Practical Region-level Attack against Segment Anything Models

Yifan Shen* Zhengyuan Li* Gang Wang
University of Illinois Urbana-Champaign
{yifan26, zli138, gangw}@illinois.edu

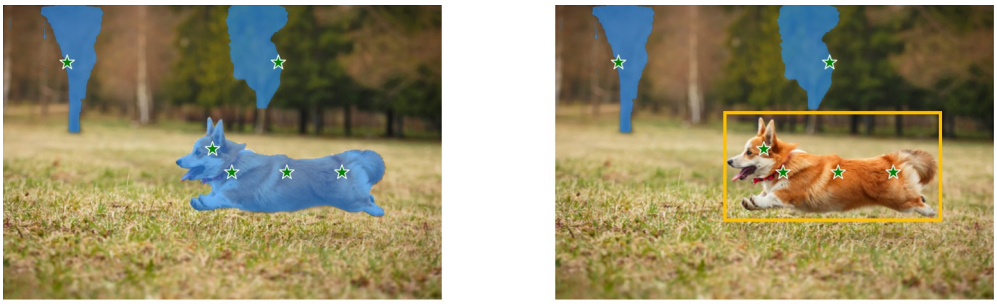

Figure 1. **Region-level Attack on a Segment Anything Model (SAM).** The left image shows the original *clean* image—objects are well segmented when a user clicks on the object region (user clicks are denoted by green stars). The right image shows the *attacked* image—the corgi in the yellow box (attack-target region) can no longer be identified by SAM no matter where the user clicks within the box. Note that, the regions outside of the yellow box in the image are not affected by the attack.

Abstract

Segment Anything Models (SAM) have made significant advancements in image segmentation, allowing users to segment target portions of an image with a single click (i.e., user prompt). Given its broad applications, the robustness of SAM against adversarial attacks is a critical concern. While recent works have explored adversarial attacks against a pre-defined prompt/click, their threat model is not yet realistic: (1) they often assume the user-click position is known to the attacker (point-based attack), and (2) they often operate under a white-box setting with limited transferability. In this paper, we propose a more practical region-level attack where attackers do not need to know the precise user prompt. The attack remains effective as the user clicks on any point on the target object in the image, hiding the object from SAM. Also, by adapting a spectrum transformation method, we make the attack more transferable under a black-box setting. Both control experiments and testing against real-world SAM services confirm its effectiveness.

*Equal Contribution

1. Introduction

Segment Anything Models (SAM) leverage foundational models for *promtable* image segmentation [16], which has shown outstanding performance. SAM can delineate objects of interest into masks based on user prompts (e.g., point of user clicks). As SAM is used for mission-critical applications such as healthcare image analysis [22, 24] and scene understanding for autonomous driving [48], the robustness of SAM (against adversarial attacks) raises concerns. For example, adversaries may manipulate the images/videos taken by autonomous vehicles to hide key objects (e.g., traffic signs, vehicles) from the image segmentation module, posting threats to driving safety.

While recent work has explored adversarial attacks against SAM [44, 50], their threat models are not yet realistic. For instance, Attack-SAM [44] proposes a white-box attack to hide an object in the image from being segmented by SAM. However, Attack-SAM assumes the precise *point location* where the user clicks and the *model parameters* are both known to the attacker. Sheng *et al*. [50] investigated another attack (TAA) under a different threat model. Their goal is to mislead SAM to output a mask of attacker-specified shape—instead of hiding the object from SAM (which is our main focus). Their results also show that this

is a challenging problem: while TAA has some transferability, the attack effect is majorly weakened after transferring.

In this paper, we introduce a region-level attack to explore a more practical threat model (see Fig. 1). The attacker's goal is to conceal the object within an attacker-specified region from SAM's segmentation. In this case, the attacker does not need to know the precise point of the click of the user—no matter which point in the region is clicked by the user, the object cannot be accurately segmented by SAM. In addition, we investigate to improve the transferability of the attack such that it can operate under a *black-box* setting. As shown in Fig. 1, after applying the adversarial perturbation to the image, clicking on any point in the yellow box (attacker-specified region) will no longer separate the corgi from the rest of the image.

Under this threat model, we first develop a Sampling-based Region Attack (S-RA), a basic method for region-level adversarial attacks, and then improve its transferability with a Transferable Region Attack (T-RA). Our design is based on two key intuitions. First, sparsely sampled points in the region can constitute a surrogate target of all pixels in the region. Second, even when the involved region goes beyond a single point, adding perturbations in the frequency domain when attacking the surrogate model can improve the transferrability [20]. Therefore, our method first applies spectrum transformation to the image in order to simulate the spectrum saliency map [20] of the victim model. Then it estimates the optimization target with evenly sampled points in the region and conducts the optimization with a PGD attack [23] to generate adversarial noises.

