Red-Teaming Segment Anything Model

Krzysztof Jankowski[1,*]   Bartlomiej Sobieski[2,*]   Mateusz Kwiatkowski[1]
Jakub Szulc[1]   Michał Janik[1]   Hubert Baniecki[1]   Przemysław Biecek[1,2]

[1]University of Warsaw   [2]Warsaw University of Technology

{kj.jankowski3, b.sobieski}@student.uw.edu.pl

Abstract

Foundation models have emerged as pivotal tools, tackling many complex tasks through pre-training on vast datasets and subsequent fine-tuning for specific applications. The Segment Anything Model is one of the first and most well-known foundation models for computer vision segmentation tasks. This work presents a multi-faceted red-teaming analysis that tests the Segment Anything Model against challenging tasks: (1) We analyze the impact of style transfer on segmentation masks, demonstrating that applying adverse weather conditions and raindrops to dashboard images of city roads significantly distorts generated masks. (2) We focus on assessing whether the model can be used for attacks on privacy, such as recognizing celebrities' faces, and show that the model possesses some undesired knowledge in this task. (3) Finally, we check how robust the model is to adversarial attacks on segmentation masks under text prompts. We not only show the effectiveness of popular white-box attacks and resistance to black-box attacks but also introduce a novel approach - Focused Iterative Gradient Attack (FIGA) that combines white-box approaches to construct an efficient attack resulting in a smaller number of modified pixels. All of our testing methods and analyses indicate a need for enhanced safety measures in foundation models for image segmentation.

1. Introduction

The emergence of foundation models [4, 18, 29] rapidly changed the landscape of the applications of artificial intelligence [21, 49]. Instead of training task-specific models from scratch, foundation models are trained on general data distributions using massive datasets and computing resources. This universal approach leads to emergent capabilities that allow for solving many challenging tasks. Foundation models can also be easily fine-tuned with domain-specific data, greatly improving the performance in complex tasks without losing previously acquired knowledge [15, 37]. However, with great power comes great responsibility, and foundation models are no exception. A highly performant model must be deeply understood so that its limitations are known prior to using it in the real world [33, 34].

Red-Teaming has recently become an integral part of foundation model deployments [37]. This research field aims to evaluate the model from various points of view in order to detect any undesirable biases, inabilities, failure cases, and misalignments [27]. It also proposes a variety of tools to fix the discovered drawbacks [9]. Performing a thorough Red-Teaming analysis is difficult, especially in the case of foundation models, as current machine learning algorithms achieve incomprehensible levels of complexity. Every attempt is doomed to shortcomings but provides important insights into possible means of improvement.

Motivated by these observations, we aim to perform a wide-ranging Red-Teaming analysis of the Segment Anything Model (SAM) [17] – a foundation model for the segmentation task. Since its release, SAM has been already found useful in various tasks, e.g. medical imaging [41]. However, SAM's limitations are still not adequately studied, therefore our goal is to provide novel universal methods that can be applied to testing SAM and other segmentation models especially focusing on robustness, privacy, and vulnerability to adversarial attacks. Specifically,

1. We propose a novel style-transfer-based approach to verify SAM's robustness to different environmental conditions, showing that SAM is prone to realistic image perturbations. This analysis emphasizes that SAM should not be used as an out-of-the-box solution in critical practical scenarios like autonomous driving.

2. We emphasize the importance of considering potential privacy security issues with SAM. Our findings reveal its capability to outperform a random classifier in facial classification tasks, highlighting concerns regarding its application in production-grade systems. This aspect necessitates careful consideration during the design phase of such systems and potential data leakage.

3. We perform an extensive evaluation of SAM with a wide

---

*equal contribution

Red-Teaming Task I: Robustness to style transfer

Phase 1 - automatically generate points of interest & Phase 2 - generate masks from K points of interest

Original image → SAM automatic mask generator → Point proposals and masks → Images in different conditions

Points of interest: center points of k biggest masks (coordinates) → SAM mask predictor (from points) → K masks for each condition / mean IOU between each condition image and normal image

Red-Teaming Task II: Robustness to attacks on privacy

target person

Original labeled images → Images of different celebrities in a 3x3 grid → SAM automatic mask generator → Masks results → precision, recall f1 score

Prompt: "Paul Burell"

