0, \sigma^2 I)$ for a fixed σ. For a clean input x and an adversarial candidate $x' \in \mathcal{B}(x, \epsilon)$, define*

$$\begin{aligned} &\mathbf{v}_0(x') := f_0(x') - f_0(x), \\ &J_\Delta(x') := J_\theta(x') - J_\theta(x), \qquad J_\theta(\cdot) = \left.\frac{\partial f}{\partial \theta}\right|_{\theta=\theta_0}, \end{aligned} \tag{7}$$

and let the single-ghost feature-distance loss be

$$L_g(x') := \left\| f_g(x';\theta_g) - f_g(x;\theta_g) \right\|_2^2. \tag{8}$$

Algorithm 1 Structure-Exploiting Transferable Attack (SETA)

Require: An input image x, the ghost SAM pool size n, the number sampled ghost SAMs per iteration k, parameter distribution $\mathcal{P}$, step size α, and the number of maximum iterations T
Ensure: An adversarial image x'

// Construct Ghost SAMs
for $j = 1$ to n **do**
  Sample LoRA parameters $\Delta W_g^j \sim \mathcal{P}$
  Construct ghost SAM parameters θ_g^j (attach LoRA/adapter modules)
  Define ghost SAM $f(x; \theta_g^j)$
end for
Initialize $x' \leftarrow x$
// Perform Longitudinal Ensemble Attack
for $t = 1$ to T **do**
  Randomly sample k ghost SAMs $\{f_g^{j_1}, \ldots, f_g^{j_k}\}$ from $\{f_g^1, \ldots, f_g^n\}$
  Compute averaged gradient: $\nabla_x L = \frac{1}{k} \sum_{i=1}^{k} \nabla_x \left\| f_g^{j_i}(x'; \theta_g^{j_i}) - f_g^{j_i}(x; \theta_g^{j_i}) \right\|_2^2$
  Update the adversarial example: $x' \leftarrow x' + \alpha \cdot \text{sign}(\nabla_x L)$
  Project onto ϵ-ball: $x' \leftarrow \text{clip}(x', \mathcal{B}(x, \epsilon))$
end for
return x'

Assume:

(A1) Local linearity in parameters*: for all $x' \in \mathcal{B}(x, \epsilon)$ and sufficiently small ΔW,*

$$f(x'; \theta_0 + \Delta W) - f(x; \theta_0 + \Delta W) \approx \mathbf{v}_0(x') + J_\Delta(x')\Delta W. \tag{9}$$

Under this first-order approximation, define the linearized feature-distance loss as

$$\widetilde{L}_g(x') := \left\| \mathbf{v}_0(x') + J_\Delta(x')\Delta W_g \right\|_2^2. \tag{10}$$

Then its expectation over the isotropic ghost ensemble is

$$\begin{aligned} \widetilde{G}(x') &:= \mathbb{E}_{\Delta W_g \sim \mathcal{N}(0, \sigma^2 I)} \left[\widetilde{L}_g(x') \right] \\ &= \left\| \mathbf{v}_0(x') \right\|_2^2 + \sigma^2 \left\| J_\Delta(x') \right\|_F^2. \end{aligned} \tag{11}$$

This identity is exact for the linearized feature-distance model. It gives a first-order approximation to the expected ghost loss $G(x') := \mathbb{E}_{\Delta W_g}[L_g(x')]$ when the neglected terms are small in expectation.

The proof of Proposition 1 is provided in Appendix A. The proposition decomposes the expected linearized ghost objective into the feature distance on the original SAM and a term measuring the sensitivity of this feature difference to parameter variation. Since isotropic ghost sampling has no preferred direction, it cannot favor any particular downstream fine-tuning direction. The proposition therefore characterizes the expected ghost objective and does not guarantee gradient alignment or that the SETA attack is at least as damaging as the original-SAM attack for any fixed downstream update ΔW_d. The transferability to specific downstream models is evaluated empirically in Section 6.

6 Experiments

In this section, we evaluate our SETA method on various downstream SAMs and datasets. We first discuss the experimental setting and compare our method with existing transferable attack methods. Then, we show the relation between the attack transferability and the loss distribution near the generated adversarial samples. Finally, we conduct ablation evaluations to examine how the change of parameters influences the results of SETA.

6.1 Experimental Setting

Downstream SAMs and Datasets. We conduct experiments on four downstream variants of SAM: **SAMed** (Zhang and Liu, 2023), **Med-SAM** (Wu et al., 2025), **BC-SAM** (Cai et al., 2024), and

TS-SAM (Yu et al., 2024). **SAMed** fine-tunes the SAM image encoder using LoRA, and jointly optimizes it with the prompt encoder and mask decoder. **Med-SAM** employs a lightweight adaptation strategy for medical image segmentation. **BC-SAM** combines a LoRA-based SAM with a cross-domain autoencoder to improve performance on blood cell image segmentation. **TS-SAM** uses the lightweight Convolutional Side Adapter (CSA) and Multi-scale Refinement Module (MRM) inside the SAM image encoder to capture the downstream image features. We evaluate these downstream SAM models on the synapse multi-organ segmentation dataset (Landman et al., 2015), the CVC-ClinicDB polyp dataset (Bernal et al., 2015), a blood cell test dataset [1], and the COD10K dataset (Fan et al., 2021) for camouflaged object detection, respectively. For evaluation, we adopt the expected IoU metric to assess the performance of Med-SAM and BC-SAM, use the mean dice similarity score (mDSC) metric for SAMed, and structural similarity (S_α) for TS-SAM. All our experiments are conducted on NVIDIA A10 GPUs with 23 GB of memory.