We evaluated the proposed attacks on multiple SAM variants including ViT-B, ViT-H and ViT-L [16] and demonstrated the effectiveness of the attacks under both white-box and black-box settings. We extensively evaluated the attack transferability to a variety of SAM architectures including EfficientSAM (S and Ti) [40], Fast-SAM (S and X)[47], MobileSAM [43], and HQ-SAM (B, L, and H) [15]. We also confirm the effectiveness of the attack (optimized with a local ViT-B) against a real-world SAM service. Our result highlights the realism of the risk and calls for new defense methods to improve the robustness of SAM.

Our contributions are summarized as follows.

- We present a region-level attack against SAM, a more practical threat model where attackers do not need to know the precise user prompt.
- We designed novel attack methods, Sampling-based Region Attack (S-RA) and Transferable Region Attack (T-RA), that undermine SAM's segmentation ability under both white-box and black-box settings.
- Extensive experiments demonstrate that S-RA and T-RA can successfully attack the original SAM and its variants.

2. Related Works

2.1. Adversarial Attacks

Deep neural networks are known to be susceptible to adversarial examples, which are samples that appear indistinguishable from genuine ones to the human eye but can mislead models into producing incorrect outputs [2, 31, 51].

Attacks are manifested in two settings: white-box and black-box. In the white-box setting, attackers can access all model knowledge, including architecture, parameters, and gradients. This setting is often used to assess model robustness rather than actual attacks [3, 10, 23]. White-box attacks, such as Fast Gradient Sign Method (FGSM) [10] and projected gradient descent (PGD) [23], allow full visibility into the target model to generate adversarial examples. In contrast, black-box attacks operate under limited knowledge [12, 36, 39, 42]. Notable techniques include updating gradients with momentum (MI-FGSM) [7], smoothing gradients with a kernel (TI-FGSM) [8], and resizing adversarial examples for input diversity (DI-FGSM) [39]. These methods are grounded in the principle of the transferability of adversarial examples exploiting vulnerabilities inherent across multiple models without specific insights into the target model's internals [42, 49].

2.2. Segment Anything Model and Variants

In the domain of image segmentation, major advancement has been made by the "Segment Anything Models" (SAM) [16]. SAM uses foundational models and capitalizes on the principles established by prior works, underscoring the importance of multi-scale features and iterative refinement for segmentation [1, 28, 46]. SAM's versatility is further explored through its application across diverse contexts, for example, medical image segmentation [21], detection of camouflaged entities [32], and semantic communication challenges [33]. SAM has been applied to both 2D and 3D environments, highlighting its potential in semantic labeling [4], object tracking [41, 45], and 3D object segmentation [6, 30].

In recent variants of SAM, SEEM [52] allows users to segment images using various "prompts", including points, markers, boxes, scribbles, text, and audio. HQ-SAM [15] enhances the ability to accurately segment any object. Semantic-SAM [17] emerges as a universal image segmentation model that enables segmentation and recognition at any desired granularity. For improved efficiency, MobileSAM [43] introduces object-aware prompt sampling, replacing the grid-search prompt sampling in the original SAM, to expedite the segmentation process. EfficientSAM [40] leverages Masked Image Pretraining to improve segmentation efficiency. Fast Segment Anything [47] speeds up the original SAM model by 50x.

Prior works have assessed the robustness of *traditional*

segmentation models against adversarial attacks [9, 14, 26, 35, 38]. More recent works have explored the problem in the context of SAM [29, 44], using *imperceivable* adversarial perturbations. Yu *et al*. [27] introduce an attack that produces *visible* image corruptions such as style changes, occlusions, and local patch attacks. However, as discussed in Sec. 1, the existing SAM attacks' threat model is not yet realistic by assuming knowing precise user prompt (under a white-box setting), and we aim to improve the realism of the attack with a region-based attack (black-box setting).

3. Preliminary

Segment Anything Model (SAM). SAM introduces a novel, promptable segmentation framework to generate precise masks for a given image and a prompt. While the original SAM [16] supports points, text, and boxes as prompts, in this work, we primarily focus on the point prompt scenario (similar to [44, 50]), leaving other types of prompts for future exploration. At its core, SAM comprises three key components: an image encoder, a prompt encoder, and a mask decoder. The image encoder leverages a Vision Transformer (ViT) architecture, pretrained using the Masked Autoencoder (MAE) [13], to extract feature representations from input images. The prompt encoder employs positional embeddings to encode prompts. The mask decoder synthesizes the outputs from both encoders to predict segmentation masks, thus determining the segmented object based on the synergy between the image and the prompt. The mask prediction process in SAM is defined as follows:

$$y = \mathit{SAM}(p, x; \theta) \tag{1}$$

where p and x denote the input prompt and image, respectively, and θ symbolizes the model's parameters. Given an image $x \in \mathbb{R}^{H \times W}$, the output y mirrors the input image's dimensions, with H, and W representing the height, and width, respectively. The pixel coordinates within image x are denoted by i and j. The mask region is delineated by the predicted values y_{ij} where values exceeding a defined threshold (e.g., 0) are classified as part of the segmented object. During inference, the final binary mask is obtained as follows, where *sign* denotes the sign function.

