Neural Networks 193 (2026) 108058

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

Region-guided attack on the segment anything model

Xiaoliang Liu[a,*], Furao Shen[b,c], Jian Zhao[d]

[a] School of Information Engineering, Wenzhou Business College, China
[b] National Key Laboratory for Novel Software Technology, Nanjing University, China
[c] School of Artificial Intelligence, Nanjing University, China
[d] School of Electronic Science and Engineering, Nanjing University, China

ARTICLE INFO

Keywords:
Segment anything model
Adversarial attacks
Perturbations
Region-guided
Black-box

ABSTRACT

The Segment Anything Model (SAM) is a cornerstone of image segmentation, demonstrating exceptional performance across various applications, particularly in autonomous driving and medical imaging, where precise segmentation is crucial. However, SAM is vulnerable to adversarial attacks that can significantly impair its functionality through minor input perturbations. Traditional techniques, such as FGSM and PGD, are often ineffective in segmentation tasks due to their reliance on global perturbations that overlook spatial nuances. Recent methods like Attack-SAM-K and UAD have begun to address these challenges, but they frequently depend on external cues and do not fully leverage the structural interdependencies within segmentation processes. This limitation underscores the need for a novel adversarial strategy that exploits the unique characteristics of segmentation tasks. In response, we introduce the Region-Guided Attack (RGA), designed specifically for SAM. RGA utilizes a Region-Guided Map (RGM) to manipulate segmented regions, enabling targeted perturbations that fragment large segments and expand smaller ones, resulting in erroneous outputs from SAM. Our experiments demonstrate that RGA achieves high success rates in both white-box and black-box scenarios, emphasizing the need for robust defenses against such sophisticated attacks. Our codes are available at https://github.com/AbeLiuXL/RGA.

1. Introduction

The Segment Anything Model (SAM) (Kirillov et al., 2023) has emerged as a leading solution in image segmentation, demonstrating remarkable adaptability and performance across diverse datasets and prompts. Its architecture allows for seamless integration with various inputs, making it a pivotal tool for applications ranging from autonomous driving (Yan et al., 2024; Zhao, 2023) to medical imaging (Mazurowski et al., 2023; Zhang & Liu, 2023). However, this versatility also exposes SAM to vulnerabilities, particularly from adversarial attacks that can significantly degrade its performance (Lu et al., 2024; Zhang et al., 2023b). These attacks leverage subtle perturbations in the input data, misleading the model into producing incorrect segmentations, thereby raising concerns about the reliability of SAM in critical contexts.

Previous adversarial attack methods, such as the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) and Projected Gradient Descent (PGD) (Madry et al., 2018), are primarily designed for classification models, where the goal is to manipulate a single prediction for the entire image. These methods typically generate image-wide perturbations, meaning small, pixel-wise modifications crafted based on the model's gradients. In classification, this approach is effective because modifying key discriminative features can shift the model's decision boundary, leading to misclassification.

While such attacks can also degrade segmentation performance, their effectiveness is often limited because segmentation models rely on spatial relationships and local consistency to assign labels to each pixel. Simply applying global perturbations may not directly disrupt object boundaries or structural coherence in a way that significantly alters the segmentation results. As a result, segmentation-specific adversarial attacks often exploit spatial dependencies and region-based manipulations to more effectively distort segmentation outputs. Recent approaches, such as Attack-SAM-K (Zhang et al., 2023b) and UAD (Lu et al., 2024), have explored these challenges by designing attacks tailored for segmentation models.

To effectively address these vulnerabilities, we propose the Region-Guided Attack (RGA), a novel adversarial attack strategy specifically designed for SAM. Unlike traditional adversarial methods that often rely on external prompts or global perturbations, RGA focuses on manipulating segmented regions directly through a Region-Guided Map (RGM). This approach allows for targeted adversarial perturbations that divide large segments into smaller fragments while merging smaller regions into larger areas, ultimately leading to misclassifications in SAM's

* Corresponding author.
E-mail addresses: 20249197@wzbc.edu.cn (X. Liu), frshen@nju.edu.cn (F. Shen), jianzhao@nju.edu.cn (J. Zhao).

https://doi.org/10.1016/j.neunet.2025.108058
Received 23 November 2024; Received in revised form 16 June 2025; Accepted 29 August 2025
Available online 2 September 2025

outputs. The innovation of RGA lies in its ability to exploit the structural dependencies within the segmentation task, leveraging the inherent characteristics of SAM to enhance the effectiveness of the attack.

The significance of RGA is twofold. First, it provides a deeper understanding of the vulnerabilities inherent in advanced segmentation models like SAM, offering insights into how adversarial perturbations can be crafted more strategically. Second, RGA presents a more refined method of inducing segmentation errors that can be applied across various segmentation frameworks, highlighting the need for robust defenses against such targeted attacks. Through extensive experiments, we demonstrate the effectiveness of RGA, revealing its capability to achieve high attack success rates in both white-box and black-box scenarios while maintaining minimal perceptual distortion in the input images. In summary, the key contributions of RGA include:

1. **Region-guided map (RGM) for adversarial guidance**: RGA introduces a novel use of the RGM to directly guide the generation of adversarial examples. By utilizing RGM to define how SAM's segmentation should be altered (i.e., splitting large regions into smaller ones and merging smaller regions into larger ones), RGA effectively guides perturbations to maximize the impact on segmentation quality.
2. **Enhanced attack success and transferability**: By leveraging RGM, RGA achieves higher attack success rates and improved transferability. The adversarial examples are crafted with a well-defined objective influenced by the segmentation output, ensuring that perturbations are systematically guided to mislead the model. Since RGA operates in a black-box setting through transfer-based methods, improving black-box attack success naturally enhances its transferability, making it more effective across different segmentation models.
3. **Independent of external prompts**: Unlike many existing methods that rely heavily on specific prompts to guide attacks, RGA operates independently of external prompts, making the adversarial process more streamlined and broadly applicable. This independence ensures that RGA can be applied in scenarios where prompts are unavailable or unpredictable.
4. **Insights into SAM's segmentation vulnerabilities**: RGA reveals particular vulnerabilities in SAM by focusing on regional manipulations instead of global input perturbations. The findings highlight how region-specific guidance, such as altering the size and boundaries of segmented areas, can degrade SAM's segmentation performance significantly. This understanding provides valuable insights for designing more resilient segmentation models.

2. Related works

2.1. Segmentation models

Image segmentation is a critical task in computer vision, aiming to partition an image into meaningful segments or objects at the pixel level. Traditional approaches relied heavily on hand-crafted features and were limited in handling complex scenes. The advent of deep learning revolutionized segmentation tasks with models like Fully Convolutional Networks (FCNs) (Long et al., 2015), which replaced fully connected layers with convolutional ones to maintain spatial information.

Building upon FCNs, the U-Net architecture (Ronneberger et al., 2015) introduced an encoder-decoder structure with skip connections, enabling precise localization and context assimilation, particularly in biomedical image segmentation. DeepLab models (Chen et al., 2018) further enhanced segmentation by incorporating atrous convolution and conditional random fields for capturing multi-scale context.

The SAM, introduced by Meta AI in 2023 (Kirillov et al., 2023), represents a significant leap in segmentation models. SAM is designed as a promptable segmentation system that can generate high-quality object masks from user input prompts, such as points, boxes, or text descriptions. Trained on a massive dataset of over one billion masks, SAM demonstrates remarkable generalization across diverse image distributions and tasks without the need for additional training.

SAM's architecture comprises three main components: an image encoder, a flexible prompt encoder, and a fast mask decoder. The image encoder processes the input image to produce an embedding, the prompt encoder transforms user prompts into embeddings, and the mask decoder combines these embeddings to generate segmentation masks. This design allows SAM to perform zero-shot generalization to new tasks and image domains, making it a foundational model for various segmentation applications.

Subsequent research has focused on adapting SAM to specific domains and improving its efficiency. For instance, efforts have been made to fine-tune SAM for medical image segmentation, where domain-specific features are crucial (Ma et al., 2024). Other studies explore integrating SAM with text-based prompts to enhance interactive segmentation capabilities (Cheng et al., 2021).

2.2. Adversarial attacks

Adversarial attacks deliberately manipulate input data to deceive machine learning models into making incorrect predictions. Initially studied in image classification (Goodfellow et al., 2015; Szegedy et al., 2013), these attacks exploit the vulnerability of deep neural networks to small, imperceptible perturbations.

The Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) was one of the first techniques introduced to generate adversarial examples by performing a one-step gradient update along the direction of the gradient's sign of the loss function with respect to the input image. Its iterative variant, the Basic Iterative Method (BIM) or I-FGSM (Kurakin et al., 2017), applies FGSM multiple times with smaller step sizes, increasing the attack's success rate.

The Projected Gradient Descent (PGD) (Madry et al., 2018) attack extends BIM by adding random initialization within the allowed perturbation radius and projecting the adversarial example back onto the feasible set after each iteration. PGD is considered a universal "first-order adversary" and is widely used due to its effectiveness in finding robust adversarial examples.

To enhance the effectiveness and transferability of adversarial examples, several advanced methods build upon FGSM, BIM, and PGD:

Momentum Iterative Fast Gradient Sign Method (MIM) (Dong et al., 2018): Incorporates a momentum term into the iterative attack process, stabilizing update directions and improving transferability against different models.

Diverse Input Iterative Fast Gradient Sign Method (DIM) (Xie et al., 2019): DIM increases the diversity of adversarial examples by applying random transformations, such as image resizing and padding, to the input image at each iteration. This randomization helps adversarial perturbations remain effective across models that process inputs of varying dimensions or padding schemes.

Translation-Invariant Iterative Fast Gradient Sign Method (TIM) (Dong et al., 2019): Crafts perturbations invariant to image translations by convolving the gradient with a predefined kernel, increasing transferability across models with different architectures and training data.

Scale-Invariant Iterative Fast Gradient Sign Method (SIM) (Lin et al., 2019): Averages gradients over multiple scaled copies of the input image, capturing scale variations and making adversarial examples effective against models processing images at various scales.

RSTAM (Liu et al., 2022), known as Random Similarity Transformation (RST) Attack Method, applies random similarity transformations (including translation, rotation, and scaling) to diversify the input image during the adversarial attack. This method increases the transferability of adversarial examples by ensuring that perturbations remain effective even under a variety of geometric transformations. EAP (Liu et al., 2024) further innovated by incorporating an image pyramid and meta-ensemble strategy into the RST framework. However, these

methods have not been specifically designed to attack the SAM. Given SAM's unique architecture and prompt-based segmentation capabilities, there is a need to explore adversarial attacks tailored to SAM. In this work, we aim to investigate and develop adversarial attack strategies specifically targeting SAM, to better understand its vulnerabilities and improve its robustness.

2.3. Adversarial attacks on segmentation models

Recent advancements in adversarial attacks on segmentation models have introduced several methods that challenge model robustness and expose vulnerabilities in complex feature extraction processes. Attack-SAM-K (ASK) (Zhang et al., 2023b) employs a global reduction of feature responses by utilizing K point prompts distributed across the entire image, with K often set to a large value such as 400. This approach is designed to manipulate the SAM model's segmentation responses on a broad scale, directly targeting its feature extraction pipeline.