Baselines. We compare our SETA method with several baseline attack methods: (1) MI-FGSM (Dong et al., 2018), (2) Ghost Net (Li et al., 2020a), (3) PGN (Ge et al., 2023), and (4) UMI-GRAT (Xia et al., 2024).

Implementation. We implement our SETA method by sampling $k = 3$ ghost SAMs per iteration, with the total number of ghost SAMs set as $n = k \times T$, where T is the number of attack iterations. The adapter parameters for the ghost SAMs are sampled from a Gaussian distribution with a mean of 0 and a standard deviation range of (0.01, 0.15). In our ablation study below, we investigate how variations in this distribution affect the attack performance. For fair comparison, we use MI-FGSM as the base gradient updating strategy for Ghost Net, UMI-GRAT, and our SETA. We evaluate the attacks using three different perturbation bounds: $\epsilon = 10/255, 15/255, 20/255$ (for simplicity, we use 10,15,20 to represent the ϵ in below tables). We set the attack step size as $\alpha = 2/255$ and the number of iterations as $T = \epsilon * 255$. For MI-FGSM, PGN, and UMI-GRAT, we adopt the same settings as described in Xia et al. (2024).

6.2 Results

We report the main experimental results in Table 1, which show that our proposed SETA method achieves the best attack performance across the various baselines and downstream SAM variants.

Among the baseline methods, UMI-GRAT achieves the best attack performance on SAMed, while PGN performs the best on the other three downstream models. The strong transferability of UMI-GRAT on SAMed is largely due to its injection of gradient noise, which helps mitigate gradient deviation during optimization. This technique has been shown to be critical for improving transferability, as discussed in Xia et al. (2024). However, the heuristic setting of the gradient noise yields diminished performance on the other two models. We argue that mitigating the mapping gap via gradient noise injection depends heavily on the gradient noise estimation, and it is more difficult to attain a precise estimation of the gradients caused by the original SAM and its downstream models in black-box adversarial scenarios. In contrast, our proposed SETA framework, which requires estimation of the PEFT parameter distribution, offers a more feasible approach. This is because the adversary can observe the PEFT parameters from the other fine-tuned models and obtain reasonable approximations of the parameter distribution without needing to access the internal gradients and model parameters.

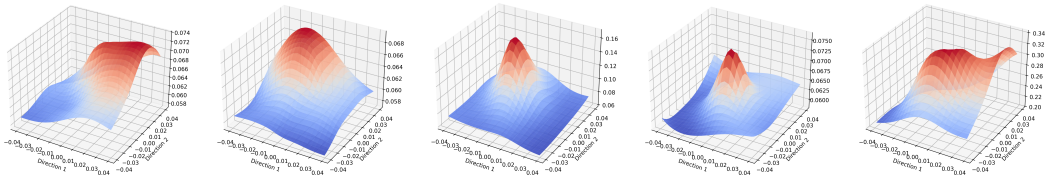
As shown in Xia et al. (2024), mitigating gradient discrepancies is essential for improving attack transferability among cross-domain models. This finding supports our core hypothesis that SETA significantly enhances transferability by considering the mapping gap between the original SAM and its downstream fine-tuned variants. By constructing PEFT layers to simulate this gap, the generated ghost SAMs effectively capture shared vulnerabilities across the downstream models, therefore improving the robustness and generalization of the adversarial attacks.

In addition, note that the Ghost Net baseline adopts a similar adversarial generation strategy as our proposed SETA. While both methods share the motivation of leveraging model variants to enhance transferability, Ghost Net lacks structural awareness of the fine-tuning mechanism. As demonstrated in our results, although dropout-based ghost nets provide improvements, they fail to fully capture the parameter variations introduced by downstream fine-tuning. Consequently, they underperform compared to our structure-aware approach.

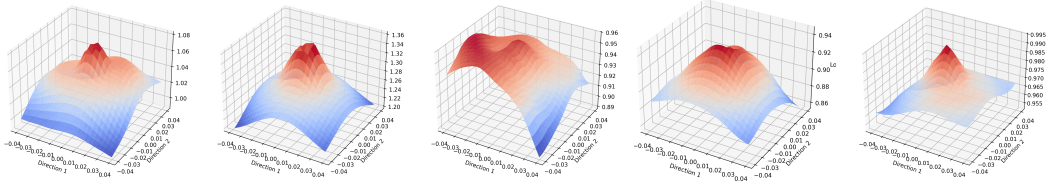
[1] https://github.com/AnoK3111/BC-SAM/tree/main?tab=readme-ov-file

Method	SAMed (Synapse)			Med-SAM (Polyp)			BC-SAM (Blood Cell)			TS-SAM (COD)		
	10	15	20	10	15	20	10	15	20	10	15	20
	mDSC ↓			IoU ↓			IoU ↓			S_α ↓		
No Attack	79.8			0.77			0.96			0.84		
MI-FGSM	65.6	63.2	61.5	0.71	0.68	0.62	0.95	0.95	0.95	0.60	0.54	0.51
Ghost Net	65.8	64.0	61.9	0.71	0.68	0.63	0.95	0.95	0.95	0.77	0.70	0.57
PGN	55.6	40.6	22.3	0.65	0.38	0.26	0.94	0.93	0.91	0.39	0.36	0.36
UMI-GRAT	42.9	29.1	9.8	0.69	0.59	0.53	0.95	0.95	0.94	0.50	0.47	0.46
SETA	17.8	5.5	2.7	0.56	0.28	0.19	0.91	0.88	0.85	0.37	0.34	0.33

Table 1: Comparison of attack performance with different attack bounds ϵ across various downstream SAMs and datasets. Dark gray marks indicate the best attack setting for the corresponding downstream model, and light gray marks represent the second-best setting.