Red-Teaming Task III: Robustness to adversarial attacks

Manually collected images → GroundingDino → Obtained bounding boxes → SAM automatic mask generator → Attacked masks

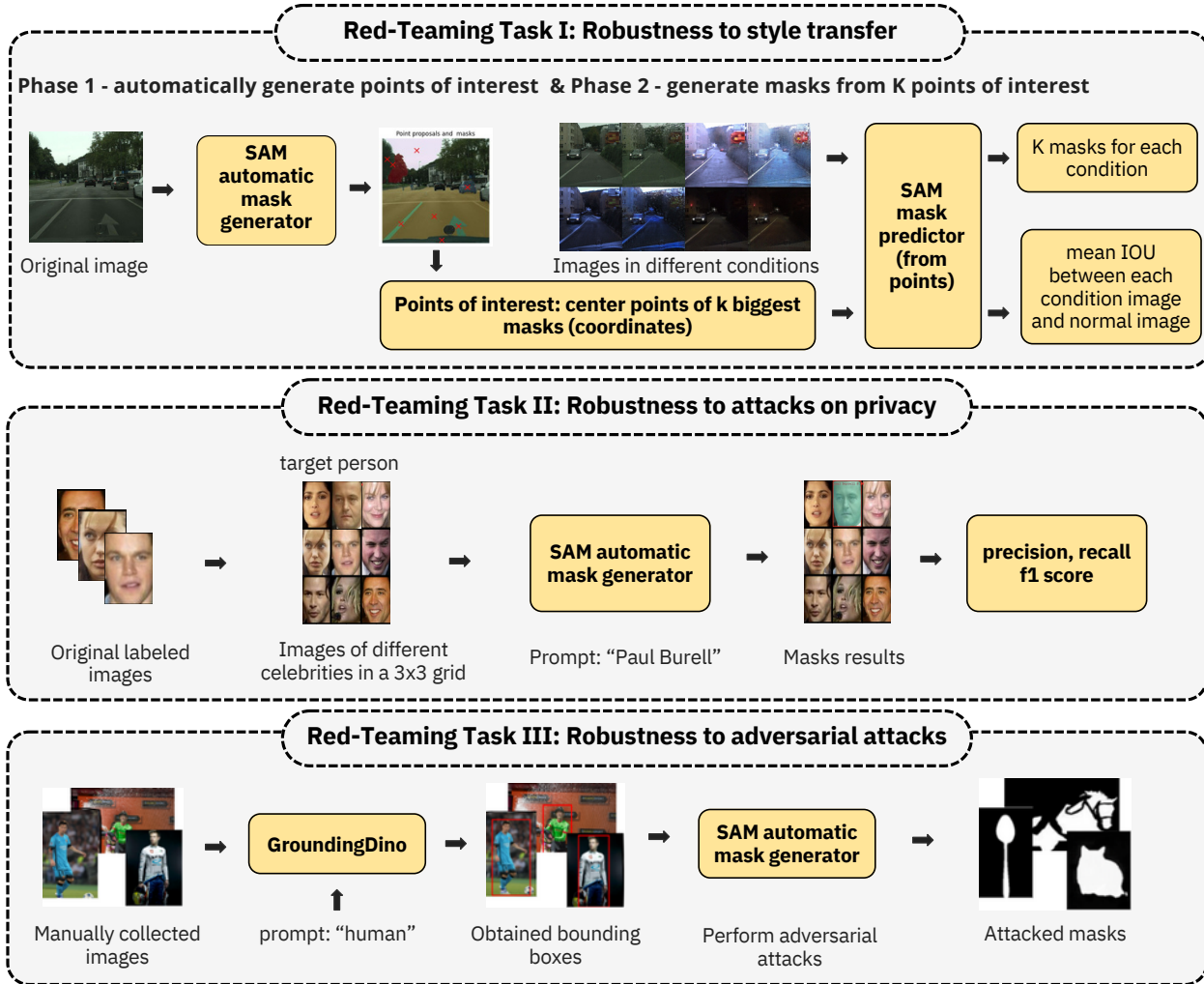prompt: "human"   Perform adversarial attacks

Figure 1. Overview of 3 Red-Teaming tasks on which Segment Anything Model is tested.

range of adversarial attacks. In addition to comparing white-box and black-box methods, we propose Focused Iterative Gradient Attack (FIGA) - a novel hybrid algorithm inspired by the observed vulnerabilities of SAM. Our approach makes the perturbations less noticeable while maintaining the attack's efficacy, further emphasizing the safety concerns.