Transferable Adversarial Perturbations (TAP) (Zhou et al., 2018) introduces perturbations that push adversarial features away from original features using Minkowski distance. By focusing on perturbation transferability, TAP highlights cross-model vulnerabilities, making it effective across various model architectures. Building upon TAP, Intermediate-Level Perturbation Decay (ILPD) (Li et al., 2023) refines this approach by maintaining an effective adversarial direction with an increased perturbation magnitude. ILPD targets intermediate-level features, thereby testing the resilience of models at deeper feature layers.

Another method, Activation Attack (AA) (Inkawhich et al., 2019), minimizes the distance between adversarial features and target image features. By manipulating the activation layers, AA achieves targeted feature alignment, granting precise control over model outputs. Prompt-Agnostic Target Attack (PATA) (Zheng & Zhang, 2023) extends AA by incorporating a regularization term to boost the feature dominance of adversarial images over randomly selected clean images. This modification enhances the flexibility of the attack, making it prompt-agnostic and applicable across diverse input conditions.

An extension of PATA, PATA++ (Zheng & Zhang, 2023), addresses the inherent conflict between feature similarity and dominance. It achieves this by selecting a new competition image during each adversarial update iteration, which dynamically adapts the attack to optimize adversarial effectiveness through iteration-based adjustments.

Recent works have explored more practical and transferable attacks on SAM. The Practical Region-level Attack (Shen et al., 2024) introduces region-based perturbations that remain effective regardless of the exact user prompt, and uses spectrum transformations to improve transferability in black-box settings, showing strong results in both controlled and real-world scenarios. Complementing this, the Cross-Prompt Adversarial Attack (Liu & Wei, 2024) proposes Omni-Attack-SAM, which generates perturbations under specific prompts that generalize well to unseen ones, significantly degrading SAM's segmentation performance without relying on ground-truth masks.

Lastly, Unsegment Anything by Simulating Deformation (UAD) (Lu et al., 2024) is a technique that focuses on disrupting the SAM model's segmentation by simulating structural deformations within the image. UAD alters the image's structural details while maintaining an effective adversarial feature distance, using an optimized differentiable deformation function. This approach enhances the robustness and transferability of adversarial examples, offering an effective means of challenging segmentation models across varying structural conditions.

Despite these advancements, the transferability of adversarial attacks specifically designed for SAM still requires improvement. Previous methods may not generalize well across different models or real-world scenarios. In this work, we aim to address this limitation by developing a novel attack strategy that enhances transferability against SAM.

3. Preliminary

In this section, we establish the foundational concepts and framework for understanding the SAM and its susceptibility to adversarial attacks.

3.1. Architecture of the segment anything model

The SAM is a promptable segmentation model designed to handle a wide range of image segmentation tasks. Its architecture comprises three primary components, as illustrated in the Fig. 1:

Image Encoder f_{θ_I}: The image encoder processes the input image $\boldsymbol{x}$ and extracts high-dimensional feature representations. These feature embeddings encapsulate essential visual details required for accurate segmentation.

Prompt Encoder h_{θ_P}: SAM accepts various forms of prompts, such as points, bounding boxes, and textual descriptions. The prompt encoder transforms these user-defined prompts P into embeddings that guide the segmentation process, enabling SAM to tackle diverse tasks without necessitating retraining.

Mask Decoder g_{θ_M}: The mask decoder integrates the outputs from the image encoder and prompt encoder to generate the final segmentation mask. This lightweight component ensures efficient and rapid mask generation.

SAM's ability to adapt to different types of inputs and tasks makes it a powerful tool, but it also exposes the model to adversarial vulnerabilities. The model's reliance on both image and prompt encoders to generate masks means that small perturbations to the input image or the prompt embeddings can lead to significant segmentation errors.

3.2. Adversarial attacks in segmentation

Adversarial attacks in segmentation involve generating small, imperceptible perturbations to the input image that mislead the model into producing incorrect segmentation results. These attacks, originally studied in classification tasks, have been extended to segmentation models. Common methods like FGSM (Goodfellow et al., 2015) and PGD

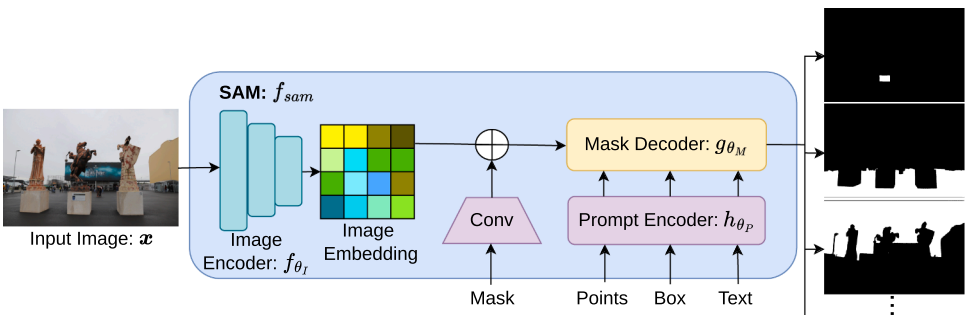


Fig. 1. Overview of the architecture of the Segment Anything Model (SAM), highlighting its key components and operational flow.

(Madry et al., 2018) target pixel-level predictions, making them applicable in segmentation tasks.

For SAM, adversarial attacks must account for how the model processes both the image and prompt embeddings. The challenge is to craft perturbations that disrupt SAM's ability to correctly combine these embeddings, leading to erroneous segmentation masks. The goal is to minimize the model's Intersection over Union (IoU) score, effectively degrading its segmentation performance.

The perturbations are typically constrained by an ℓ_∞-norm, which limits their magnitude to remain imperceptible to human observers while causing significant model errors. Formally, the adversarial objective can be defined as:

$$\delta = \underset{\delta, \|\delta\|_\infty \leq \epsilon}{\operatorname{argmix}} \mathbb{E}_{\theta_I, \theta_P, \theta_M} \left[\mathrm{IoU}(g_{\theta_M}(f_{\theta_x}(\boldsymbol{x}+\delta), h_{\theta_P}(P)), g_{\theta_M}(f_{\theta_x}(\boldsymbol{x}), h_{\theta_P}(P))) \right] \tag{1}$$

where δ is the adversarial perturbation, $\boldsymbol{x}$ is the input image, P is the prompt, and ϵ is the perturbation limit. The function g_{θ_M} represents the mask decoder, f_{θ_I} the image encoder, and h_{θ_P} the prompt encoder.

The adversary aims to find a perturbation δ that minimizes the IoU between the original mask and the mask resulting from the perturbed image, effectively disrupting SAM's output while keeping the perturbation imperceptible.

3.3. Challenges in adversarial attacks on SAM

Adversarial attacks on SAM present unique challenges compared to standard segmentation models:

- **Prompt-agnostic attacks**: SAM's versatility in handling different types of prompts makes it challenging to design adversarial examples that generalize across various prompts. Unlike traditional segmentation models, where the input is static, SAM's output depends on the specific prompts provided by the user, increasing the complexity of the attack.
- **Transferability**: Attacks designed for SAM must also transfer effectively across different segmentation models, including those with varying architectures and prompt types. Many existing attacks fail to generalize across different models, limiting their practical impact.
- **Feature-level perturbations**: Attacks on SAM must focus on disrupting the model's feature representations. Since SAM relies on both image and prompt features to generate masks, the adversary must craft perturbations that target the image feature space without being overly dependent on a specific prompt.

These challenges motivate the need for novel adversarial strategies, such as the Region-Guided Attack, which is designed to exploit SAM's segmentation vulnerabilities more effectively.

4. Approach

In this section, we present the details of our proposed Region-Guided Attack (RGA) targeting the Segment Anything Model (SAM). The method is designed to manipulate SAM's segmentation capabilities by systematically altering how the model interprets regions within an image. By focusing on both large and small segmented areas, we can induce errors in SAM's output with minimal and imperceptible perturbations to the input image. The RGA framework leverages SAM's prompt-based segmentation architecture to achieve targeted adversarial attacks while maintaining a high degree of transferability to other segmentation models.

4.1. Overview

The Region-Guided Attack (RGA) targets the SAM by manipulating how it segments image regions, causing segmentation errors through strategic adversarial perturbations. The attack process involves querying SAM for an initial segmentation, generating a guidance map based on the model's output, and using the Segmentation and Dilation Strategy (SAD) to craft perturbations that alter SAM's segmentation boundaries. This manipulation divides large regions into smaller segments and merges small regions into larger ones, disrupting the segmentation process.

As illustrated in the Fig. 2, the RGA framework is structured around three key steps:

- Single Query: The process begins by feeding the original image into SAM and retrieving the segmentation result. This result provides the basis for generating the adversarial perturbations.
- Region-Guided Map (RGM) : RGM is constructed based on the initial segmentation result from SAM. It serves as a guide for how adversarial perturbations should be applied to influence SAM's output. Specifically, the RGM is used to misguide SAM by splitting originally

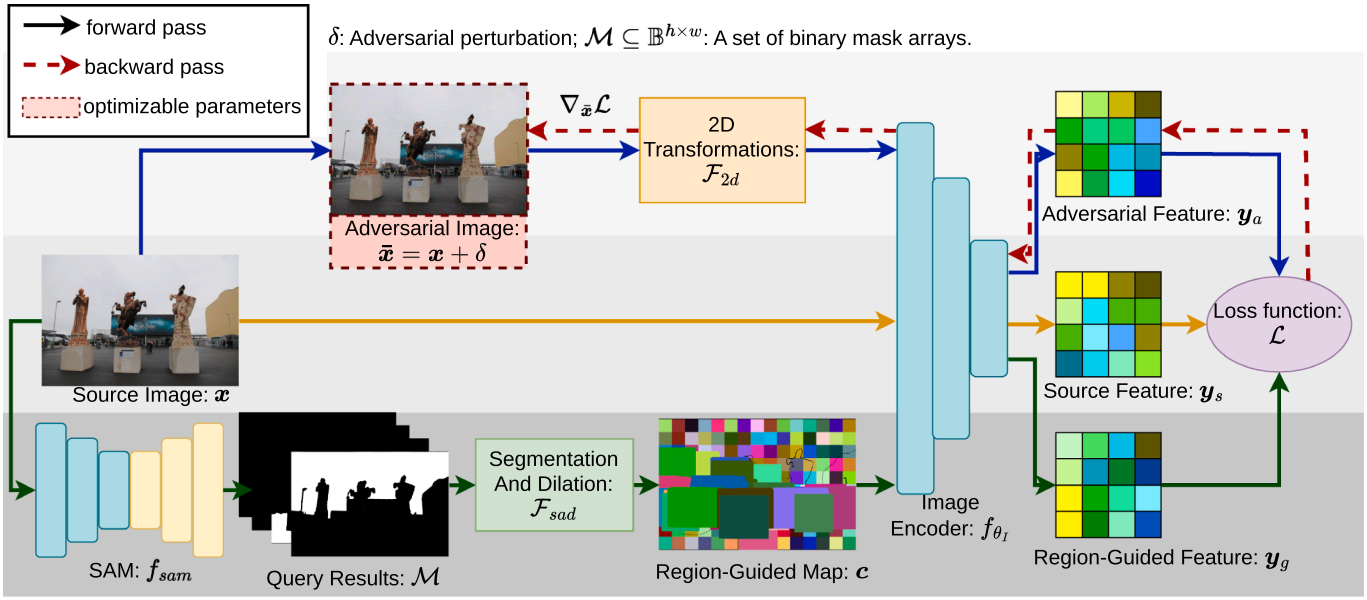


Fig. 2. Overview of the region-guided attack (RGA) framework on SAM.

segmented large regions into smaller fragments, while smaller regions are expanded into larger ones. This is achieved by leveraging the SAD strategy, which tailors the perturbation process to different region sizes - applying grid-based segmentation for large areas and dilation-based enhancement for smaller areas. The RGM effectively guides SAM to make segmentation errors, altering the size and boundaries of regions in a manner that leads to incorrect segmentation outputs.