$$M_{pred} = \mathit{sign}(y) \tag{2}$$

Projected Gradient Descent Attack. Projected gradient descent (PGD) attack [23] is a popular adversarial attack method. It utilizes the first-order gradient and iteratively finds the solution to the optimization problem within the allowed perturbation set. The algorithm can be concisely represented as follows:

$$x^{(t+1)} = \mathit{Clip}_{x,\epsilon}\left(x^{(t)} + \alpha \cdot \mathit{sign}(\nabla_x L(\theta, x^{(t)}, y))\right) \tag{3}$$

where $x^{(t)}$ is the adversarial example at iteration t, α is the step size, L is the loss function defined by the specific task, θ represents the model parameters, and y is the ground-truth label. The function $\mathit{Clip}_{x,\epsilon}$ ensures that the perturbed image remains within an ϵ-neighborhood of the original image, which means $\|\delta\|_\infty < \epsilon$. In other words, ϵ controls the magnitude of the adversarial perturbation. The method is untargeted (i.e., the adversarial example aims to obtain any other labels that are not the ground-truth label y). It serves as a fundamental building block of our attack method.

4. Method

In this section, we introduce the threat model and our detailed attack method. We start with the basic point-level attack setting and then build the idea of region-level attacks on top of it, and then use spectrum transformation to further improve the transferability.

4.1. Threat Movel and Attack Overview

Point-level Attack. Point-level attack [44] was previously proposed to attack SAM, assuming the attacker knows the user prompt (i.e., click point). The goal of the attack is to conceal a target object, which is formulated as minimizing the prediction values of the SAM-generated mask y. Formally, given an image x and a set of permissible adversarial perturbations, S, constrained by predefined attack strength parameters, the goal is to compute the adversarial perturbation δ that obliterates the mask when the model is prompted with a specific point (p). The loss function is defined as:

$$L(x, p) = \|\mathit{Clip}\left(\mathit{SAM}(p, x+\delta; \theta), \right. \\ \left. \mathit{min} = Neg_{th}\right) - Neg_{th}\|^2, \tag{4}$$

where θ is the parameters of the target model, and Neg_{th} is a negative hyperparameter threshold used by SAM to segment an object (see Sec. 3). Following [44], we set $Neg_{th}$to -10, as the non-mask regions often have values around this threshold. The point-level threat model is the following optimization problem:

$$\delta^* = \arg\min_{\delta \in S} L(x, p) \tag{5}$$

where δ^* represents the optimal value of δ, which is the perturbation that achieves the best adversarial effect. The application of the clip function is strategic, preventing the predictive values from becoming excessively negative, which could inadvertently impede the optimization process. This ensures that predicted values, $\mathit{SAM}(p, x+\delta)$, are coerced towards being less than or equal to Neg_{th}, with clipping applied to maintain values above this threshold.

Region-level Attack. Region-level attack allows the user to specify a region R where SAM should fail to segment the objects regardless of the user prompts. Formally,

$$\delta^* = \arg\min_{\delta \in S} \mathbb{E}_{p \sim \text{uniform}(R)}[L(x, p)] \tag{6}$$

which means for a point uniformly sampled from R, we minimize the expectation of SAM's segmentation mask's area. In this work, we assume that R is a rectangle that covers the object specified by the attacker.

4.2. Sampling-based Region Attack (S-RA)

First, we discuss a *white-box* attack. Directly optimizing Eq. (6) is computationally intensive due to the large number of pixels in an image. Alternatively, we sample points in the region and create the substitute loss function. We uniformly select points by partitioning the R into a grid where m points are chosen along the horizontal axis and n points along the vertical axis, resulting in a total of $m \times n$ points. The loss function for the point set is

$$L_{SRA}(x, P) = \frac{1}{m \times n} \sum_{p \in P} L(x, p) \tag{7}$$

Compared with random sampling used in previous work [50], this structured selection process ensures comprehensive coverage of the targeted region. Each point within this grid is subsequently targeted with the point attack strategy.

In the later experiments (Sec. 5), we will evaluate the attack effectiveness by *randomly* selecting a point within the region R as a prompt to examine the segmentation result. Note that, the newly selected point during testing time is not necessarily (unlikely) among these sampled points used for attack optimization, due to the sparsity of sampling.