We believe that our work may inspire a more wide-sighted view on further Red-Teaming of SAM, as well as other foundation models used in critical practical scenarios. We share our code on GitHub.[1]

2. Related work

Segment Anything Model (SAM) [17] is a foundation segmentation model consisting of an image encoder (pre-

trained Vision Transformer (ViT) [8]), flexible prompt encoder, and a transformer decoder. These building blocks enable prompting the model to generate masks from points of interest, bounding boxes, text prompts, or segment the whole image.

Thanks to great performance and ease of fine-tuning, SAM found use cases in many tasks ranging from image editing [33], medical images segmentation [41], or even intelligent vehicles [22, 43]. Due to being used in many critical areas, the model gained attention from researchers focusing on the robustness analysis.

One popular type of analysis covers the model's robustness to different image augmentations [39] unveiling SAM's greater susceptibility to some distortions particularly Gaussian blur, radial blur, and chromatic aberration. This analysis gives more insight into the robustness of SAM, however, most of such distortions will never hap-

---

[1] https://github.com/JankowskiChristopher/red-teaming-segment-anything-model

pen in real life, therefore we focus more on segmentation in style transfer setting [28], especially motivated by another work [45] showing that contrary to human vision, SAM is biased more towards texture rather than shape.

As prior work mostly focuses on changing the texture of the object which might lead to out-of-distribution data and happen extremely rarely in the real world, we shift our focus to a real-life scenario of different weather conditions in street photographs. Similar analyses were performed [35] where adverse weather conditions were generated using Photoshop. However, these modified images are far from actually simulating real-world photos.

Evaluating the model explanation correctness is often performed with the use of bounding boxes that indicate areas on which the explanation should focus. However, in many practical scenarios, such boxes are not available. Merging images into grids is a simple technique that has proven useful in this context [3, 31]. It allows for creating synthetic images with user-defined regions that contain the complete knowledge required to perform a correct prediction. For example, a grid can be constructed from three images of dogs and one image of a cat and used as model input. If the animal classifier predicts the cat class, the model explanation should indicate only the cat image from the grid. Our work relates to this line of research by exploiting grids of images in red-teaming to test SAM's robustness to privacy attacks with no connection to other explanation techniques.

SAM has been also subject to different attacks including white-box [46], black-box [48], and even Universal Adversarial Perturbation (UAP) [25] where a small perturbation resulted in degradation of mask quality in many images [14]. Neural network-based models are particularly vulnerable to adversarial attacks [36], so in our work, we decided to test several approaches to find potential weaknesses in SAM. One of the primary algorithms in this field is the Fast Gradient Sign Method (FGSM) [12], which uses gradients computed over input and, based on these, generates noise to disrupt the model performance. This method is not always immediately successful, so it can be used iteratively, introducing small perturbations [23]. As this is a very popular algorithm, its effectiveness has already been tested on SAM [46]. We decided to make changes to the update rule to try to achieve any desired mask at a lower cost. However, always changing all pixels can sometimes be undesirable, so we also examined attacks on a limited area of the image. An extreme case of this approach is the Jacobian-based Saliency Map Attack (JSMA) [40], in which only one pixel is selected for attack.

The most versatile type of attacks are black-box attacks, as we can also apply them when we do not have access to the weights of a particular version of the model. All we need is the possibility of multiple inferences. Many tools are used in this field to find the right direction of attack, these include random orthonormal vectors [13] and surrogate models [5].

3. Methods

3.1. Robustness to style transfer

Robustness of SAM has been extensively tested on different perturbed images [16, 35, 39] however because these perturbations usually consist of artificial image filters, they are out of distribution data and won't appear in real-life scenarios. Therefore our analysis focuses on testing the robustness of segmentation masks in real-life scenario image perturbations.

To accomplish this goal, we test the quality of generated masks under different styles from the Multi-weather-city dataset [26] which is a modified version of the Cityscapes dataset [6] consisting of dashboard images of streets in different cities. The original Cityscapes images were modified using a set of GAN [11] and CycleGAN [50] methods to obtain image variants under 7 generated adverse weather conditions: night, snow, wet, and additional variants with drops on the windshield.