- Adversarial Perturbation: The adversarial perturbations are generated through a gradient-based optimization process. The loss function calculates the difference between SAM's initial segmentation and the desired incorrect segmentation based on the guidance map. By applying these perturbations to the input image, the model is forced to produce incorrect segmentations without visually altering the input image.

The workflow involves a forward pass, where SAM generates the segmentation based on the perturbed image, and a backward pass to optimize the perturbations using the guidance map. The adversarial perturbations are then applied to the image, resulting in an adversarial image that misleads SAM, as shown in the Fig. 2.

4.2. Problem formulation

The goal of the RGA is to generate adversarial perturbations for an input image $\boldsymbol{x}$ such that the SAM, denoted as f_{sam}, produces incorrect segmentation results that align with a predefined guidance map. Given an input image $\boldsymbol{x}$ and SAM f_{sam}, we want to generate a perturbation δ such that when applied to the image $\boldsymbol{x}$, the perturbed image $\boldsymbol{x}+\delta$ leads SAM to produce a segmentation output that deviates significantly from the original segmentation output, as defined by the guidance map $\boldsymbol{c}$. The aim is to guide SAM into splitting large segments into smaller fragments and merging smaller regions into larger ones, ultimately compromising segmentation accuracy. The optimization objective can be described as follows:

$$\delta = \underset{\delta,\|\delta\|_\infty \le \epsilon}{\operatorname{argmix}} \mathcal{L}\Big(f_{\theta_I}(\boldsymbol{x}), f_{\theta_I}(\mathcal{F}_{2\text{d}}(\boldsymbol{x}+\delta)), f_{\theta_I}(\boldsymbol{c})\Big) \tag{2}$$

where δ represents the adversarial perturbation applied to the image $\boldsymbol{x}$, ϵ denotes the perturbation bound that ensures the perturbation remains imperceptible to human observers, f_{θ_I} is the image encoder function used in SAM to generate feature embeddings, $\mathcal{F}_{2\text{d}}(\cdot)$ refers to the 2D transformation function that augments the perturbed image to increase its diversity, and $\boldsymbol{c}$ is the guidance map generated using the SAD strategy, which serves as a target for directing the segmentation output towards incorrect boundaries.

The goal is to find an adversarial perturbation δ that maximizes the loss function $\mathcal{L}$, which measures the divergence between SAM's original segmentation and the target segmentation defined by $\boldsymbol{c}$. This ensures that the segmentation output is altered according to the desired adversarial effect.

4.3. Segmentation and dilation strategy

In this work, we introduce the Segmentation and Dilation (SAD) strategy for efficiently processing binary mask arrays and applying random colorization based on the size of the segmented regions. This strategy addresses varying scales of segmented areas through two distinct approaches: **grid-based segmentation** for large regions and **dilation-based enhancement** for smaller regions, ensuring accurate and non-overlapping color application across the mask.

The grid size, denoted as $grid\_size$, is calculated based on the image dimensions (width w and height h) and the granularity parameter γ:

$$grid\_size = \lfloor \min(w,h) \times \gamma \rfloor \tag{3}$$

where w and h represent the width and height of the image, respectively, and γ controls the size of the grid blocks used for large-region segmentation.

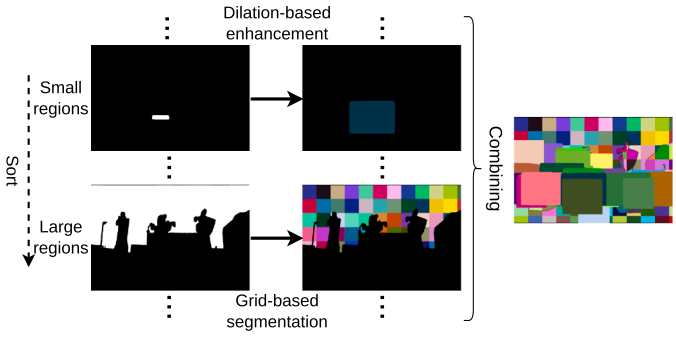


Fig. 3. Illustration of the SAD strategy for generating the Region-Guided Map (RGM). Small regions segmented by SAM are enhanced via dilation and added first. Large regions are then subdivided using grid-based segmentation and added later, typically forming the background.

For **small regions**, defined as those with an area smaller than $grid\_size^2$, a dilation-based enhancement operation is applied. This operation, called DilateRegion, is performed using a predefined structuring element over n iterations, where n is the number of dilation steps. A random RGB color R_1 is assigned to each small region, and the dilated mask is updated by applying the color only to previously uncolored regions:

$$\boldsymbol{c} \leftarrow \boldsymbol{c} + (R_1 \times \text{DilateRegion}(\boldsymbol{M}_j, n)) \cdot (\boldsymbol{c} == 0). \tag{4}$$

Here, $\boldsymbol{c} \in \mathbb{R}^{h\times w\times 3}$ denotes the color mask, which is initialized to zeros, meaning that no areas are colored initially. DilateRegion($\boldsymbol{M}_j, n$) denotes the mask $\boldsymbol{M}_j$ after n iterations of dilation, and $\boldsymbol{c} == 0$ ensures that the random color R_1 is applied solely to uncolored areas in the final mask $\boldsymbol{c}$.

For **large regions**, where the segmented area exceeds $grid\_size^2$, the region is subdivided into grid blocks of size $grid\_size$. Each grid block containing at least one positive pixel is assigned a unique random RGB color R_2. The color is applied to each grid block, ensuring no overlap, and the colored blocks are combined to form the final result:

$$\boldsymbol{c} \leftarrow \boldsymbol{c} + (R_2 \times \text{SegmentGrid}(\boldsymbol{M}_j, grid\_size)) \cdot (\boldsymbol{c} == 0), \tag{5}$$

where SegmentGrid($\boldsymbol{M}_j, grid\_size$) refers to the grid-based segmentation of the large-region mask $\boldsymbol{M}_j$.

By combining these two techniques-dilation-based enhancement with n iterations for small regions and grid-based segmentation for large regions-the SAD strategy effectively adapts to regions of various sizes. This dual approach ensures that the final mask $\boldsymbol{c}$ is accurately colorized without overlap, while enhancing the visibility of smaller regions that might otherwise be neglected.As shown in Fig. 3, this is an example of using the SAD strategy to generate a region-guided map. For more detailed descriptions, please refer to Algorithm 1.

4.4. 2D transformations

Previous research has shown that increasing the diversity of input images during the generation of adversarial examples enhances both the effectiveness and transferability of the attacks. Techniques such as DIM (Xie et al., 2019), TIM (Dong et al., 2019), SIM (Lin et al., 2019), and RSTAM (Liu et al., 2022) have introduced methods to manipulate input images, enabling adversarial examples to generalize across different models and tasks.

In this work, we implement the Random Similarity Transformation Strategy from RSTAM (Liu et al., 2022, 2024), which applies random translations, rotations, and scaling to input images. This method ensures that the generated perturbations remain effective against geometric transformations. Additionally, we adopt the Scale-Invariant Strategy from SIM (Lin et al., 2019), which averages gradients across multiple scaled versions of the image. By integrating these strategies, we improve the robustness and generalization of our adversarial examples,

Algorithm 1 Segmentation and dilation strategy.

Require: $\mathcal{M} \subseteq \mathbb{B}^{h \times w}$: A set of binary mask arrays
w, h: Width and height of the image
γ: Granularity parameter for grid size
n: Number of dilation iterations
Ensure: $\boldsymbol{c}$: Colored mask
Initialize color mask $\boldsymbol{c} \leftarrow \text{zeros}(h, w, 3)$ ▷ All zeros
Compute grid size $grid\_size \leftarrow \lfloor \min(w, h) \times \gamma \rfloor$
for each mask $\boldsymbol{M}_j$ in $\mathcal{M}$ **do**
  if area of $\boldsymbol{M}_j \leq grid\_size^2$ **then** ▷ Small region detected
    Perform dilation on small region: DilateRegion($\boldsymbol{M}_j, n$)
    Generate random color R_1
    Apply color $\boldsymbol{c} \leftarrow \boldsymbol{c} + (R_1 \times \text{DilateRegion}(\boldsymbol{M}_j, n)) \cdot (\boldsymbol{c} == 0)$
  else ▷ Large region detected
    Perform grid-based segmentation: $\boldsymbol{B}_{\text{grid}} \leftarrow \text{SegmentGrid}(\boldsymbol{M}_j, grid\_size)$
    for each grid block B in $\boldsymbol{B}_{\text{grid}}$ **do**
      if block B contains any positive pixel **then**
        Generate random color R_2
        Apply color $\boldsymbol{c} \leftarrow \boldsymbol{c} + (R_2 \times B) \cdot (\boldsymbol{c} == 0)$
      end if
    end for
  end if
end for
return $\boldsymbol{c}$

enhancing their transferability across various models and real-world applications.

4.5. Loss function

To guide the adversarial attack, we define the following loss function $\mathcal{L}$, which aims to reduce the similarity between the adversarial image and the source image while encouraging its similarity to a guidance map:

$$\mathcal{L} = \frac{< \boldsymbol{y}_a, \boldsymbol{y}_s >}{||\boldsymbol{y}_a||^2 \cdot ||\boldsymbol{y}_s||^2} - \lambda \frac{< \boldsymbol{y}_a, \boldsymbol{y}_g >}{||\boldsymbol{y}_a||^2 \cdot ||\boldsymbol{y}_g||^2} \tag{6}$$

where:

- $< \cdot, \cdot >$ denotes the inner product of the vectors.
- $\boldsymbol{y}_a = f_{\theta_I} \mathcal{F}_{2d}(\bar{\boldsymbol{x}})$ denotes the features extracted from the adversarial image $\bar{\boldsymbol{x}}$ after applying a 2D transformation $\mathcal{F}_{2d}$.
- $\boldsymbol{y}_s = f_{\theta_I}(\boldsymbol{x})$ represents the features extracted from the source image $\boldsymbol{x}$.
- $\boldsymbol{y}_g = f_{\theta_I}(c)$ corresponds to the features of the guidance map $\boldsymbol{c}$.
- f_{θ_I} is the image encoder from the SAM used for feature extraction.
- λ is a regularization parameter that controls the contribution of the second term. The default value of λ is set to 1.