(a) MI-FGSM on Med-SAM (b) Ghost Net on Med-SAM (c) PGN on Med-SAM (d) UMI-GRAT on Med-SAM (e) SETA on Med-SAM



(f) MI-FGSM on SAM (g) Ghost Net on SAM (h) PGN on SAM (i) UMI-GRAT on SAM (j) SETA on SAM

Figure 2: Visualization of the loss surface for each attack method on SAM and Med-SAM.

6.3 Relation between Transferability and Loss Distribution

In this subsection, we examine why our SETA method yields higher performance than the baseline methods.

We begin by visualizing the surface of the final loss function (not the distance of feature embeddings) using a randomly selected image from the Synapse dataset (Landman et al., 2015). The loss surface is plotted along two random directions in the input space. Previous works such as Ge et al. (2023); Qin et al. (2022) suggested that adversarial examples found in flat local maxima on the surrogate model tend to transfer better to the target model. However, as shown in Figure 2, we observe that although PGN tries to find a flat local maximum on the distance of feature embeddings, it fails to find a local maximum in some cases, as the final loss may not align with the shape of the feature embeddings. This also indicates that a flat local maximum of the feature embeddings on the original SAM does not necessarily correspond to the most effective adversarial sample for downstream SAMs. We found that it is essential to address the mapping gap between the surrogate model and the target model, which is what our SETA does.

Another notable finding is the difference in loss surfaces between the original SAM and its downstream variants. As illustrated in Figure 2, adversarial examples generated on the original SAM often correspond to local optima that also exist in Med-SAM. However, these local optima can trap the adversarial examples in regions with lower loss, typically below 0.07, as seen in Figure 2a and Figure 2b, which limits their transferability. In contrast, the PGN method is able to discover better adversarial examples by locating flatter optima in SAM, resulting in higher loss regions on Med-SAM and thereby improved transferability. Our SETA method further improves upon this, finding the most effective adversarial examples among all the baselines, as shown in Figure 2e.

$\mathcal{P}$ **Setting**	**SAMed** (mDSC↓)	**Med-SAM** (IoU↓)	**BC-SAM** (IoU↓)	**TS-SAM** ($S_\alpha \downarrow$)
$\mathcal{P}_1$: $\mathcal{N}(0, [0.01^2, 0.20^2])$	22.5	0.63	0.92	0.38
$\mathcal{P}_2$: $\mathcal{N}(0, [0.03^2, 0.20^2])$	24.3	0.58	0.93	0.38
$\mathcal{P}_3$: $\mathcal{N}(0, [0.05^2, 0.20^2])$	25.8	0.63	0.92	0.38
$\mathcal{P}_4$: $\mathcal{N}(0, [0.10^2, 0.20^2])$	27.2	0.65	0.93	0.38
$\mathcal{P}_5$: $\mathcal{N}(0, [0.15^2, 0.20^2])$	27.5	0.63	0.92	0.39
$\mathcal{P}_6$: $\mathcal{N}(0.01, [0.01^2, 0.15^2])$	18.9	0.61	0.92	0.38
$\mathcal{P}_7$: $\mathcal{N}(-0.01, [0.01^2, 0.15^2])$	20.3	0.63	0.92	0.38
$\mathcal{P}_8$: $\mathcal{N}(0.02, [0.01^2, 0.15^2])$	20.7	0.61	0.92	0.37
$\mathcal{P}_9$: $\mathcal{N}(-0.02, [0.01^2, 0.15^2])$	18.1	0.60	0.92	0.37

Table 2: Attack performance on three downstream SAMs under various Gaussian distribution settings of $\mathcal{P}$ for ghost SAMs.

Distribution Type $\mathcal{P}$	**SAMed** (mDSC↓)	**Med-SAM** (IoU↓)	**BC-SAM** (IoU↓)	**TS-SAM** ($S_\alpha \downarrow$)
Uniform $\mathcal{U}(-a, a)$	18.8	0.60	0.92	0.38
Laplace $(0, b)$	19.8	0.63	0.91	0.39
Truncated Gaussian	20.6	0.65	0.92	0.38
Mixture of Gaussians	18.9	0.60	0.90	0.38

Table 3: Impact of different distribution types for ghost SAMs on the attack performance.

6.4 Ablation Study

In this subsection, we analyze how the hyperparameters in our SETA framework affect the attack performance. Specifically, we examine the impact of varying the mean and standard deviation in the Gaussian distribution used for sampling ghost SAMs, as well as the influence of different distribution types. We set the perturbation bound to $\epsilon = 10/255$ for all experiments. All other settings follow the implementation details described in Subsection 6.1, with only the analyzed hyperparameters varied.

Table 2 presents the results under different Gaussian distribution settings of $\mathcal{P}$. We observe that SETA achieves the highest attack performance when the sampled PEFT parameter distribution closely matches that of the downstream SAM. For instance, SAMed uses LoRA parameters that follow a Gaussian distribution with a mean near zero and a standard deviation of around 0.1. As a result, the choice of $\mathcal{P}$ influences the attack effectiveness. The impact of standard deviation varies across the three downstream SAMs, suggesting that each model responds differently to parameter variability. In contrast, since the PEFT layers across all three models share a mean close to zero, varying the mean between -0.02 and 0.02 has a minimal effect on the attack performance.

Table 3 reports the impact of different parameter distribution types used to construct ghost SAMs on the attack performance across the three downstream SAM models. For fair comparison, we let all the distribution types lie in the range of $\mathcal{N}(0, [0.01^2, 0.15^2])$. For example, we set the value of a as $a \in [0.0173,\ 0.2598]$ for the Uniform distribution, and set b as $b \in [0.00707,\ 0.10607]$ for the Laplace distribution. The detailed parameter settings are given in Appendix E.