To reliably assess the quality of generated masks under different styles, we have to make sure that we correctly compare the masks among images. To achieve this goal, as illustrated in Figure 1 Task I, our pipeline starts with original photos and uses SAM automatic mask generation mode to generate all the masks for each image. We use a heuristic of picking k largest masks as they hold information about the most important objects and calculate the center point for each mask. These points are treated as mask IDs to later reliably compare the corresponding masks.

In the second part, we take all the images with the changed style and corresponding coordinates of k points from the original image and we calculate the masks using the SAM mask predictor from points mode. We obtain masks for each image and calculate the mean Intersection over Union (IoU) [32] between the obtained masks and the original masks. We recognize that style transfer changes might lead to alterations in images, including changes in colors and potentially different masks. However, we consider these variations to be minor and not of significant concern.

3.2. Robustness to attacks on privacy

With the growing popularity of computer vision models in downstream tasks, potential privacy concerns have been raised [47]. As SAM can be prompted with text prompts, we examine its capacity to classify celebrity faces. This inquiry is crucial, as it underscores potential misuse implications associated with unintended classification.

Although SAM is open-source, textual prompting is not available in the original source code. Therefore, we use

LangSAM [24], an open–source implementation that connects SAM with GroundingDINO [19] which serves as a text encoder instead of originally used CLIP [30] making textual prompting possible. It also strongly shows how flexible the prompt encoder is and how easily it can be changed.

Our analysis is based on the CelebA dataset [20] with selected 16 different celebrities, evenly split by gender and from diverse backgrounds, to ensure broad representation in our study. As illustrated in Figure 1 Task II, we initiate each iteration of the experiment by randomly selecting nine images of different celebrities' faces and putting them in a 3×3 grid. This deliberate choice of using multiple faces, as opposed to just one, prevents the model from "cheating" by consistently segmenting the sole person in an image; thus, making this method more reliable. We then prompt the LangSAM model with the full name of the celebrity and as a result, obtain the indices of the grid in which the model thinks that the celebrity is located. We only consider indices with bounding boxes where a mask is present as valid.

We evaluate the task as a binary classification i.e., for each face in a grid, we determine whether it is a true positive, true negative, false positive, or false negative. Subsequently, we measure precision, recall, and the F1 score over multiple iterations. To see whether the model robustly classifies celebrities, we repeat the tests and permute faces within each grid 5 times.

We approach LangSAM as an end-to-end system, evaluating its ability to accurately classify celebrities while considering potential biases holistically. These biases could largely originate from the GroundingDINO prompt encoder. However, as we have previously argued, the flexible nature of the prompt encoder, which can be easily interchanged, underscores the importance of analyzing its performance to minimize risks when such a model is deployed.

3.3. Robustness to adversarial attacks

Adversarial attacks aim at producing an imperceptible perturbation in the image that forces the model to return clearly inaccurate outputs. Specifically, for a given model f and image $\mathbf{x}$, adversarial attacks can be generally understood as solutions to the following optimization problem:

$$\min_{\delta} -\|f(\mathbf{x}) - f(\mathbf{x} + \delta)\| + \lambda \cdot \|\delta\|, \qquad (1)$$

i.e. we want to find the smallest perturbation δ that results in the biggest difference between the prediction before and after adding this perturbation to the image. In Equation 1, λ accounts for the trade-off between the norm of the difference of predictions and of the perturbation. In terms of *targeted* attacks, we additionally assume that the perturbation should lead to a specified prediction of the model.

The methods for creating these perturbations can be generally categorized as *white-box*, if they assume access to the model's weights and *black-box* otherwise.

In this paper we focus on white-box methods: Fast Gradient Sign Method (FGSM) [12], Jacobian-based Saliency Map Attack (JSMA) [40] and black-box methods: Simple Black-box Adversarial Attacks (SIMBA) [13] and Ensemble-based Blackbox Attacks (EBAD) [5].