The loss function $\mathcal{L}$ drives the adversarial image to move away from the source image while moving closer to the guidance map, allowing for control over the attack's behavior through the regularization parameter λ.

The RGA approach enables targeted, region-specific adversarial perturbations that disrupt SAM's segmentation without requiring external prompts. By focusing on the internal feature structure, RGA achieves a high degree of transferability and robustness against various segmentation models. A detailed algorithmic description of the RGA approach is provided in Algorithm 2.

The computational complexity of RGA is primarily determined by the SAD strategy and iterative adversarial optimization. The SAD strategy contributes $O(hwn)$ due to dilation and segmentation operations, while adversarial optimization incurs $O(TE)$, where T is the number of adversarial iterations and E represents the computational cost of SAM's feature extraction and backpropagation. Additional operations, such as

Algorithm 2 Region-guided attack (RGA).

Require: Source image $\boldsymbol{x}$.
Require: Segment Anything Model (SAM) f_{sam}; SAM image encoder f_{θ_I}.
Require: 2D transformation function $\mathcal{F}_{2d}$; Segmentation and dilation function $\mathcal{F}_{\text{sad}}$.
Require: Granularity parameter for grid size γ; Number of dilation iterations n; Perturbation step size α; Perturbation bound ϵ; Decay factor for momentum μ; Number of adversarial iterations T.
Require: Loss function $\mathcal{L}$.
Ensure: Adversarial perturbation: δ.
Initialize the gradient: $\boldsymbol{g}_0 \leftarrow 0$
Initialize the adversarial image with uniform noise: $\bar{\boldsymbol{x}} \leftarrow \boldsymbol{x} + \mathcal{U}(-\epsilon, \epsilon)$
Query SAM with the source image: $\mathcal{M} \leftarrow f_{\text{sam}}(\boldsymbol{x})$
Generate the guidance map using the SAD strategy: $\boldsymbol{c} \leftarrow \mathcal{F}_{\text{sad}}(\mathcal{M}; \gamma, n)$
for $t = 0$ to $T - 1$ **do**
  Compute the gradient of the loss function:
  $\boldsymbol{g}_t^{'} \leftarrow \nabla_{\bar{\boldsymbol{x}}_t} \mathcal{L}\left(f_{\theta_I}(\boldsymbol{x}), f_{\theta_I} \mathcal{F}_{2d}(\bar{\boldsymbol{x}}_t), f_{\theta_I}(\boldsymbol{c})\right)$
  Update the gradient using momentum:
  $\boldsymbol{g}_{t+1} \leftarrow \mu \cdot \boldsymbol{g}_t + \boldsymbol{g}_t^{'}$
  Update the adversarial image using the sign of the gradient:
  $\bar{\boldsymbol{x}}_{t+1} \leftarrow \underset{[\boldsymbol{x}-\epsilon, \boldsymbol{x}+\epsilon]}{\text{Clip}} \left(\bar{\boldsymbol{x}}_t + \alpha \cdot \text{sign}(\boldsymbol{g}_{t+1})\right)$
end for
Compute the adversarial perturbation: $\delta \leftarrow \bar{\boldsymbol{x}}_T - \boldsymbol{x}$
return δ

gradient updates and perturbation clipping, contribute $O(Thw)$. Overall, the total complexity is $O(hwn) + O(TE)$, with the primary bottleneck being gradient computation in SAM.

5. Experiments

5.1. Experiment settings

Datasets: For evaluation, we utilized the SA-1B dataset (Kirillov et al., 2023), introduced in the Segment Anything project by Meta AI . This dataset comprises approximately 11 million diverse, high-resolution images (approximately 1024×1024 pixels), each accompanied by segmentation masks, resulting in over 1 billion masks in total. The images were sourced from licensed and privacy-protecting repositories, ensuring a wide variety of environments, object types, and lighting conditions.

From this extensive dataset, we curated a subset of 1000 images, labeled sequentially from sa_1.jpg to sa_1000.jpg. This subset was selected to maintain the diversity and complexity of the original dataset, encompassing various scenes and objects. Collectively, these 1000 images contain a total of 98,875 segmentation masks, providing a substantial amount of data to enhance the robustness and statistical significance of our findings.

Compared baselines: To assess the effectiveness of our proposed method, we conducted comparisons with several established adversarial attack techniques on segmentation models. The baselines included:

1. Clean: The original segmentation results without any adversarial perturbations, serving as a baseline for comparison.
2. Attack-SAM-K (ASK) (Zhang et al., 2023b): Utilizes a large number of prompts to globally alter SAM's feature responses, challenging its segmentation robustness.

3. Transferable Adversarial Perturbations (TAP) (Zhou et al., 2018): Focuses on generating adversarial perturbations with high transferability across different models, exposing general model vulnerabilities.
4. Intermediate-Level Perturbation Decay (ILPD) (Li et al., 2023): Extends TAP by increasing perturbation magnitude at intermediate feature levels, further testing model resilience.
5. Activation Attack (AA) (Inkawhich et al., 2019): Aligns adversarial features with target features, offering precise control over model output alterations.
6. Prompt-Agnostic Target Attack (PATA) (Zheng & Zhang, 2023): Enhances adversarial feature dominance without reliance on specific prompts, making it adaptable to varied input conditions.
7. PATA++ (Zheng & Zhang, 2023): An improved version of PATA that dynamically adjusts competition images, optimizing adversarial effects through iterative updates.
8. Unsegment Anything by Simulating Deformation (UAD) (Lu et al., 2024): Alters image structure to disrupt segmentation, leveraging deformation to enhance adversarial robustness and transferability.
9. DarkSAM (Zhou et al., 2025): A prompt-free Universal Adversarial Perturbation (UAP) attack against SAM, disrupting segmentation by manipulating both spatial semantics and high-frequency textures. It effectively degrades SAM's performance across diverse images and prompts with a single perturbation.

Evaluation Metrics: To assess the performance of our adversarial attack method, we employed the following evaluation metrics:

Mean Intersection over Union (mIoU): The mIoU measures the average overlap between the predicted segmentation and the ground truth masks. It is calculated by taking the mean of the Intersection over Union (IoU) across all masks. The mIoU is defined as:

$$\text{mIoU} = \frac{1}{N}\sum_{m=1}^{N}\frac{|P_m \cap G_m|}{|P_m \cup G_m|} \tag{7}$$

where: N is the total number of masks, P_m is the predicted segmentation for mask m, and G_m is the corresponding ground truth segmentation for mask m.

Attack Success Rate at IoU $\leq 50\,\%$ (ASR@50): ASR@50 measures the proportion of SAM-generated masks, across all adversarial examples, where the IoU with the ground truth mask is less than or equal to 50 %. This metric evaluates the effectiveness of the attack by indicating how frequently it significantly reduces segmentation accuracy. The formula for ASR@50 is:

$$\text{ASR@50} = \frac{1}{N}\sum_{m=1}^{N}\mathbb{I}(\text{IoU}_m \leq 0.50) \tag{8}$$

where N is the total number of masks across all adversarial examples, IoU_m is the IoU for mask m, and $\mathbb{I}$ is the indicator function, equal to 1 if $\text{IoU}_m \leq 0.50$, and 0 otherwise.

Attack Success Rate at IoU $\leq 10\,\%$ (ASR@10): Similar to ASR@50, ASR@10 measures the proportion of masks for which the IoU with the ground truth mask is less than or equal to 10 %, across all adversarial examples. This metric captures cases where the attack is highly effective, causing significant degradation in segmentation accuracy. The formula for ASR@10 is:

$$\text{ASR@10} = \frac{1}{N}\sum_{m=1}^{N}\mathbb{I}(\text{IoU}_m \leq 0.10) \tag{9}$$

where the terms N, IoU_m, and $\mathbb{I}$ are as defined above, with the threshold adjusted to 10 %.

Implementation Details: In our experiments, we set the following default parameters. The granularity parameter for grid size, γ, is set to 0.1 to ensure adequate detail in the segmentation process, and the number of dilation iterations, n, is set to 100 to expand regions effectively. For all adversarial attacks, we set the perturbation bound, ϵ, to 8/255 and the perturbation step size, α, to 2/255, with the number of adversarial iterations, T, set to 40 to balance effectiveness and computational efficiency. The momentum decay factor for MI is set to 0.4, the number of scale copies for SI is set to 3, and the transformation range factor for RST is set to 0.01.All experiments are conducted on four NVIDIA GeForce GTX 3090 GPUs, and the models are implemented using the PyTorch (Paszke et al., 2019) framework.

5.2. Quantitative evaluation

We evaluate the performance of our proposed RGA method against several state-of-the-art adversarial attack techniques, including ASK (Zhang et al., 2023b), TAP (Zhou et al., 2018), ILPD (Li et al., 2023), AA (Inkawhich et al., 2019), PATA (Zheng & Zhang, 2023), PATA++ (Zheng & Zhang, 2023), and UAD (Lu et al., 2024). The evaluation metrics used are mean Intersection over Union (mIoU) (lower is better), Attack Success Rate at thresholds 50 % (ASR@50), and 10 % (ASR@10) (higher is better). The experiments are conducted on four models: SAM-B (white-box), SAM-L, SAM-H, and FastSAM (Zhao et al., 2023). The results are summarized in Table 1.

As shown in Tables 1 and 2, our RGA method significantly outperforms existing adversarial attack methods across all models and evaluation metrics. Specifically, on the SAM-B (white-box) model, RGA achieves a mIoU of 26.64 ± 31.95, which is substantially lower than the closest competitor, UAD, with a mIoU of 51.53 ± 34.00. This indicates that RGA is more effective in degrading the segmentation quality of the model. Furthermore, RGA achieves the highest rates of ASR @ 50 and ASR @ 10 of 72.08 % and 55.22 %, respectively, demonstrating superior attack success.

Similar trends are observed on the SAM-L and SAM-H models, where RGA consistently achieves lower mIoU values and higher ASR rates compared to other methods. Notably, on the FastSAM model,

Table 1
Quantitative comparison of different adversarial attack methods on SAM-B, SAM-L, and SAM-H. The metrics include mean Intersection over Union (mIoU), Attack Success Rate at thresholds 50 % (ASR@50), and 10 % (ASR@10). Lower mIoU and higher ASR values indicate better attack performance. The best performances in each block are shown in **bold**.