4.3. Transferable Region Attack (T-RA)

Under the *black-box* setting, attackers need to compute adversarial perturbations based on a local substitute model and then apply the perturbation to attack a different target model. The above sample-based region attack (S-RA) has shown limited transferability (see Sec. 5), which motivates us to improve it for black-box attacks. More specifically, we introduce a transferable region attack (T-RA) by adapting spectrum transformation (ST) [20]. While spectrum transformation was initially devised to improve adversarial attacks targeting *image classifiers*, we have discovered that its effectiveness extends to attacking SAM variants as well.

To improve transferability, model augmentation [19] utilizes loss-preserving transformations on the image to avoid the adversarial attack overfitting the current model. Spectrum Transformation (ST) [20] is a form of model augmentation that perturbs the image in the frequency domain. The intuition is the following: regions of high and low frequency

Algorithm 1 Transferable Region Attack (T-RA)

Input: SAM model f, image $\boldsymbol{x}$, sampled points P in the region, original segmentation y, perturbation limit ϵ, negative threshold neg_th, number of steps N, number of spectrum transformed samples M, PGD attack step size α, spectrum transformation hyperparameters ρ and η
Output: Adversarial image $\boldsymbol{x}'$

procedure T-RA
  $\mathcal{L}_{\text{best}} \leftarrow \infty$
  $\boldsymbol{x}' \leftarrow \boldsymbol{x}$
  $\delta \leftarrow 0$
  for $step = 1$ To N **do**
    $\delta_{sum} \leftarrow 0$
    for $i = 1$ To M **do**
      $\boldsymbol{x}_1 \leftarrow \text{ST}(x, \rho, \eta) + \delta$
      $L \leftarrow L_{SRA}(x_1, P)$
      $\delta_{temp} \leftarrow \text{sign}(\frac{\partial L}{\partial x_1}) * \alpha$
      $\delta_{sum} \leftarrow \delta_{sum} + \delta_{temp}$
    end for
    $\delta \leftarrow \delta_{sum}/M + \delta$
    $x' \leftarrow \text{Clip}(x + \delta, x - \epsilon, x + \epsilon)$
  end for
  return $\boldsymbol{x}'$
end procedure

in the image correspond to areas of significant and minor pixel variations respectively. High-frequency areas often represent edges and textures, indicative of rapid pixel intensity changes, while low-frequency areas denote smoother, homogeneous regions that often encompass entire objects. Different SAM variants depend on different frequency domains of interest to make predictions. By manipulating the spectrum of the image, the idea is to simulate and exploit feature variations of different victim models to enhance the transferability of adversarial attacks. The Discrete Cosine Transform (DCT) and inverse Discrete Cosine Transform (iDCT) are utilized to transform the image back and forth in the spatial and spectrum space. The transformation is formalized as follows:

$$ST(x, \rho, \eta) = iDCT(DCT(x + \eta) \odot M(\rho)), \tag{8}$$

where x denotes the original image, η is a noise vector drawn from a normal distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$, and $M(\rho)$ a mask with elements sampled from a uniform distribution $\mathcal{U}(1-\rho, 1+\rho)$. The operation $\odot$ represents element-wise multiplication. Note that ρ controls the strength of perturbation. When ρ is too high, the resulting image may not preserve the semantics of the original image; in contrast, when ρ is too low, the adversarial example loses transferability.

We detail the implementation of our T-RA attack in Algorithm 1. The algorithm takes in a set of parameters including the target model f, an original image $\boldsymbol{x}$, and a pre-

defined attack region R. The process iterates over a predefined number of steps (N) and a predefined number of spectrum-transformed samples (M), dynamically adjusting the perturbation δ to minimize the loss L, thereby maximizing the adversarial effect. The parameter ϵ represents the maximum allowable change for each pixel value in the image, ensuring that the perturbations remain imperceptible to the human eye. The algorithm applies spectrum simulations, transforming the image $\boldsymbol{x}$ into $\boldsymbol{x}_1$ (line 8), for improving the transferability of the final adversarial example $\boldsymbol{x}'$. Following [44], we use PGD for attack optimization.

5. Evaluation

5.1. Experimental Setup

Our experiments are conducted primarily using two variants of the SAM models [16]: ViT-B (91M parameters), and ViT-H (636M parameters). A third ViT-L model (308M parameters) will be used only for selective experiments. More specifically, ViT-B will be used for white-box evaluation. Then for the the black-box evaluation, we run transferred attacks from the smaller ViT-B model to the larger ViT-H model. Finally, in Sec. 6, we further explore the attack transferability to four more SAM variants.

We evaluate both the sample-based region attack (S-RA) and practical region attack (T-RA). We begin by defining a target region within the image and then test *clean images* with both ViT-B and ViT-H models. We set the width and height of regions to be one-third of the original image. On the clean image, we perform click-based segmentation on a randomly selected point (p) within the predefined region, resulting in a segmentation mask termed $Mask_{clean}$. Subsequently, we apply the attack method (either S-RA or T-RA) to attack the images and retest them at the same point (p) in both ViT-B and ViT-H models. This attacked image yields another segmentation mask $Mask_{adv}$. A comparative analysis of $Mask_{clean}$ and $Mask_{adv}$ is conducted to assess the efficacy of the attack.