3.3.1 White-box attacks

FGSM operates by making uniform adjustments to all pixels in an image based on the gradient of the loss with respect to the input image. This method is known for its efficiency, in rapidly generating adversarial examples, but it can lead to overt modifications that are easier to detect due to the broad scope of its perturbations. In our approach, we modify the original loss to:

$$X := X - \epsilon \cdot \text{sgn}(\nabla_X \|\sigma(\text{SAM}(p, X)) - \mathbf{Y}\|_2^2) \qquad (2)$$

where X is the image, Y is the target, p is the text prompt, and the epsilon value for pixels in the range $[0, 255]$ is set to 1. The modified formula is better suited for segmentation problems, as it can be useful in solving more sophisticated tasks. We denote the modified algorithm as FGSM*.

On the other end of the spectrum, JSMA takes a more meticulous route than FGSM. It focuses on altering individual pixels one at a time, guided by a saliency map that ranks pixels based on their impact on the model's output. While this approach allows for more discreet perturbations, it is significantly slower and computationally more intensive than FGSM.

In order to reduce the shortcoming of the mentioned methods, we introduce a novel approach Focused Iterative Gradient Attack (FIGA). FIGA is designed to bridge the gap between these extremes. It targets the top k pixels guided by the saliency map and adjusts their values by a fixed amount ϵ in the direction indicated by the gradient.

FIGA allows for a tailored balance: at $k = 1$ the method is equivalent to JSMA and at k equal to the total number of pixels it mirrors FGSM. Choosing a k between the two extremes achieves a strategic balance between efficiency and stealthiness. By targeting a select group of pixels rather than adjusting them all uniformly or individually, our method can generate adversarial examples that are less noticeable than those produced by FGSM but more quickly and with less computational cost than those generated by JSMA. For a pseudocode of a single iteration of our method, see Algorithm 1.

3.3.2 Black-box attacks

SIMBA operates as a black-box technique, employing iterative pixel perturbations with random orthonormal vectors. One of the components of SIMBA is the Discrete Cosine Transform (DCT) [1] that filters out high noise frequencies.

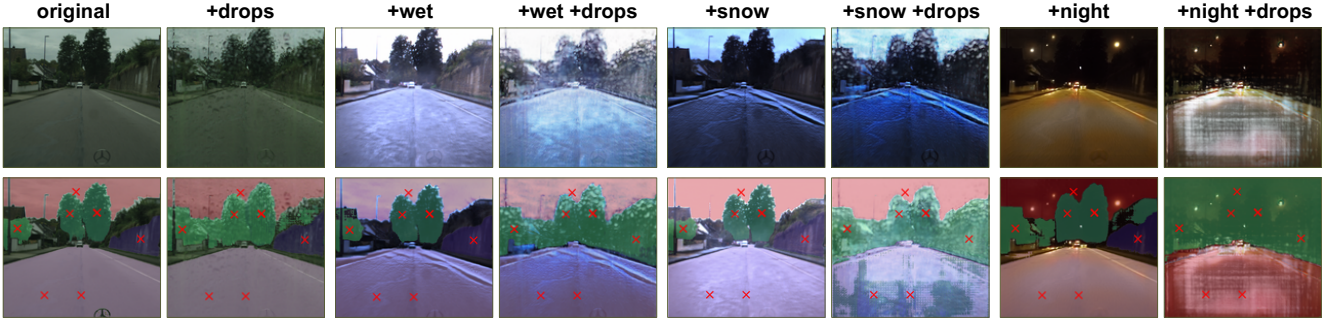| original | +drops | +wet | +wet +drops | +snow | +snow +drops | +night | +night +drops |

Figure 2. Example SAM predictions for the original image and with added different weather conditions. Respective masks have the same colors, but due to opacity and changing colors of the images the masks' colors change as well. Red crosses correspond to the points of interest for which the corresponding masks were generated.

Algorithm 1 FIGA iteration

1: **Input:**
2: X, Y ▷ Input image and target mask
3: $v = \nabla_X J(f(X), Y))$ ▷ gradient v w.r.t. loss J
4: k ▷ Number of pixels to modify
5:
6: **function** FIGA(X, v, k)
7:     topk $\leftarrow$ sorted$(|v|, \text{desc}))[k]$
8:     **for** $i = 1, 2, \ldots$ **do**
9:         **if** $|v[i]| < $ topk **then**
10:             $v[i] \leftarrow 0$
11:         **end if**
12:     **end for**
13:     $X \leftarrow X - \epsilon \cdot \text{sign}(v)$
14:     **return** X
15: **end function**

This component proved effective in enhancing the method's accuracy against sophisticated models like SAM.