Method	SAM-B(white-box)			SAM-L			SAM-H		
	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑
Clean	82.01 ± 27.27	10.14	1.75	85.73 ± 27.26	7.57	1.26	85.97 ± 25.26	7.42	1.12
ASK (Zhang et al., 2023b)	68.07 ± 28.65	24.10	6.87	77.14 ± 25.05	14.46	4.13	78.71 ± 24.02	12.93	3.51
TAP (Zhou et al., 2018)	63.49 ± 32.58	29.69	13.12	75.12 ± 27.83	16.96	6.68	77.36 ± 26.21	14.55	5.31
ILPD (Li et al., 2023)	63.21 ± 32.54	30.15	13.09	75.17 ± 27.75	16.82	6.59	77.52 ± 26.02	14.37	5.18
AA (Inkawhich et al., 2019)	61.06 ± 32.33	32.48	13.11	70.70 ± 29.74	21.49	8.67	72.87 ± 28.61	19.13	7.39
PATA (Zheng & Zhang, 2023)	61.36 ± 32.31	32.23	13.04	70.81 ± 29.73	21.27	8.62	73.07 ± 28.49	18.66	7.30
PATA++ (Zheng & Zhang, 2023)	61.54 ± 32.22	32.00	12.94	71.02 ± 29.55	21.16	8.38	73.16 ± 28.44	18.84	7.22
UAD (Lu et al., 2024)	51.53 ± 34.00	43.89	20.79	66.07 ± 32.04	26.44	12.27	68.96 ± 30.87	23.42	10.23
DarkSAM (Zhou et al., 2025)	75.36 ± 31.93	15.91	4.77	79.08 ± 32.17	12.56	3.89	80.17 ± 31.84	11.45	3.36
RGA(Ours)	**26.64 ± 31.95**	**72.08**	**55.22**	**28.63 ± 32.59**	**69.93**	**52.09**	**30.98 ± 33.55**	**67.38**	**48.82**

Table 2
Quantitative comparison of different adversarial attack methods on various segmentation models. The best performances in each block are shown in **bold**.

Method	FastSAM (Zhao et al., 2023)		EfficientSAM (Xiong et al., 2024)		MobileSAM (Zhang et al., 2023a)		HQ-SAM (Ke et al., 2023)	
	mIoU↓	ASR@10↑	mIoU↓	ASR@10↑	mIoU↓	ASR@10↑	mIoU↓	ASR@10↑
Clean	42.43	42.38	77.96	2.95	70.67	4.20	79.10	2.69
ASK (Zhang et al., 2023b)	38.13	48.43	73.55	4.81	65.14	8.70	65.58	8.41
AA (Inkawhich et al., 2019)	32.64	55.10	69.62	6.77	60.39	12.34	59.53	12.90
PATA (Zheng & Zhang, 2023)	32.74	54.97	70.01	6.63	60.71	12.34	60.54	12.10
PATA++ (Zheng & Zhang, 2023)	32.85	54.65	69.66	7.11	60.90	11.98	60.37	12.21
UAD (Lu et al., 2024)	28.83	59.63	66.44	9.02	58.22	15.80	49.56	20.89
DarkSAM (Zhou et al., 2025)	28.52	59.18	72.54	6.32	63.03	11.36	72.34	5.79
RGA(Ours)	**5.32**	**89.56**	**42.63**	**33.98**	**30.99**	**51.72**	**20.61**	**59.20**

Fig. 4. Qualitative comparison of segmentation results under different adversarial attacks on Meta AI's online SAM and FastSAM models in a black-box setting. The comparison includes various prompts, with Point, Box, and Everything on Meta AI's online SAM, and Text ("dog") on FastSAM. The perturbation bound ϵ is reduced to 4/255. While most other attack methods are largely ineffective, our method continues to successfully disrupt segmentation.

RGA achieves an exceptionally low mIoU of 5.32 ± 14.98 and ASR@50 and ASR@10 rates of 96.91 % and 89.56 %, respectively. This highlights the effectiveness of RGA in both white-box and black-box settings.

The superior performance of RGA can be attributed to its ability to generate regionally optimized adversarial perturbations that effectively disrupt the segmentation process of the models. These results confirm the efficacy of our approach in compromising the robustness of segmentation models to adversarial attacks.

5.3. Qualitative evaluation

We conducted qualitative evaluations to assess the effectiveness of the proposed RGA in a black-box setting, specifically targeting Meta AI's online SAM model and FastSAM. In our experiments, we validated the attack on Meta AI's online SAM model using various prompt types: Point, Box, and Everything. For FastSAM, a Text prompt set to "dog" guided the segmentation process.

In previous experiments, we used a default perturbation bound of $\epsilon = 8/255$. In this qualitative evaluation, however, for all attack methods, the perturbation bound ϵ is reduced to 4/255 to ensure a more challenging and subtle attack scenario, allowing us to better compare the effectiveness of each method under constrained conditions.

Fig. 4 presents visual results comparing the segmentation masks generated under normal conditions and under adversarial attacks by RGA, ASK, PATA++, and UAD. The comparison illustrates how RGA effectively disrupts segmentation performance across all prompt types on Meta AI's online SAM, producing segmentation errors by fragmenting large regions and expanding small regions. This validates RGA's capability to induce substantial segmentation errors while maintaining visually subtle perturbations to the input image.

In comparison, FastSAM's segmentation result using the Text prompt "dog" demonstrates the versatility and transferability of RGA in handling different prompt-based black-box models. Our results indicate that RGA consistently outperforms existing adversarial attacks in reducing segmentation quality across both Meta AI's online SAM and FastSAM models.

Fig. 5 further illustrates the effectiveness of RGA by visualizing the adversarial perturbations generated according to the RGM. The RGM acts as the target for RGA, guiding the perturbations to specifically manipulate segmented regions identified during the attack. By comparing the perturbations applied by RGA against those from ASK, PATA++, and UAD, we can observe distinct patterns that reveal how each method influences segmentation outputs. RGA's use of the RGM allows for strategic alterations-fragmenting large segments and expanding small ones-thereby reinforcing its ability to induce significant segmentation errors.

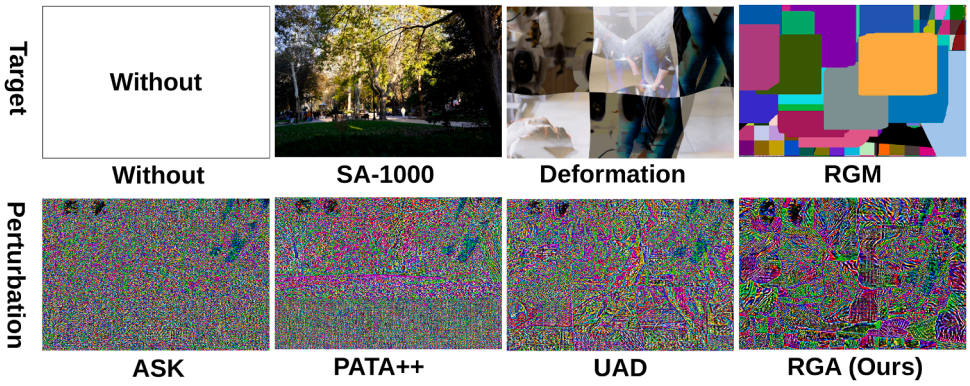


Fig. 5. Visualizations of generated adversarial perturbations from different methods, corresponding to different targets used during the attack on segmentation models.

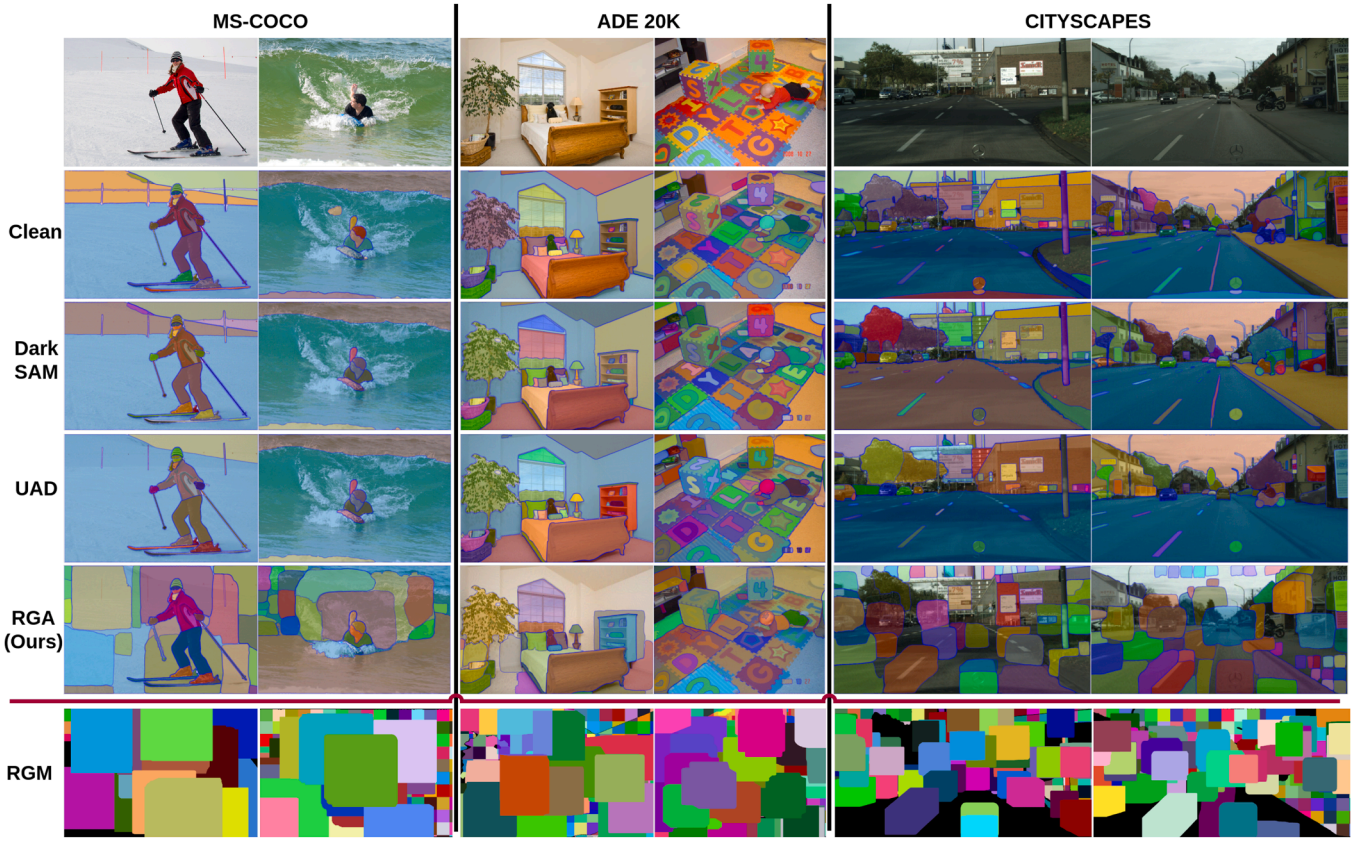


Fig. 6. Qualitative comparison of segmentation results under different scenarios on Meta AI's online SAM in a black-box setting. The perturbation bound ϵ is set to 4/255.

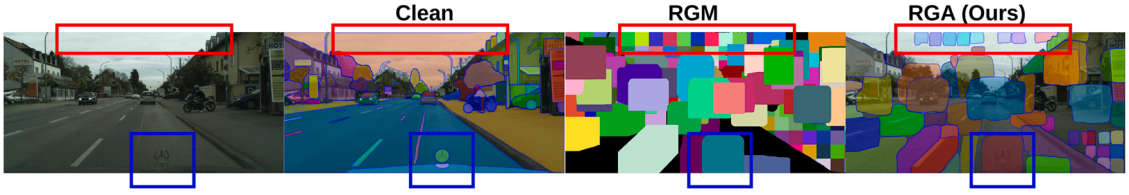


Fig. 7. Visualization of Region-Guided Attack (RGA) on SAM under different region types. The background (e.g., sky) is misled to be fragmented into small regions, while small foreground objects are expanded into abnormally large regions. This demonstrates RGA's capability to selectively disrupt the segmentation of different region types in a targeted manner.