Baseline. Regarding the baseline, given our attack has a novel threat model, we *adapt* Attack-SAM [44] to our threat model. More specifically, we run Attack-SAM to attack the *center of the region*, and then during testing time, the attack is evaluated on a randomly sampled point in the region (following the same region-level attack protocol). We acknowledge that Attack-SAM is not designed for region-level attack—the purpose of the experiment is to show whether the attack optimized for a pre-defined prompt can generalize to other (nearby) points in the region.

Dataset. For our evaluation dataset, we randomly select 200 images from the SA-1B dataset [16]. To induce perturbations in the images, we constrain the magnitude of adversarial perturbation by setting the value of ϵ to four distinct levels: 2/255, 4/255, 8/255, and 16/255. These values serve as upper bounds for the perturbation, ensuring controlled and quantifiable levels of adversarial noise. The experiments are conducted on NVIDIA A100 GPU to systematically assess the robustness of the SAM model against varying degrees of constrained perturbations.

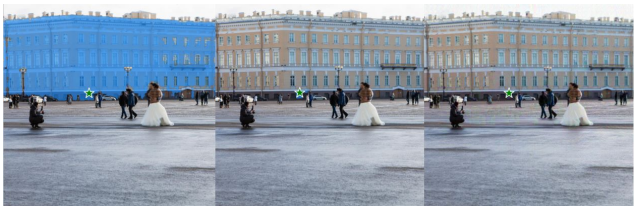

Figure 2. Image segmentation results under different attack methods on the ViT-B model. The left image is the original clean image. The middle image is attacked by S-RA and the right image is attached by T-RA. The attack strength is $\epsilon = 8/255$.

ϵ	2/255	4/255	8/255	16/255
AttackSAM [44]	20.56	10.48	4.28	3.69
S-RA	**2.99**	**1.75**	**1.52**	**1.27**

Table 1. mIoU (%) of the white-box experiment on the ViT-B model under the S-RA and AttackSAM [44] with varying attack strengths (ϵ).

Evaluation Metrics. Following [5, 44], we use the mean Intersection over Union (mIoU) as our primary evaluation metric. mIoU is a commonly used metric for evaluating image segmentation. It measures the overlap between the predicted and the ground-truth segmentation masks. Then it takes the mean of the IoU values across all test samples. The IoU for a single sample is the ratio of the intersection of the predicted segmentation mask $Mask_{adv}$ and the ground truth mask $Mask_{clean}$ to their union. Mathematically, the mIoU is expressed as:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU}\left(\text{Mask}_{adv}^{(i)},\ \text{Mask}_{\text{clean}}^{(i)}\right) \tag{9}$$

where N is the total number of samples in the test set. The value of mIoU ranges from 0 to 1, with a lower value indicating a more effective attack.

5.2. Qualitative and Quantitative Results

We conduct the evaluation under white-box and black-box settings, respectively. For the white-box setting, we run the sample-based region attack (S-RA) on the ViT-B model, allowing full access to the model's architecture and parameters (in comparison with the baseline AttackSAM). We confirm the attack is highly successful. As shown in Tab. 1, a subtle perturbation ($\epsilon = 2/255$) from S-RA can already

Figure 3. Visualization of white-box and black-box attack results (attack strength ϵ =8/255). The first row shows the original clean images segmented using the ViT-B model. The second row shows S-RA attack (white-box) trained on ViT-B model and the segmentation results on the same ViT-B model. The result confirms the effectiveness of S-RA attack under a white-box setting. The third row shows S-RA attack (black-box) trained on ViT-B model and the segmentation results on a different ViT-H model. The result shows the lack of transferability of S-RA under a black-box setting. The fourth row shows T-RA attack (black-box) trained on ViT-B model and the segmentation results on a different ViT-H model. The result shows T-RA transfer well and ViT-H cannot segment correctly under this attack.

effectively remove most of the mask, hiding the target object from SAM. This is reflected by the minimal overlap between the generated mask and the ground-truth mask (mIoU=2.99%). Comparing with the baseline, we show S-RA outperforms the adapted AttackSAM [44] across all attack strengths. Fig. 2 shows example attack images from this experiment.

Next, we evaluate the attacks under the black-box setting: the adversarial examples are first computed with the ViT-B model and then used to attack a more complicated ViT-H model. The results are shown in Tab. 2 and example images are shown in Fig. 3. The result first confirms the lack of transferability of the basic S-RA for black-box attacks (with high mIoU ranging from 31.64% to 46.32%). Similarly, the baseline AttackSam [44] also does not transfer well. Then we show the improved transferability of the T-RA strategy with much lower mIoU. For example, when attack strength is ϵ =8/255, mIoU is below 10%. The result confirms the effectiveness of T-RA for black-box attacks.