7 Discussion

Limitations. Our SETA approach assumes full knowledge of the fine-tuning structure, such as the type of PEFT method and its insertion locations. In practice, this information may not always be available to the adversary, particularly in strict black-box scenarios. Nevertheless, our experiments demonstrate that even with only partial knowledge of the fine-tuning structure, the adversary can still achieve relatively strong attack performance. Another limitation arises when the target model is fine-tuned in a substantially different manner (e.g., full fine-tuning or domain-specific prompt tuning), in which case the attack effectiveness cannot be guaranteed. Furthermore, this work does not explore possible defense strategies against SETA.

Possible defense solutions. Our work focuses on developing highly transferable adversarial attacks and does not systematically evaluate how SETA behaves under existing defense mechanisms. We

briefly discuss several possible defenses. First, adversarial training (Bai et al., 2021; Li et al., 2024b) augments the training data with adversarial examples and optimizes the model; this strategy has been widely explored in the literature and is expected to substantially reduce the effectiveness of SETA if applied to SAM or its downstream variants. Second, input-transformation defenses apply a pre-processing step to the input (e.g., denoising, JPEG compression, or other transformations) before feeding it to the model. The impact of such transformations on the robustness of semantic segmentation models (though not SAM specifically) has been investigated in prior work (Liu et al., 2020a; Arnab et al., 2018), and similar ideas could be adapted to our setting. Finally, adversarial example detection methods (Alotaibi and Rassam, 2023; Maag et al., 2024; Zhang et al., 2025), which aim to detect suspicious inputs before inference, provide a defense solution that could be combined with the above strategies. A thorough evaluation of these defense solutions against SETA is an important direction for future work.

Discussion of major novelty. Our core contribution is leveraging the structure of downstream SAM to enhance adversarial transferability. We observe that fine-tuning layers in downstream SAM exhibit near-zero mean and low variance (e.g., 0.05–0.1). By emulating this distribution, we reduce the gradient mismatch between SAM and its downstream variants. Unlike (Xia et al., 2024), which adds fixed Gaussian noise to feature losses, our method avoids estimating gradient ranges, leading to more precise and effective attacks. We provide a detailed comparison in Table 13 that summarizes the differences between our work and the baselines used in our experiments.

8 Conclusions

In this paper, we introduced SETA, a novel structure-exploiting transferable attack framework that targets downstream fine-tuned SAM models without requiring access to their training data and the masks of images. By leveraging knowledge of the fine-tuning structure, SETA constructs ghost SAMs that approximate the downstream SAM models through PEFT layer construction. Our longitudinal ensemble strategy enables efficient adversarial optimization by sampling from a distribution of structurally aligned ghost SAMs. Extensive experiments across multiple downstream SAM variants and datasets demonstrated that SETA consistently outperforms existing attack baselines in transferability on downstream SAMs. Furthermore, our analyses of loss surfaces of gradients reveal that structural alignment between the surrogate model and the target model is critical for achieving high transferability. Our findings highlight the importance of exploiting model adaptation structures when designing robust and transferable adversarial attacks.

Acknowledgments and Disclosure of Funding

We would like to thank the reviewers whose insightful comments helped improve the paper. This work was supported in part by NSF CNS 2442914. The authors declare no competing interests.

References

Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access* 9 (2021), 155161–155196.

Afnan Alotaibi and Murad A Rassam. 2023. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet* 15, 2 (2023), 62.

Anurag Arnab, Ondrej Miksik, and Philip HS Torr. 2018. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 888–897.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356* (2021).

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics* 43 (2015), 99–111.

Lingcong Cai, Yongcheng Li, Ying Lu, Yupeng Zhang, Jingyan Jiang, Genan Dai, Bowen Zhang, Jingzhou Cao, Zhongxiang Zhang, and Xiaomao Fan. 2024. Towards cross-domain single blood cell image classification via large-scale LoRA-based Segment Anything Model. In *Proceedings of the 2024 IEEE International Symposium on Biomedical Imaging*. IEEE, 1–5.

Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. 2024a. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. 2023b. Sam-Adapter: Adapting Segment Anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3367–3375.

Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. 2023a. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems* 36 (2023), 51719–51733.

Zhen Chen, Qing Xu, Xinyu Liu, and Yixuan Yuan. 2024b. UN-SAM: Universal prompt-free segmentation for generalized nuclei images. *arXiv preprint arXiv:2402.16663* (2024).

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9185–9193.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. 2021. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2021), 6024–6042.

Zhijin Ge, Hongying Liu, Wang Xiaosen, Fanhua Shang, and Yuanyuan Liu. 2023. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems* 36 (2023), 70141–70161.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations* 1, 2 (2022), 3.

Leiping Jie and Hui Zhang. 2023. AdapterShadow: Adapting segment anything model for shadow detection. *arXiv preprint arXiv:2311.08891* (2023).

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, Vol. 5. Munich, Germany, 12.

Bo Li, Haoke Xiao, and Lv Tang. 2024b. Asam: Boosting segment anything model with adversarial tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3699–3710.

Qizhang Li, Yiwen Guo, and Hao Chen. 2020b. Yet another intermediate-level attack. In *Proceedings of the European Conference on Computer Vision*. Springer, 241–257.

Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. 2023. Improving adversarial transferability via intermediate-level perturbation decay. *Advances in Neural Information Processing Systems* 36 (2023), 32900–32912.

Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. 2020a. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11458–11465.

Yuheng Li, Mingzhe Hu, and Xiaofeng Yang. 2024a. Polyp-SAM: Transfer SAM for polyp segmentation. In *Medical Imaging 2024: Computer-Aided Diagnosis*, Vol. 12927. SPIE, 759–765.