Table 3
Ablation study results for the RGA on SAM-B (white-box), SAM-L, and SAM-H models. Each row shows the effect of enabling different components of the RGA framework, including Region-Guided Map (RGM), Momentum Iteration (MI), Random Similarity Transformation (RST), and Scale-Invariance (SI). The best performances in each block are shown in **bold**.

RGM	MI	RST	SI	SAM-B(white-box)			SAM-L			SAM-H		
				mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑
×	×	×	×	49.98	46.82	23.66	71.49	21.15	7.92	75.24	16.90	5.41
✓	×	×	×	45.71	50.67	32.83	70.66	22.14	9.30	74.54	17.42	6.85
✓	✓	×	×	46.63	49.77	31.67	69.98	22.57	10.19	73.89	17.76	7.42
✓	✓	✓	×	29.71	68.84	52.31	33.38	64.75	47.53	38.11	59.57	41.57
×	✓	✓	✓	30.27	71.59	40.39	32.02	69.40	38.14	33.22	68.04	36.76
✓	×	✓	✓	27.64	71.12	54.47	29.43	69.03	51.63	32.74	65.68	47.77
✓	✓	✓	✓	**26.87**	**72.99**	**55.60**	**28.27**	**69.64**	**52.66**	**31.15**	**67.41**	**48.81**

Table 4
Impact of integrating the Region-Guided Map (RGM) with traditional adversarial attack methods. The best performances in each block are shown in **bold**.

Method	SAM-B(white-box)			SAM-L			SAM-H		
	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑
MIM (Dong et al., 2018)	54.64	40.92	19.01	72.27	20.09	7.45	75.44	16.69	5.52
MIM + RGM	**46.59**	**49.70**	**32.04**	**70.09**	**22.20**	**10.05**	**74.13**	**17.36**	**7.05**
DIM (Xie et al., 2019)	33.46	67.95	36.16	44.23	53.85	26.45	47.79	50.13	24.04
DIM + RGM	**28.32**	**70.04**	**54.14**	**40.51**	**56.50**	**39.93**	**46.34**	**49.89**	**32.96**

This targeted approach, depicted in Fig. 5, underscores RGA's robustness in maintaining subtle visual integrity while effectively disrupting segmentation performance across various conditions. Together, these figures highlight RGA's capability to perform sophisticated adversarial attacks tailored to the structural characFteristics of segmentation models.

Fig. 6 and Fig. 7 present a qualitative comparison of segmentation results under different scenarios on Meta AI's online SAM in a black-box setting. These scenarios are drawn from three diverse datasets: MS-COCO (Lin et al., 2014), ADE 20K (Zhou et al., 2017), and CITYSCAPES (Cordts et al., 2016), covering a range of object categories and scene complexities. The results demonstrate that while most existing adversarial attack methods struggle to significantly impact segmentation, RGA remains highly effective, consistently inducing substantial segmentation errors across different datasets. This highlights the robustness and transferability of RGA in real-world applications, further underscoring the necessity for improved defenses against adversarial threats in segmentation models.

5.4. Ablation studies

To comprehensively evaluate the contributions of various components in the RGA framework, we conduct a series of ablation studies focusing on three main aspects: the individual impact of the proposed components, the integration of RGA with traditional adversarial attack methods, and the evaluation of different adversarial target types. The ablation experiments are conducted on SAM-B (white-box), SAM-L, and SAM-H models, and the results are presented in three tables.

5.4.1. Component analysis

To better understand the impact of different components in our RGA framework, we evaluate the following: Region-Guided Map (RGM), Momentum Iteration (MI), Random Similarity Transformation (RST), and Scale-Invariance (SI). Table 3 provides a summary of the results, highlighting the effectiveness of each component when applied individually or in combination.

As observed in Table 3, the baseline version of RGA, where none of the components (RGM, MI, RST, SI) are applied, demonstrates relatively low Attack Success Rates (ASR) and high mIoU values, indicating limited effectiveness in degrading the segmentation performance of the SAM models. Specifically, the baseline achieves a mIoU of 49.98 for SAM-B, 71.49 for SAM-L, and 75.24 for SAM-H, with low ASR@10 values.

Adding the RGM significantly improves the performance, resulting in decreased mIoU and increased ASR@50 and ASR@10 across all models. Incorporating Momentum Iteration (MI) and Random Similarity Transformation (RST) further enhances the attack success, achieving a reduction in mIoU to 29.71 for SAM-B and 33.38 for SAM-L, accompanied by notable improvements in ASR values. The final configuration, where all components (RGM, MI, RST, SI) are included, achieves the best performance, with SAM-B's mIoU dropping to 26.87, and ASR@50 and ASR@10 reaching 72.99 % and 55.60 %, respectively. This demonstrates the synergy between the components and their collective contribution to the overall performance of RGA.

5.4.2. Integrating RGM with traditional methods

We further evaluate the impact of integrating the RGM component with traditional adversarial attack methods, such as MIM (Dong et al., 2018) and DIM (Xie et al., 2019). Table 4 presents the results, demonstrating the effectiveness of RGM in enhancing the transferability and robustness of these traditional methods.

From Table 4, it is evident that adding RGM to both MIM and DIM results in substantial performance gains across all metrics on SAM-B (white-box), SAM-L, and SAM-H models. For example, the mIoU for MIM on SAM-B decreases from 54.64 to 46.59 when integrated with RGM, while ASR@50 and ASR@10 improve from 40.92 % to 49.70 % and 19.01 % to 32.04 %, respectively. Similarly, DIM + RGM achieves a reduction in mIoU to 28.32 and improvements in ASR metrics, underscoring the effectiveness of incorporating RGM to enhance traditional attack approaches.

5.4.3. Evaluating different target types

To further evaluate the versatility of our RGA framework, we conduct experiments using different target types for adversarial perturbations, including black, white, random noise, and randomly selected samples from the SA-1000 dataset. The results are summarized in Table 5, which shows that our RGM approach consistently outperforms other target types in terms of mIoU and ASR metrics across all models.

As shown in Table 5, RGM achieves the lowest mIoU and highest ASR values across SAM-B, SAM-L, and SAM-H models. For example, on the SAM-B (white-box) model, RGM achieves a mIoU of 26.87, which is

Table 5
Comparison of different target types for adversarial perturbations, including black, white, random noise, and samples from the SA-1000 dataset. The best performances in each block are shown in **bold**.

Target	SAM-B(white-box)			SAM-L			SAM-H		
	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑
black	34.17	59.90	51.44	40.18	54.79	42.24	44.60	50.18	36.34
white	33.10	61.00	52.90	39.16	55.89	43.78	44.17	50.95	36.60
random noise	32.15	65.65	46.73	37.22	58.75	39.96	42.28	54.91	31.36
SA-1000	30.05	68.71	48.64	34.27	63.62	43.59	36.89	60.85	40.79
RGM	**26.87**	**72.99**	**55.60**	**28.27**	**69.64**	**52.66**	**31.15**	**67.41**	**48.81**

significantly better than the other target types, such as SA-1000, which yields a mIoU of 30.05. The ASR@50 and ASR@10 values for RGM are also notably higher, demonstrating the superior ability of RGM to degrade segmentation quality effectively.

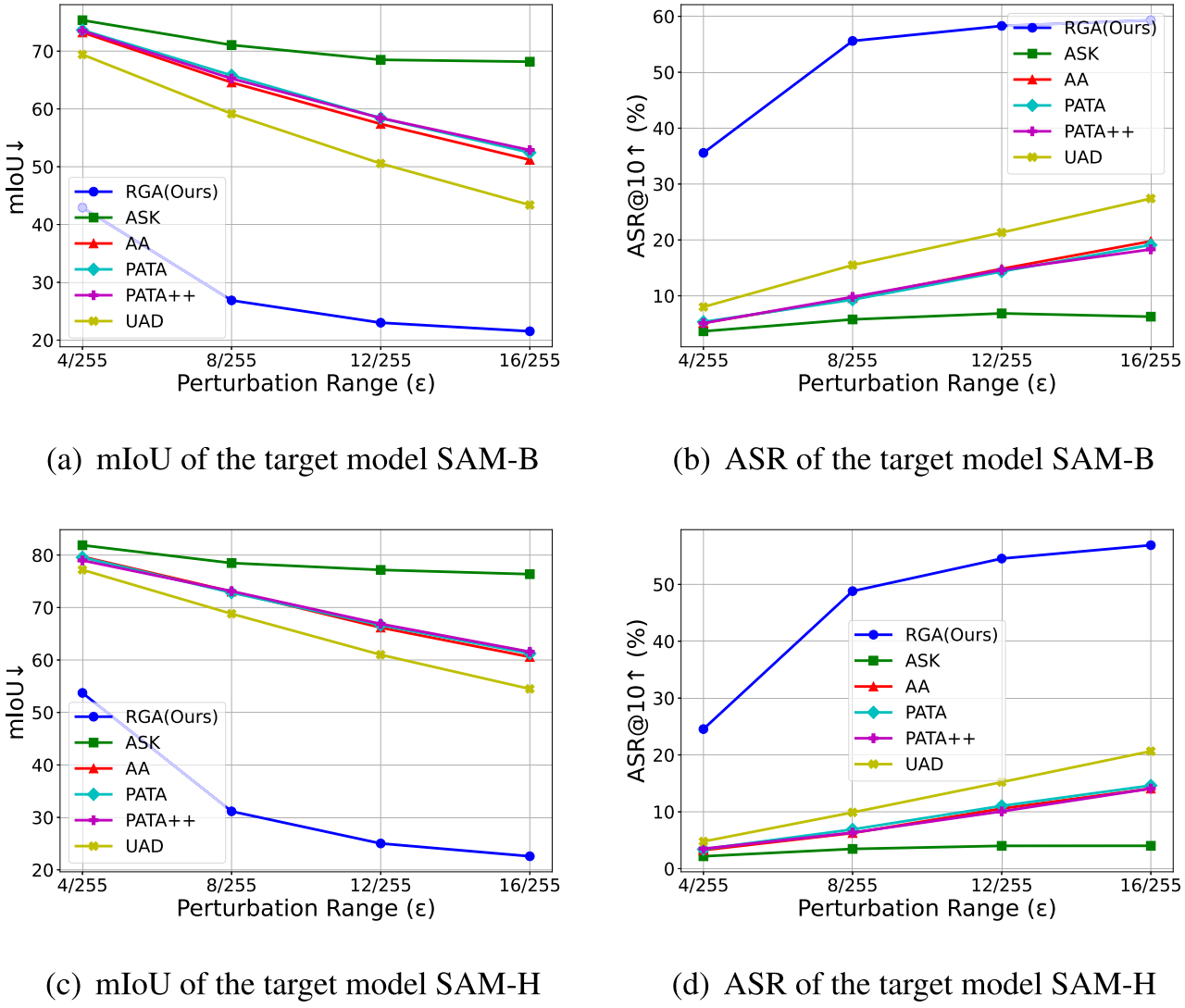
In summary, the ablation studies conducted across the three experiments highlight the importance of each component in the RGA framework, the effectiveness of integrating RGM with traditional adversarial methods, and the superior performance of RGM over other target types. The results validate the strength of our approach in improving the transferability and effectiveness of adversarial attacks against segmentation models.