ϵ	2/255	4/255	8/255	16/255
AttackSAM [44]	59.58	48.49	43.95	32.47
S-RA	46.32	43.32	40.75	31.64
T-RA (ρ=0.1)	**45.01**	**17.70**	**9.34**	**11.43**
T-RA (ρ=0.3)	54.91	31.15	9.84	10.16

Table 2. mIoU (%) result of the AttackSAM [44], S-RA and T-RA on under black-box settings (trained on ViT-B, tested on ViT-H). For S-RA, the attack is not as strong as the white-box attack (see Tab. 1). Comparing S-RA and T-RA under the black-box setting, T-RA has a stronger attack result with an overall lower mIoU under various settings (especially when $\rho = 0.1$). This confirms the effect of spectrum transformation of T-RA.

ρ	0.01	0.05	0.1	0.2	0.3
mIoU (%)	19.32	18.31	**17.70**	22.98	30.51

Table 3. mIoU of varying the ρ parameter in the T-RA under the black-box setting (trained on ViT-B; tested on ViT-H), with a fixed attack strength $\epsilon = 4/255$. The results indicate that a ρ value of 0.1 yields the most effective attack, achieving the lowest mIoU.

5.3. Ablation Study

ρ of the T-RA. We conduct an ablation study on the ρ parameter of the T-RA, assessing its impact on the effectiveness of adversarial attacks (trained on ViT-B; tested on ViT-H). The study systematically varies ρ across a set of values: 0.01, 0.05, 0.1, 0.2, and 0.3, while maintaining a constant ϵ value of 4/255. This investigation allows us to discern the optimal range for ρ. The mIoU percentage results for different ρ values in the T-RA are presented in Tab. 3. Additionally, Fig. 4 illustrates the visual differences in the adversarial examples generated with varying ρ values. The result shows that the attack generally works well under these ρ values, and the best-performing value is 0.1.

Figure 4. Visualization of the segmentation results under different ρ values for the T-RA under a black-box setting (trained ViT-B; tested on the ViT-H). The attack strength is fixed as $\epsilon = 4/255$. The first column shows the original clean image's segmentation result. The second column is a zoomed-in view to highlight the mask on the clean image. The subsequent columns display the segmentation results for ρ values of 0.01, 0.1, 0.2, and 0.3, respectively. We find that $\rho = 0.1$ resulting in the most effective degradation of segmentation accuracy.

ρ, ϵ (/255)		0.1, 4	0.1, 8	0.3, 4	0.3, 8
ViT-L	mIoU (%)	13.37	**4.40**	26.80	6.34
ViT-H	mIoU (%)	17.70	**9.33**	30.51	9.84

Table 4. mIoU (%) result of the effectiveness T-RA under a black-box setting under various ϵ and ρ settings. The attack is trained on ViT-B and tested on ViT-L and ViT-H, respectively. Bold values highlight the most successful attack configurations. Note that the ϵ values are represented as "4" and "8" for brevity, but they correspond to $4/255$ and $8/255$, respectively.

T-RA on ViT-L and ViT-H. We then extend our investigation of the T-RA to explore its transferability across other ViT models from SAM [16], namely the 308M parameters ViT-L and the 636M parameters ViT-H. Given the substantial variance in model size and complexity, from the 91M parameters ViT-B to these larger architectures, it is imperative to assess the robustness of our adversarial strategy. We conduct experiments with ϵ values of 8/255 and 16/255, and ρ parameters set to 0.1 and 0.3. We show the results in Tab. 4, with the impact of ϵ and ρ settings on the mIoU across the ViT-L and ViT-H models. The results confirm the transferability of both models. Also, the transferability is consistently higher with larger perturbations (ϵ = 8/255).

Density of Attack Points. Recall that our attack samples attack points from the target region to compute adversarial examples. Here, we focus on T-RA and investigate the impact of the attack point density on the attack effectiveness. We define the density using a parameter λ, which represents the number of pixels between consecutive attack points sampled in both horizontal and vertical directions. Recall that we sample $m \times n$ attack points using a grid. Formally, $m = \frac{W}{\lambda}$ and $n = \frac{H}{\lambda}$. In our experiments, we test different values of λ: 50, 60, 70, and 80 pixels. This means that for a given λ, an attack point is placed every λ pixels along both axes, forming a grid-like pattern of attack points within the region. The experiments are carried out using images with $\epsilon = 8/255$ and $\rho = 0.1$ to assess the impact of λ on the mIoU metric. The results are presented in Tab. 5. As expected, a higher density of sampled points (i.e., lower λ) leads to a more effective attack.