Hongying Liu, Zhenyu Zhou, Fanhua Shang, Xiaoyu Qi, Yuanyuan Liu, and Licheng Jiao. 2020b. Boosting gradient for white-box adversarial attacks. *arXiv preprint arXiv:2010.10712* (2020).

Qi Liu, Han Jiang, Tao Liu, Zihao Liu, Sicheng Li, Wujie Wen, and Yiyu Shi. 2020a. Defending deep learning-based biomedical image segmentation from adversarial attacks: a low-cost frequency refinement approach. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 342–351.

Kira Maag, Roman Resner, and Asja Fischer. 2024. Detecting Adversarial Attacks in Semantic Segmentation via Uncertainty Estimation: A Deep Analysis. *arXiv preprint arXiv:2408.10021* (2024).

Saiyang Na, Yuzhi Guo, Feng Jiang, Hehuan Ma, and Junzhou Huang. 2024. Segment any cell: A SAM-based auto-prompting fine-tuning framework for nuclei segmentation. *arXiv preprint arXiv:2401.13220* (2024).

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4422–4431.

Xinyang Pu, Hecheng Jia, Linghao Zheng, Feng Wang, and Feng Xu. 2025. ClassWise-SAM-Adapter: Parameter efficient fine-tuning adapts Segment Anything to SAR domain for semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2025).

Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. 2022. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *Advances in Neural Information Processing Systems* 35 (2022), 29845–29858.

Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. 2025. Medical SAM Adapter: Adapting Segment Anything Model for medical image segmentation. *Medical Image Analysis* 102 (2025), 103547.

Song Xia, Wenhan Yang, Yi Yu, Xun Lin, Henghui Ding, Lingyu Duan, and Xudong Jiang. 2024. Transferable adversarial attacks on sam and its downstream models. *Advances in Neural Information Processing Systems* 37 (2024), 87545–87568.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610* (2018).

Yang Yu, Chen Xu, and Kai Wang. 2024. TS-SAM: Fine-tuning segment-anything model for downstream tasks. In *Proceedings of the 2024 IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.

Chaoning Zhang, Fachrina Dewi Puspitasari, Sheng Zheng, Chenghao Li, Yu Qiao, Taegoo Kang, Xinru Shan, Chenshuang Zhang, Caiyan Qin, Francois Rameau, et al. 2023. A survey on Segment Anything Model (SAM): Vision foundation model meets prompt engineering. *arXiv preprint arXiv:2306.06211* (2023).

Kaidong Zhang and Dong Liu. 2023. Customized Segment Anything Model for medical image segmentation. *arXiv preprint arXiv:2304.13785* (2023).

Zhentong Zhang, Xinde Li, Pengfei Zhang, Wang Kui, Tianrong Gao, and Tao Shen. 2025. AttackTracer: Semantic-level adversarial attack location traceability via evidential diffusion model. *Neurocomputing* (2025), 131535.

Anqi Zhao, Tong Chu, Yahao Liu, Wen Li, Jingjing Li, and Lixin Duan. 2023. Minimizing maximum model discrepancy for transferable black-box targeted attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8153–8162.

Zihan Zhong, Zhiqiang Tang, Tong He, Haoyang Fang, and Chun Yuan. 2024. Convolution meets LoRA: Parameter efficient finetuning for Segment Anything Model. *arXiv preprint arXiv:2401.17868* (2024).

Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. 2023. Can sam segment polyps? *arXiv preprint arXiv:2304.07583* (2023).

Ziqi Zhou, Yifan Hu, Yufei Song, Zijing Li, Shengshan Hu, Leo Yu Zhang, Dezhong Yao, Long Zheng, and Hai Jin. 2026. Vanish into thin air: Cross-prompt universal adversarial attacks for sam2. *Advances in Neural Information Processing Systems* 38 (2026), 142839–142861.

Ziqi Zhou, Yufei Song, Minghui Li, Shengshan Hu, Xianlong Wang, Leo Yu Zhang, Dezhong Yao, and Hai Jin. 2024. Darksam: Fooling segment anything model to segment nothing. *Advances in Neural Information Processing Systems* 37 (2024), 49859–49880.

Yanfei Zhu, Yaochi Zhao, Zhuhua Hu, Tan Luo, and Like He. 2024. A review of black-box adversarial attacks on image classification. *Neurocomputing* (2024), 128512.

A Proof of Proposition

Proof. For a fixed $x' \in \mathcal{B}(x, \epsilon)$, we use the linearized feature-distance loss in Proposition 1. Expanding the squared norm gives

$$\begin{aligned}\widetilde{L}_g(x') = \big\|\mathbf{v}_0(x')\big\|_2^2 + 2\,\mathbf{v}_0(x')^\top J_\Delta(x')\Delta W_g \\ + \Delta W_g^\top J_\Delta(x')^\top J_\Delta(x')\Delta W_g.\end{aligned} \tag{12}$$

Since $\Delta W_g \sim \mathcal{N}(0, \sigma^2 I)$, we have $\mathbb{E}[\Delta W_g] = 0$ and $\mathbb{E}[\Delta W_g \Delta W_g^\top] = \sigma^2 I$. Thus, the cross term has zero expectation:

$$\mathbb{E}\big[2\,\mathbf{v}_0(x')^\top J_\Delta(x')\Delta W_g\big] = 0. \tag{13}$$

For the quadratic term, the trace identity yields

$$\begin{aligned}&\mathbb{E}\big[\Delta W_g^\top J_\Delta(x')^\top J_\Delta(x')\Delta W_g\big] \\ &\quad= \mathrm{tr}\big(J_\Delta(x')^\top J_\Delta(x')\,\mathbb{E}[\Delta W_g \Delta W_g^\top]\big) \\ &\quad= \sigma^2\,\mathrm{tr}\big(J_\Delta(x')^\top J_\Delta(x')\big) = \sigma^2\big\|J_\Delta(x')\big\|_F^2.\end{aligned} \tag{14}$$