5.5. Sensitivity analysis of hyper-parameters

In this section, we conduct a sensitivity analysis to evaluate the impact of key hyper-parameters on the performance of our RGA framework. The hyper-parameters examined include the perturbation bound ϵ, the number of adversarial iterations T, and the parameters associated with the Segmentation and Dilation Strategy, specifically the granularity parameter for grid size γ and the number of dilation iterations n.

5.5.1. Perturbation bound ϵ

The perturbation bound ϵ plays a critical role in defining the



(a) mIoU of the target model SAM-B

(b) ASR of the target model SAM-B

(c) mIoU of the target model SAM-H

(d) ASR of the target model SAM-H

Fig. 8. Sensitivity analysis results of the perturbation bound, ϵ.

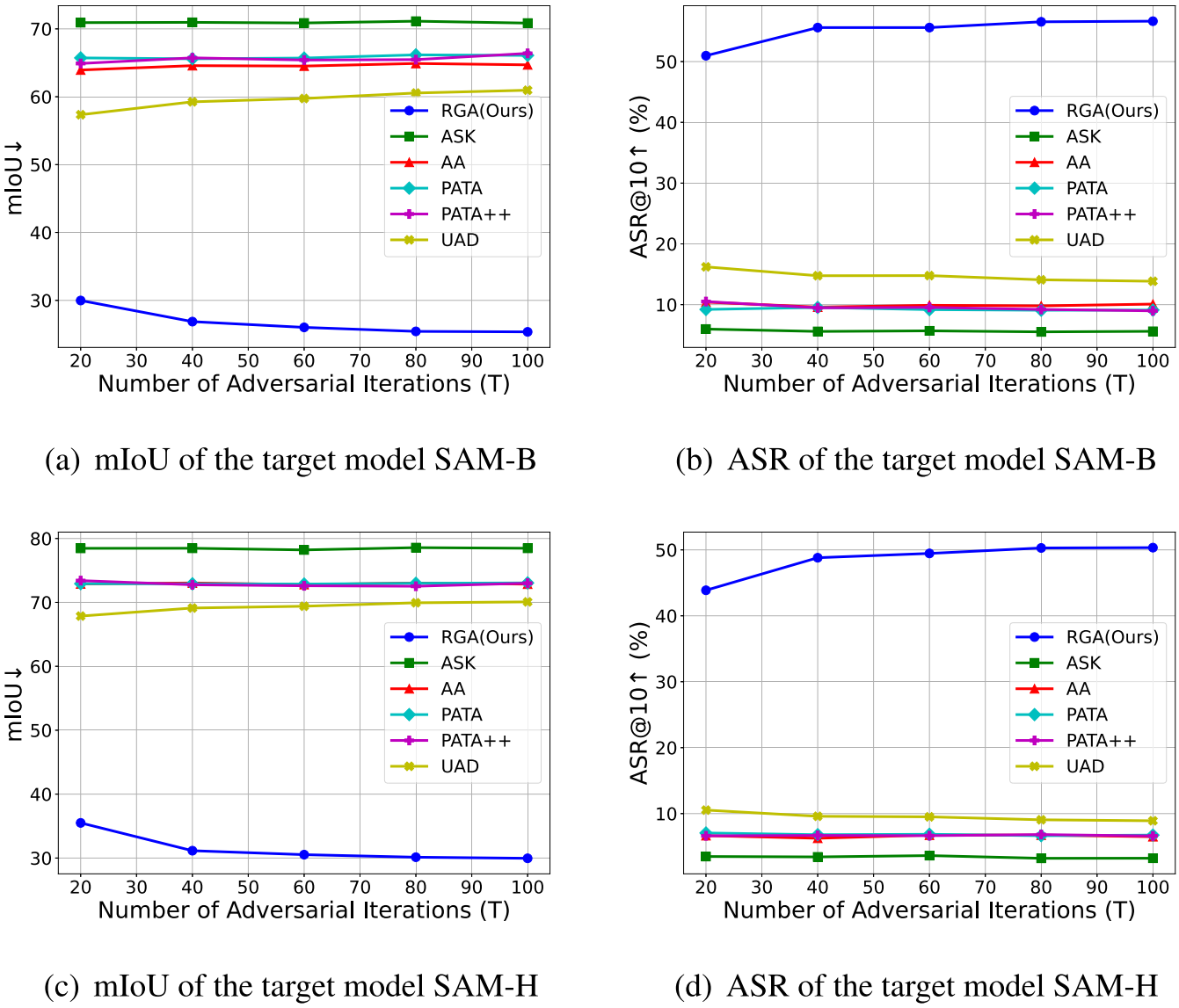


Fig. 9. Sensitivity analysis results of the number of adversarial iterations, T.

maximum allowable distortion applied to the input images. As shown in Fig. 8, we evaluate various values of ϵ to determine how the extent of perturbation affects the attack's effectiveness. Lower values of ϵ generally result in more subtle perturbations, which can help maintain the perceptual quality of the input while still achieving a significant degradation in segmentation performance. Conversely, higher values of ϵ can lead to more aggressive attacks, potentially compromising the model's integrity more effectively but also risking the visibility of the perturbations.

5.5.2. Number of adversarial iterations T

The number of adversarial iterations T refers to the count of optimization steps taken during the perturbation generation process. As shown in Fig. 9, increasing T initially improves the attack's success, as it provides more opportunity for refining the adversarial perturbations. However, after a certain point, increasing T further may lead to overfitting to the specific model, thereby reducing the generalizability of the perturbations. To balance effectiveness and computational cost while avoiding overfitting, we set the default value of T to 40, as it offers a good compromise between attack success and robustness.

5.5.3. Granularity parameter for grid size γ and number of dilation iterations n

Fig. 10 presents the sensitivity analysis results for the granularity parameter for grid size γ and the number of dilation iterations n. The granularity parameter for grid size γ and the number of dilation iterations n are used to generate the Region-Guided Map (RGM), which plays a crucial role in misleading the SAM model. These parameters determine how the segmentation results are manipulated to create the adversarial effect.

The parameter γ controls the size of the grid blocks used for dividing large segmented regions. If γ is too small, the large regions may be excessively subdivided, which could result in an ineffective adversarial influence that fails to mislead SAM consistently. On the other hand, if γ is too large, the subdivision may be too coarse, reducing the effectiveness of targeting specific parts of the large regions, thereby limiting the ability of RGM to mislead SAM effectively. Hence, finding an appropriate value for γ is key to achieving a balance that maximizes the adversarial impact while maintaining the perceptual quality of the guidance.

Similarly, the number of dilation iterations n influences how much smaller segmented regions are expanded. If n is too small, the dilation effect may be insufficient, limiting the impact on the smaller

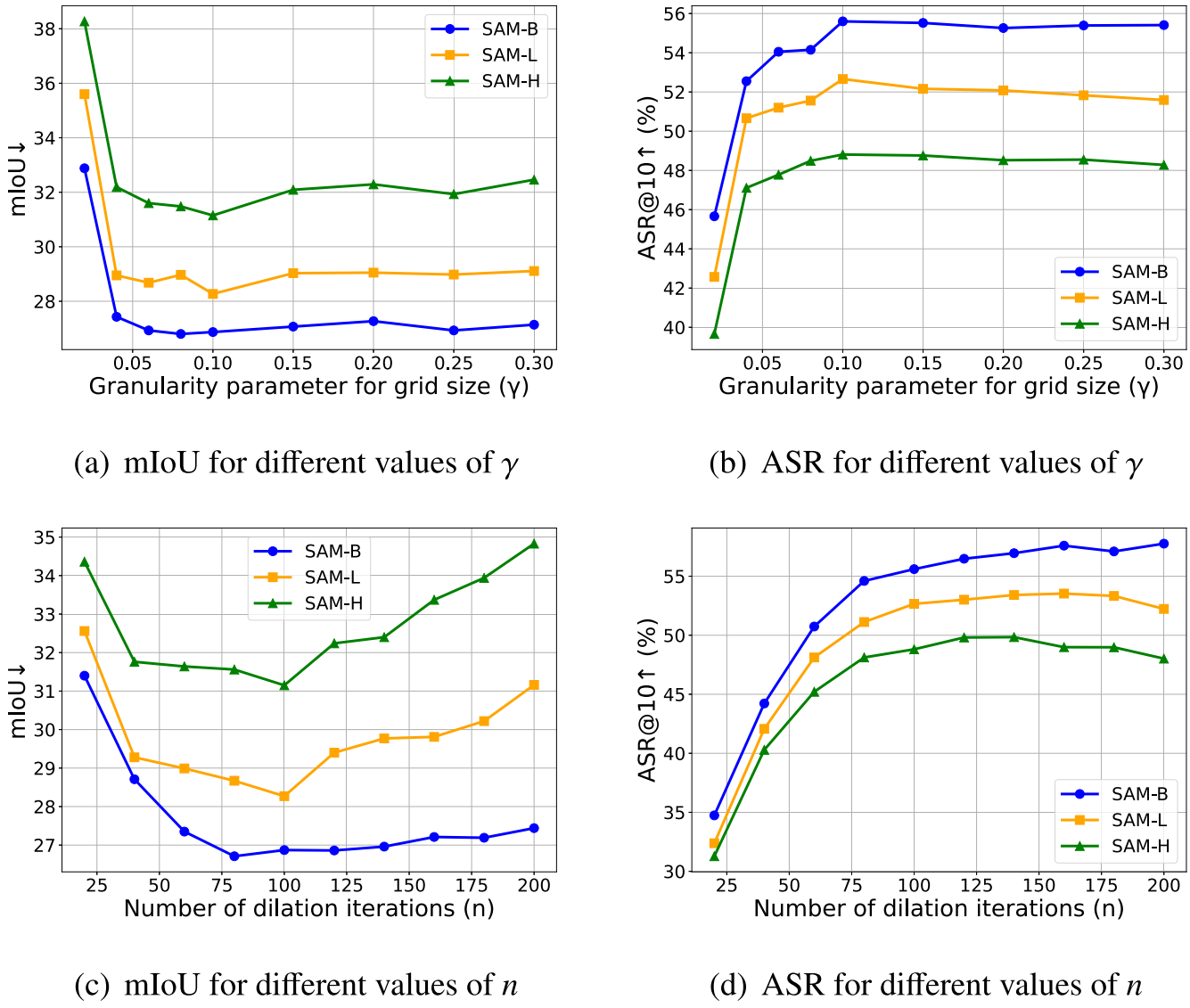


Fig. 10. Sensitivity analysis results of the granularity parameter for grid size, γ, and number of dilation iterations n.

regions that need to be expanded for effective manipulation of the segmentation. Conversely, if n is too large, the dilation could overly expand these small regions, making the adversarial perturbation too obvious or even causing unintended merging of regions, which reduces the precision of the adversarial attack. Thus, an appropriate value for n is essential to ensure that the expansion is effective enough to influence the segmentation without introducing unintended side effects.