λ	50	60	70	80
mIoU (%)	8.97	10.65	11.50	12.69

Table 5. Impact of varying attack point density on the mIoU metric for the T-RA attack. A lower λ represents a higher point density.

6. Cross Model Transferability

In this section, we further evaluate the transferability of our adversarial examples with a broader set of SAM variants with different architectures, under the black-box setting. The adversarial examples are generated with $\epsilon = 8/255$ and $\rho = 0.1$ using T-RA, all trained on ViT-B. The testing SAM models include EfficientSAM (S and Ti) [40], Fast-SAM (S and X)[47], MobileSAM [43], and HQ-SAM (B, L, and H) [15]. These testing models vary in architecture and complexity, from lightweight versions like EfficientSAM (Ti) to

ϵ	2/255	4/255	8/255	16/255
EfficientSAM (S) [40]	11.80	11.50	11.62	**11.38**
EfficientSAM (Ti) [40]	8.06	7.92	7.75	**7.13**
Fast-SAM (S) [47]	40.06	35.14	25.74	**21.42**
Fast-SAM (X) [47]	67.51	60.15	55.38	**37.78**
MobileSAM [43]	72.92	**52.65**	57.01	58.40
HQ-SAM (B) [15]	51.58	21.48	2.61	**0.01**
HQ-SAM (L) [15]	69.25	36.25	**22.60**	31.47
HQ-SAM (H) [15]	71.01	42.11	**29.50**	37.10

Table 6. mIoU (%) of transferred attack T-RA with different attack strength ϵ and a fixed $\rho = 0.1$ on different SAM variants. All attacks are trained by the ViT-B model, and then tested on each of the listed SAM variants under the black-box setting. The lowest mIoU value (i.e., the most successful attack) for each model is highlighted in bold.

more robust versions like HQ-SAM (H).

Tab. 6 shows the mean Intersection over Union scores at different attack strengths. The results confirm the overall transferability of our attack on these SAM variants. Notably, HQ-SAM (B) exhibits a major drop in mIoU at $\epsilon = 8/255$, indicating a high level of susceptibility to our attack. In contrast, models like MobileSAM maintain higher mIoU scores, demonstrating some level of resilience to stronger attacks. In most models, the segmentation capability weakens as the attack strength increases. However, for HQ-SAM, the weakest segmentation capability is observed at $\epsilon = 8/255$ in two experimental groups. This is likely due to the architectural similarities between HQ-SAM and the original SAM model (see its pattern in Tab. 2), leading to a similar pattern of vulnerability at this attack strength.

Testing on Real-world SAM Services. We tested our adversarial example on a real-world SAM Service[1], under a black-box setting. Due to limitations in testing through their web interface, a comprehensive quantitative evaluation was not feasible. Instead, we hand-picked a few images, generated adversarial examples with $\epsilon = 8/255$ and $\rho = 0.1$ on the ViT-B model and uploaded the images to the SAM service to manually inspect the result. We observed that online services are indeed more robust (against transferred attacks): they may have implemented countermeasures or image preprocessing steps. For example, not all the points in the attack region can successfully trigger the adversarial effect. We still found some successful images (under the "point click" prompt). Examples are shown in Fig. 5. The experiment was conducted *ethically*: these images were uploaded to our personal account and immediately deleted after the tests (i.e., not affecting other users or the service).

[1]Meta AI SAM: https://segment-anything.com/

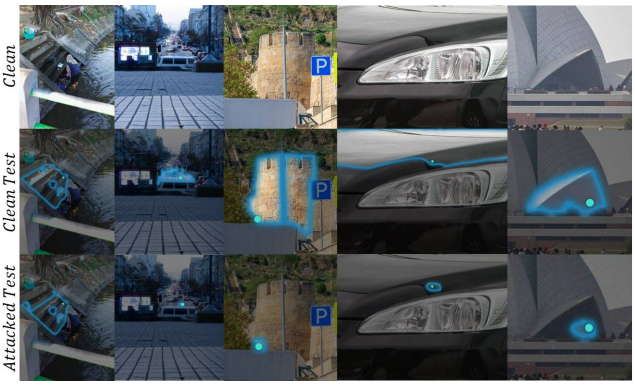

Figure 5. Visualization of adversarial examples with $\epsilon = 8/255$ and $\rho = 0.1$, computed on ViT-B and tested on a real-world SAM service. The blue dot is the test point and the highlighted area is the output mask.

7. Conclusion and Future Work

In this paper, we introduce a more practical region-level adversarial attack against Segment Anything Models (SAM). We show that the proposed methods can effectively generate *transferable* adversarial examples that compromise SAM's segmentation ability within attacker-defined regions. Through extensive experiments, we demonstrate the feasibility of region-level attacks in both white-box and black-box settings and confirm the effectiveness of attacks on multiple SAM variants. The result calls for more robust SAM models to withstand such adversarial threats.