Taking the expectation of Eq. (12) and combining these two identities proves Eq. (11). The equality concerns the linearized loss only; it does not bound the neglected terms in the original nonlinear feature-distance loss or imply a comparison for a fixed downstream target. □

B Comparison with Recent SAM-specific Attacks

In Table 4, we further compare our SETA method with two recent SAM-specific attack methods, DarkSAM (Zhou et al., 2024) and UAP-SAM2 (Zhou et al., 2026), on the four downstream SAMs. DarkSAM and UAP-SAM2 generate universal adversarial perturbations, while our SETA generates adversarial perturbations for individual images by exploiting the downstream fine-tuning structures. We report the results under the prompt-free downstream setting using the perturbation budgets and evaluation metrics in Table 1. As shown in Table 4, SETA achieves lower segmentation scores than both methods on SAMed, BC-SAM, and TS-SAM across all three perturbation budgets, indicating stronger attack performance. On Med-SAM, SETA achieves lower IoU at $\epsilon = 15/255$ and $20/255$, while its reported IoU at $\epsilon = 10/255$ is close to that of UAP-SAM2 (0.56 and 0.5585, respectively). These results further demonstrate the effectiveness of exploiting fine-tuning structures for adversarial transferability in the evaluated downstream setting.

Method	**SAMed (Synapse)**			**Med-SAM (Polyp)**			**BC-SAM (Blood Cell)**			**TS-SAM (COD)**		
	10	**15**	**20**	**10**	**15**	**20**	**10**	**15**	**20**	**10**	**15**	**20**
	mDSC ↓			IoU ↓			IoU ↓			S_α ↓		
DarkSAM	73.29	69.76	68.86	0.7222	0.6791	0.3760	0.9530	0.9448	0.9314	0.8316	0.8031	0.7865
UAP-SAM2	64.03	50.24	30.29	0.5585	0.3715	0.2403	0.9522	0.9447	0.9322	0.7491	0.5599	0.4860
SETA	17.8	5.5	2.7	0.56	0.28	0.19	0.91	0.88	0.85	0.37	0.34	0.33

Table 4: Comparison with recent SAM-specific attacks across downstream SAMs and datasets. The columns 10, 15, and 20 denote perturbation budgets $\epsilon = 10/255$, $15/255$, and $20/255$, respectively. Lower segmentation scores indicate stronger attacks. SETA results are reproduced from Table 1 at their original reporting precision.

# of k	SAMed (mDSC↓)	Med-SAM (IoU↓)	BC-SAM (IoU↓)	TS-SAM ($S_\alpha \downarrow$)
$k = 1$	27.3	0.58	0.93	0.37
$k = 2$	20.6	0.57	0.92	0.37
$k = 3$	17.8	0.56	0.91	0.37
$k = 4$	15.5	0.56	0.91	0.36
$k = 5$	15.5	0.55	0.92	0.36

Table 5: Impact of varying the number of ghost SAMs k for the SETA method on the attack performance.

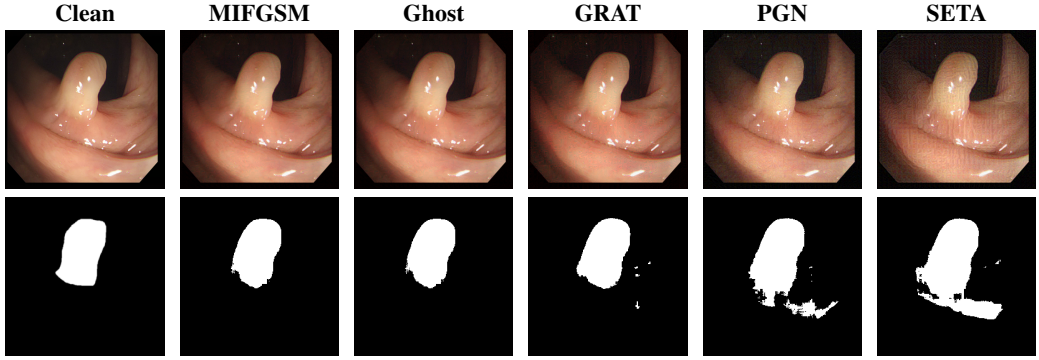


Figure 3: Comparison of different attack methods on the polyp dataset. The top row shows the input/perturbation images; the bottom row shows the predicted masks (or ground truth).

C Additional Ablation Study

Method	Time (s) on SAMed
SETA ($k = 1$)	0.55
SETA ($k = 2$)	1.12
SETA ($k = 3$)	1.62
SETA ($k = 4$)	2.17
SETA ($k = 5$)	2.70
MI-FGSM	0.11
Ghost Networks	1.85
PGN	1.12
UMI-GRAT	0.11

Table 6: The computational overhead on **SAMed**.

In Table 5, we demonstrate how the number of ghost SAMs used in each attack iteration affects the attack performance of our SETA method. Increasing the number of ghost SAMs may enhance the transferability of the adversarial examples, similar to how using multiple models during attack generation can improve transferability in standard adversarial settings.

We also report the computational overhead for each attack iteration of our SETA method, compared to other baseline approaches. For SETA, we vary the number of ghost SAMs used, since each ghost SAM requires both a forward pass and backpropagation during each attack iteration. As shown in Table 6, the computational overhead increases approximately linearly with the number of ghost SAMs. In contrast, methods like MI-FGSM and UMI-GRAT require only a single gradient update, resulting in significantly lower computational cost. Our SETA method, by leveraging multiple ghost SAMs, results in higher computation time. All timing results are measured on 512×512 input images using an NVIDIA A10 GPU and the SAMed model based on SAM-ViT-B.