In summary, the values of γ and n directly impact the construction of the RGM, which is used to guide SAM to make incorrect segmentations. Both parameters must be tuned carefully to strike a balance between effective misleading of the model and maintaining the subtlety of the perturbations.

5.6. RGA under JPEG compression defense

To evaluate the effectiveness of RGA against input transformation defenses, we examine its performance under JPEG compression. As shown in Table 6, RGA's success rate drops under this defense, as expected. However, when combined with BPDA (Athalye et al., 2018) using a differentiable approximation of the JPEG layer, RGA exhibits improved performance, with ASR@10 exceeding 30 % across all evaluated models. These results indicate that although JPEG compression mitigates RGA's effectiveness, the attack remains potent when equipped with adaptive gradient estimation.

6. Discussion

The Region-Guided Attack (RGA) employs a powerful approach to adversarial attacks on segmentation models by generating the Region-Guided Map (RGM) through its Segmentation and Dilation (SAD) strategy. While RGA demonstrates considerable strength in degrading segmentation quality by manipulating regions based on size, certain limitations restrict its effectiveness in more specialized segmentation scenarios. Here, we discuss two key limitations: challenges with overlapping regions and the difficulty of achieving desired errors with subtle perturbations.

1. Limitations with overlapping regions: In tasks where regions frequently overlap-such as medical imaging, multi-layered materials in industrial inspections, or semi-transparent objects in computer vision-segmentation often requires intricate, multi-layered outputs that capture the depth and interaction of overlapping structures. RGA's SAD strategy, which applies binary transformations (dilation for small regions and fragmentation for large ones), is not inherently equipped to

Table 6
Performance of RGA under JPEG compression defense. The table reports mIoU and ASR@10 for three SAM variants. Results show that JPEG defense weakens RGA, but combining RGA with BPDA (Athalye et al., 2018) partially restores attack effectiveness. Best results in each block are in **bold**.

Method	SAM-B(white-box)		SAM-H		EfficientSAM	
	mIoU↓	ASR@10↑	mIoU↓	ASR@10↑	mIoU↓	ASR@10↑
Clean	67.15	6.02	71.37	5.05	64.75	7.54
RGA	60.07	9.95	64.17	9.88	58.54	10.39
RGA + BPDA (Athalye et al., 2018)	**49.23**	**35.43**	**40.45**	**28.62**	**39.83**	**31.23**

manage these complex, overlapping relationships within the segmentation output.

Overlapping regions require an adversarial approach that can distinguish between intersecting areas and selectively perturb them without compromising the multi-layered representation. However, SAD's current design lacks the granularity to handle overlapping regions effectively. When dealing with complex intersections, SAD's dilation and fragmentation may lead to coarse distortions that fail to influence the segmented regions in a meaningful way. For example, in a layered tissue sample in medical imaging, disrupting the representation of one tissue layer without affecting the layers above or below requires finer control than SAD's region-specific transformations currently offer. As a result, RGA's impact may be limited in these scenarios, as it cannot fully exploit the segmentation model's sensitivity to overlapping structures.

Future iterations of RGA could address this limitation by incorporating more advanced segmentation-sensitive perturbations that can account for multi-layered and semi-transparent relationships within the data. Approaches like hierarchical perturbation strategies or adaptive transformations that respect the contextual interactions of overlapping layers could improve RGA's effectiveness in these high-precision tasks.

2. Subtle perturbations may not yield desired errors: RGA's perturbation approach, while effective for disrupting larger segmented regions or inducing shifts in boundaries, may struggle in cases where minimal yet precise boundary modifications are necessary to degrade the model's output meaningfully. In domains like high-resolution medical imaging or satellite analysis, segmentation accuracy hinges on the precise identification of fine-grained boundaries, where even small misclassifications can have significant implications. However, the SAD strategy's binary handling of regions lacks the subtlety needed to create these precise boundary distortions.

SAD's transformations-dilating small regions and fragmenting large ones-are relatively coarse and may introduce perceptible yet ineffective changes in the segmentation output. When applied to highly detailed regions, these transformations could result in visible but inconsequential alterations that do not meaningfully impact the model's accuracy. For example, in tasks that require distinguishing between intricate, closely related textures or subtle edge boundaries, RGA's SAD-based perturbations may produce overly generalized errors that fail to impact model predictions at the desired level of detail.

Addressing this limitation would require an enhancement to SAD's perturbation methods, allowing it to achieve fine-grained distortions capable of subtly shifting boundaries without creating overtly visible artifacts. Future research could explore integrating more sophisticated techniques such as texture-sensitive perturbations or boundary-preserving modifications to better control the precision of attacks. These enhancements would enable RGA to more effectively target detailed segmentation tasks that demand high fidelity and minimal perceptual distortion.

In summary, while RGA offers a robust, region-guided framework for adversarial attacks, its current SAD strategy presents challenges in tasks involving overlapping regions and in applications where subtle, precise boundary distortions are essential. Future developments in RGA could benefit from adaptive, multi-layered perturbation techniques to handle overlapping regions more effectively and from refined boundary-sensitive transformations to achieve desired segmentation errors with greater subtlety. By addressing these limitations, RGA's applicability could be extended to a broader range of specialized, high-precision segmentation tasks.

7. Conclusion

In this work, we introduced the Region-Guided Attack (RGA), an innovative adversarial attack method designed specifically for segmentation models like the SAM. RGA stands out by generating a Region-Guided Map (RGM) through the Segmentation and Dilation (SAD) strategy, enabling a prompt-agnostic approach that disrupts SAM's segmentation accuracy. SAD tailors the perturbation strategy according to the size and structure of each segmented region, dilating small regions to exaggerate their presence and fragmenting larger regions to destabilize their boundaries. The resulting RGM acts as a blueprint that guides adversarial perturbations in a spatially targeted manner, enhancing the effectiveness and precision of the attack.

Our quantitative and qualitative evaluations demonstrate that RGA significantly degrades segmentation performance across multiple models and scenarios, achieving high attack success rates in both white-box and black-box settings. By leveraging the feature-based structure of SAM, RGA ensures that perturbations are highly transferable and effective even without prompt-based guidance, highlighting a fundamental vulnerability in segmentation models that rely on spatial coherence.

The success of RGA underscores the need for future research to address region-specific adversarial threats within segmentation frameworks. Defensive measures should focus on enhancing segmentation model robustness against feature-guided perturbations, which are particularly challenging to detect and mitigate. Overall, RGA presents a significant advancement in adversarial attack strategies, providing valuable insights into the structural vulnerabilities of segmentation models and setting a foundation for developing more resilient and secure segmentation algorithms in critical applications.

CRediT authorship contribution statement

Xiaoliang Liu: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization; **Furao Shen:** Supervision, Resources, Project administration, Funding acquisition; **Jian Zhao:** Writing – review & editing, Validation, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported in part by the STI 2030-Major Projects of China under Grant 2021ZD0201300, and by the National Science Foundation of China under Grant 62276127.

References

Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning* (pp. 274–283). PMLR.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Cheng, B., Schwing, A., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, *34*, 17864–17875.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

Dong, Y., Liao, F., Pang, T., Su, H., Hu, J. Z., & Li, X. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185–9193).

Dong, Y., Pang, T., Su, H., & Zhu, J. (2019). Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4312–4321).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International conference on learning representations*.

Inkawhich, N., Wen, W., Li, H. H., & Chen, Y. (2019). Feature space perturbations yield more transferable adversarial examples. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Ke, L., Ye, M., Danelljan, M., Tai, Y.-W., Tang, C.-K., & Yu, F. (2023). Segment anything in high quality. *Advances in Neural Information Processing Systems*, *36*, 29914–29934.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026).

Kurakin, A., Goodfellow, I., & Carlin, D. (2017). Adversarial examples in the physical world. In *Proceedings of the international conference on learning representations (ICLR)*.

Li, Q., Guo, Y., Zuo, W., & Chen, H. (2023). Improving adversarial transferability by intermediate-level perturbation decay. In *NeurIPS*.

Lin, J., Song, C., & He, K. (2019). Nesterov accelerated gradient and scale invariance for adversarial attacks. In *Proceedings of the british machine vision conference*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th european conference, zurich, switzerland, september 6–12, 2014, proceedings, part v 13* (pp. 740–755). Springer.

Liu, X., Shen, F., Zhao, J., & Nie, C. (2022). RSTAM: An effective black-box impersonation attack on face recognition using a mobile and compact printer. arXiv preprint arXiv:2206.12590

Liu, X., Shen, F., Zhao, J., & Nie, C. (2024). EAP: An effective black-box impersonation adversarial patch attack method on face recognition in the physical world. *Neurocomputing*, *580*, 127517.

Liu, Y., & Wei, P. (2024). Cross-prompt adversarial attack on segment anything model. In *Proceedings of the 2024 12th international conference on communications and broadband networking* (pp. 34–39).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Lu, J., Yang, X., & Wang, X. (2024). Unsegment anything by simulating deformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 24294–24304).

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications*, *15*(1), 654.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International conference on learning representations*.

Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., & Zhang, Y. (2023). Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis 89*, 102918.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems* (pp. 8026–8037). (*vol. 32*).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Shen, Y., Li, Z., & Wang, G. (2024). Practical region-level attack against segment anything models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 194–203).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., & Yuille, A. (2019). Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2730–2739).

Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F. et al. (2024). Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16111–16121).

Yan, J. et al. (2024). Segment-anything models achieve zero-shot robustness in autonomous driving. arXiv preprint arXiv:2408.09839

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., & Hong, C. S. (2023a). Faster segment anything: Towards lightweight SAM for mobile applications. arXiv:2306.14289

Zhang, C., Zhang, C., Kang, T., Kim, D., Bae, S.-H., & Kweon, I. S. (2023b). Attack-sam: Towards evaluating adversarial robustness of segment anything model. arXiv preprint arXiv:2306.14289

Zhang, K., & Liu, D. (2023). Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., & Wang, J. (2023). Fast segment anything. arXiv preprint arXiv:2306.12156

Zhao, Z. (2023). Enhancing autonomous driving with grounded-segment anything model: Limitations and mitigations. In *2023 IEEE 3rd international conference on data science and computer application (ICDSCA)* (pp. 1258–1265).

Zheng, S., & Zhang, C. (2023). Black-box targeted adversarial attack on segment anything (SAM). arXiv preprint arXiv:2310.10010

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ADE20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5122–5130).

Zhou, W., Hou, X., Chen, Y., Tang, M., Huang, X., Gan, X., & Yang, Y. (2018). Transferable adversarial perturbations. In *Proceedings of the european conference on computer vision (ECCV)* (pp. 452–467).

Zhou, Z., Song, Y., Li, M., Hu, S., Wang, X., Zhang, L. Y., Yao, D., & Jin, H. (2025). Darksam: Fooling segment anything model to segment nothing. *Advances in Neural Information Processing Systems*, *37*, 49859–49880.