Defense and Future Work. Potential defense mechanisms to enhance the robustness of SAM models include applying adversarial training techniques [23, 34], the use of input transformation methods to reduce the effectiveness of adversarial perturbations [11, 18, 37], and the exploration of novel SAM architectures that are inherently more resistant to adversarial manipulation [25]. Investigating these defenses and their effectiveness in mitigating the threats posed by adversarial attacks on SAM models is an important area for future research.

In this paper, we primarily focus on SAM and its variants that support point prompts. The transferability and effectiveness of our approach on other *segmentation models* and other *prompt types* remain to be explored, which can be a potential direction for future work. In addition, the real-world impact of our attack may be further influenced by factors such as image quality, image preprocessing, and the presence of countermeasures.

Acknowledgments. This work was supported in part by NSF grants 2055233 and 2229876. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2

[2] Battista Biggio, Giorgio Fumera, and Fabio Roli. Security evaluation of pattern classifiers under attack. *IEEE transactions on knowledge and data engineering*, 26(4):984–996, 2013. 2

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 2

[4] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. https://github.com/fudan-zvg/Semantic-Segment-Anything, 2023. 2

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5

[6] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2

[8] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 2

[9] Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox. Adversarial examples for semantic image segmentation. *arXiv preprint arXiv:1703.01101*, 2017. 3

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[11] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 8

[12] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. *Advances in neural information processing systems*, 33:85–95, 2020. 2

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[14] Xu Kang, Bin Song, Xiaojiang Du, and Mohsen Guizani. Adversarial attacks for image segmentation on multiple lightweight models. *IEEE Access*, 8:31359–31370, 2020. 3

[15] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 2, 7, 8

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3, 5, 7

[17] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 2

[18] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1778–1787, 2018. 8

[19] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019. 4

[20] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *European Conference on Computer Vision*, pages 549–566. Springer, 2022. 2, 4

[21] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[22] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 3, 8

[24] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 1

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9078–9086, 2019. 8

[26] Utku Ozbulak, Arnout Van Messem, and Wesley De Neve. Impact of adversarial examples on deep learning models for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 300–308. Springer, 2019. 3

[27] Yu Qiao, Chaoning Zhang, Taegoo Kang, Donghun Kim, Shehbaz Tariq, Chenshuang Zhang, and Choong Seon Hong. Robustness of sam: Segment anything under corruptions and beyond. *arXiv preprint arXiv:2306.07713*, 2023. 3

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted*

Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015. 2

[29] Jérôme Rony, Jean-Christophe Pesquet, and Ismail Ben Ayed. Proximal splitting adversarial attack for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20524–20533, 2023. 3

[30] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild, 2023. 2

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[32] Lv Tang, Haoke Xiao, and Bo Li. Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709*, 2023. 2

[33] Shehbaz Tariq, Brian Estadimas Arfeto, Chaoning Zhang, and Hyundong Shin. Segment anything meets semantic communication. *arXiv preprint arXiv:2306.02094*, 2023. 2

[34] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 8

[35] Zhen Wang, Buhong Wang, Yaohui Liu, and Jianxin Guo. Global feature attention network: Addressing the threat of adversarial attack for aerial image semantic segmentation. *Remote Sensing*, 15(5):1325, 2023. 3

[36] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*, 2020. 2

[37] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017. 8

[38] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 3

[39] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 2

[40] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *arXiv preprint arXiv:2312.00863*, 2023. 2, 7, 8

[41] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 2

[42] Chaoning Zhang, Philipp Benz, Adil Karjauv, Jae Won Cho, Kang Zhang, and In So Kweon. Investigating top-k white-box and transferable black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15085–15094, 2022. 2

[43] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 2, 7, 8

[44] Chenshuang Zhang, Chaoning Zhang, Taegoo Kang, Donghun Kim, Sung-Ho Bae, and In So Kweon. Attack-sam: Towards evaluating adversarial robustness of segment anything model. *arXiv preprint arXiv:2305.00866*, 2023. 1, 3, 5, 6

[45] Zhenghao Zhang, Zhichao Wei, Shengfan Zhang, Zuozhuo Dai, and Siyu Zhu. Uvosam: A mask-free paradigm for unsupervised video object segmentation via segment anything model. *arXiv preprint arXiv:2305.12659*, 2023. 2

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[47] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. 2, 7, 8

[48] Zihao Zhao. Enhancing autonomous driving with grounded-segment anything model: Limitations and mitigations. In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*, pages 1258–1265. IEEE, 2023. 1

[49] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems*, 34:6115–6128, 2021. 2

[50] Sheng Zheng and Chaoning Zhang. Black-box targeted adversarial attack on segment anything (sam). *arXiv preprint arXiv:2310.10010*, 2023. 1, 3, 4

[51] Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2384–2391, 2023. 2

[52] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 2