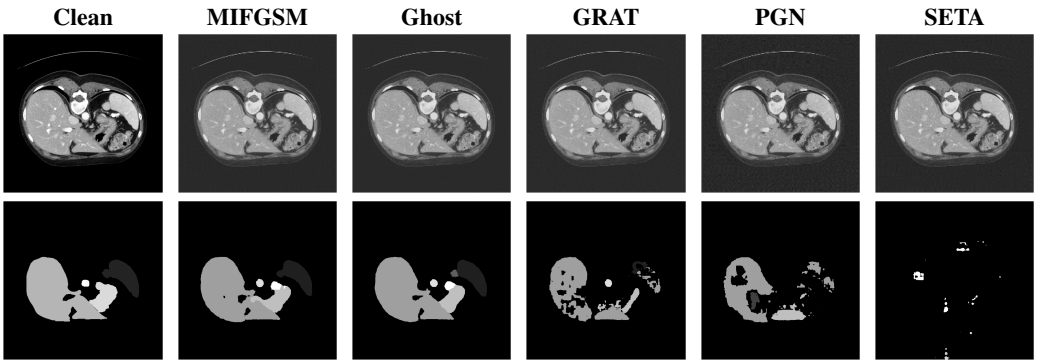


Figure 4: Comparison of different attack methods on Synapse dataset. The top row shows the input/perturbation images; the bottom row shows the predicted masks (or ground truth).

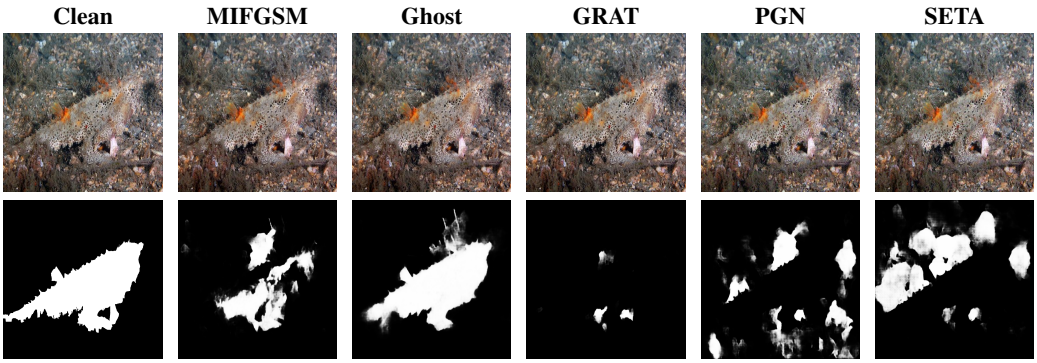


Figure 5: Comparison of different attack methods on COD10K dataset. The top row shows the input/perturbation images; the bottom row shows the predicted masks (or ground truth).

Our SETA requires the adversary to know the fine-tuning structure of the downstream SAM. To demonstrate that SETA can also enhance attack transferability when the adversary has only partial knowledge of the structure, we report the attack performance in Table 7 under the same attack setting as Table 1, with attack bound $\epsilon = 10/255$. The performance is evaluated when different numbers of fine-tuning blocks are correctly placed, while the remaining blocks are omitted from the downstream SAM. The correctly placed blocks are sequentially positioned from the input side of the image encoder toward the output side. The results in Table 7 show that the more fine-tuning blocks are correctly placed in the SETA ghost SAM, the better attack performance SETA achieves.

In Table 8, we compare the attack performance of SETA with and without the longitudinal ensemble. The longitudinal ensemble is designed to reduce computational overhead rather than to improve attack transferability. Without the longitudinal ensemble, we aggregate the gradients from all ghost SAMs at every iteration to generate adversarial examples. Although this yields better attack performance, it requires roughly $\frac{n}{k}$ times more computation, where n is the total number of ghost SAMs and k is the ghost SAMs sampled every iteration in the longitudinal ensemble.

We also implement our attacks and baselines on the AdapterShadow (Jie and Zhang, 2023). Table 11 compares the BER (Balance Error Rate) of different adversarial attacks on AdapterShadow under increasing perturbation budgets ϵ. A higher BER indicates a stronger attack, where

$$\text{BER} = \big(1 - 0.5(\text{TP}/N_p + \text{TN}/N_n)\big) \times 100,$$

with TP/N_p and TN/N_n denoting the true positive rates for shadow and non-shadow pixels, respectively. Across all budgets, SETA consistently achieves the highest BER, demonstrating the strongest attack effectiveness among the compared methods.

In Table 9, we report the attack performance when directly adding noise to the original SAM parameters using the same noise distribution adopted in SETA, as described in Table 1. We observe

Model	2 Blocks	4 Blocks	6 Blocks	8 Blocks	10 Blocks
SAMed (mDSC↓)	66.99	38.18	25.32	22.30	19.84
Med-SAM (IoU↓)	0.325	0.294	0.293	0.291	0.289
BC-SAM (IoU↓)	0.962	0.917	0.916	0.916	0.915
TS-SAM (S_α↓)	0.813	0.623	0.421	0.382	0.362

Table 7: Performance under different numbers of correctly placed blocks.

that directly perturbing the SAM parameters with such distribution noise does not achieve strong attack performance.

$\epsilon * 255$	**SAMed** (mDSC↓)	**Med-SAM** (IoU↓)	**BC-SAM** (IoU↓)	**TS-SAM** (S_α ↓)
10	17.8/9.4	0.56/0.46	0.91/0.91	0.37/0.35
15	5.5/2.8	0.28/0.24	0.88/0.88	0.34/0.33
20	2.7/1.6	0.19/0.12	0.85/0.85	0.33/0.31

Table 8: Attack performance of SETA with and without the longitudinal ensemble. Each entry reports the results with/without the ensemble.

$\epsilon * 255$	**SAMed** (mDSC↓)	**Med-SAM** (IoU↓)	**BC-SAM** (IoU↓)	**TS-SAM** (S_α ↓)
10	67.8	0.646	0.958	0.539
15	67.3	0.382	0.954	0.397
20	66.9	0.295	0.946	0.391

Table 9: The attack performance when adding noise to SAM parameters.

Table 10 reports the attack performance on ViT-H based Med-SAM and TS-SAM. Overall, SETA consistently achieves the strongest attack performance across both downstream SAM variants.

Table 11 further evaluates the attack performance on AdapterShadow (Jie and Zhang, 2023) using BER as the evaluation metric, where higher BER indicates stronger attack effectiveness. SETA consistently achieves the highest BER across all perturbation budgets compared with the baseline methods. In particular, at $\epsilon = 20/255$, SETA achieves a BER of 1.38, significantly outperforming PGN and UMI-GRAT. Moreover, the attack performance gap becomes larger as the perturbation budget increases. These results further demonstrate the effectiveness and transferability of SETA on downstream camouflaged object detection models.

D Visualization of the Generated Adversarial Images and Predictions

We show the generated adversarial images produced by different attack methods under a fixed perturbation bound, along with their corresponding predicted masks. The visualizations include randomly selected examples from the Synapse, Polyp, and COD10K datasets. As illustrated in Figure 3, Figure 4, and Figure 5, the results are consistent with the quantitative findings in Table 1, confirming that our SETA method achieves the most effective attack performance with its generated perturbation images.

E Details of hyperparameters

Table 12 provides the parameter settings for Table 3. The distributions of the ghost SAMs impact the attack performance of our SETA.

Method	Med-SAM (Polyp)			TS-SAM (COD)		
	IoU ↓			S_α ↓		
	10	15	20	10	15	20
No Attack	0.82			0.89		
MI-FGSM	0.72	0.68	0.63	0.62	0.55	0.52
Ghost Net	0.69	0.66	0.59	0.80	0.75	0.61
PGN	0.66	0.45	0.29	0.47	0.39	0.36
UMI-GRAT	0.67	0.58	0.52	0.53	0.47	0.46
SETA	0.60	0.34	0.24	0.44	0.40	0.36

Table 10: Attack performance on Med-SAM (ViT-H) and TS-SAM (ViT-H) under different perturbation budgets ϵ.

$\epsilon * 255$	**MI-FGSM** (BER↑)	**Ghost Net** (BER↑)	**PGN** (BER↑)	**UMI-GRAT** (BER↑)	**SETA** (BER↑)
10	0.64	0.63	0.68	0.67	0.69
15	0.64	0.66	0.70	0.71	0.81
20	0.65	0.73	0.83	0.85	1.38

Table 11: Attack performance on **AdapterShadow** (Jie and Zhang, 2023) across different attack methods. BER is reported for different perturbation budgets ϵ, with an average of 0.61 BER when there is no attack. The higher the BER, the better the attack performance.

Distribution Type and Description	**Parameter Value(s)**
Uniform $\mathcal{U}(-a, a)$	$a \in [0.0173,\ 0.2598]$
Laplace $\mathcal{L}(0, b)$	$b \in [0.0071, 0.1061]$
Truncated Gaussian $\mathcal{N}(0, \sigma^2)$ truncated at $\pm 2\sigma$	$\sigma \in [0.01, 0.15]$
Mixture of Gaussians: $0.5 \cdot \mathcal{N}(\mu_1, \sigma_1^2)$ + $0.5 \cdot \mathcal{N}(\mu_2, \sigma_2^2)$	$(\mu_1, \sigma_1) = (-0.1, 0.05)$ $(\mu_2, \sigma_2) = (0.1, 0.05)$

Table 12: Parameter settings used in Table 3.

F The Use of Large Language Models (LLMs)

We use LLMs to assist with writing tasks, including grammar checking and improving readability.

Main idea	SETA	Ghost Net	MI-FGSM / PGN	UMI-GRAT
Ensemble multiple SAM variants	✓	✓	×	×
Meta-initialization and gradient noise	×	×	×	✓
Stable gradient direction	✓	×	✓	✓

Table 13: Comparison of main ideas used by different transferable attacks.

Model	**Best-fit dist.**	**#Layers**	**Avg. mean**	**Avg. std**
TS-SAM	Gaussian	309	0.22400	0.08488
TS-SAM	Laplace	192	0.13000	0.03082
TS-SAM	Uniform	33	-0.01140	0.12060
SAMed	Gaussian	48	0.00019	0.09221
BC-SAM	Gaussian	36	-0.00004	0.03128
BC-SAM	Laplace	12	-0.00003	0.00659
Med-SAM	Gaussian	79	0.00010	0.03293
Med-SAM	Uniform	65	-0.05335	0.10648

Table 14: Parameters distribution summary across different medical SAM variants used in our experiments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are reflected in later sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in Sec. 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed not to penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the necessary assumptions and proof in Appendix Sec. A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the needed hyperparameter settings in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the code, and the link to the code can be found in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We show the details of experiments in Sec. 6 and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We run each experiment and report the average results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computer resource in Sec. 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that this work complies with the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper proposes an adversarial attack for improving the attack transferability of image segmentation models under adversarial settings. We do not anticipate any direct negative societal impacts arising from this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our proposed framework does not have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use publicly available datasets, models, and research codebases. All assets are properly cited, and we follow their respective licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper declares the usage of LLM in the appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.