Transfer-based attacks use a group of similar surrogate models to create adversarial attacks on a black-box target model. The EBAD algorithm adapts this for image segmentation, enhancing the transferability of attacks from the surrogate models to the target. In EBAD, each iteration starts with generating a perturbation using the PGD [23] method to deceive multiple surrogate models simultaneously, by optimizing a combined prediction score.

4. Results

4.1. Robustness to style transfer

Thanks to its general architecture, our testing method can be applied to various images and style transfer techniques. However, we specifically utilize the Multi-weather-city [26] dataset to evaluate the performance of systems using SAM for intelligent vehicles [43] across diverse road conditions which is crucial for deploying such systems safely.

Figure 3 presents the distributions of mean IOUs between masks for different weather conditions. Each histogram represents the results of the comparison between original images and images with the selected adverse condition applied. We can observe that when conditions are not that harsh i.e. snow and night, the means of the distributions are 0.87 and 0.80 respectively. In wet conditions, the mean decreases to 0.70. When drops are present on the windshield, the performance drops even further to 0.39 (night + drops) and masks often are destroyed as can be seen in Figure 2. All distributions approximately follow the normal distribution and the results align with intuition, showing that as the conditions deteriorate, the performance also declines.

We can further analyze different weather conditions by noticing that conditions: "snow", "night" and "wet" predominantly change the colors but added conditions "drops" influence the texture of the images. This aligns with research showcasing that contrary to intuition SAM is more biased towards texture rather than shape [45]. Therefore masks generated for images with "drops" are significantly worse.

As the presence of the drops on the images is so crucial, we believe that it has to be taken into account when designing algorithms for autonomous vehicles which might not work under heavy rain conditions posing serious risks.

4.2. Robustness to attacks on privacy

The results in Table 1 demonstrate that the model's performance varies significantly depending on the celebrity name used in the prompt.

Notably, certain celebrities, such as *Prince William* and *Paul Burrell*, achieve substantially higher precision, recall, and F1 scores, which is particularly intriguing given their association with the British royal family. Conversely, some celebrities, like *Keanu Reeves*, exhibit scores markedly lower than those expected from a random classifier.

Additionally, the experiment reveals some unique anomalies: names such as *Orlando Bloom*, *Madonna*, and

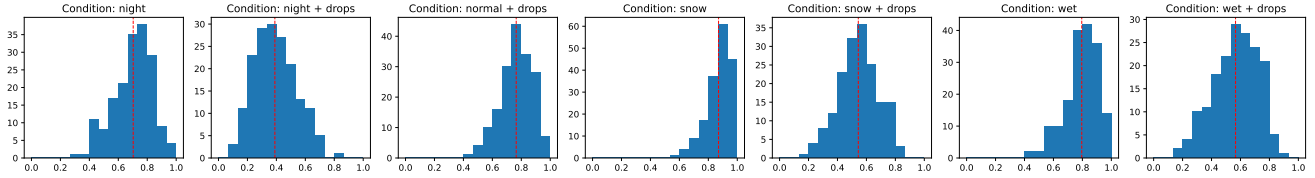| night	night + drops	normal + drops	snow	snow + drops	wet	wet + drops
0.70 ± 0.13	0.39 ± 0.14	0.76 ± 0.11	0.87 ± 0.08	0.54 ± 0.14	0.80 ± 0.11	0.57 ± 0.16

Figure 3. Distributions of mean IOUs between original image masks and augmented image masks for all weather conditions. Red vertical line represents the mean of the distribution. Exact values of mean and standard deviation are presented above the histograms.

Celebrity ♀	Precision (%)	Recall (%)	F1 (%)
Angelina Jolie	22.9 ± 0.8	73.4 ± 2.4	34.9 ± 1.1
Britney Spears	23.2 ± 0.5	$\mathbf{96.6 \pm 1.7}$	37.4 ± 0.8
Jennifer Aniston	15.6 ± 1.5	40.7 ± 4.0	22.6 ± 2.2
Madonna	11.2 ± 0.1	$\mathbf{99.7 \pm 0.5}$	20.1 ± 0.1
Salma Hayek	17.3 ± 0.4	87.4 ± 2.4	28.8 ± 0.7
Jennifer Lopez	24.9 ± 1.7	53.6 ± 3.9	34.0 ± 2.3
Keira Knightley	25.0 ± 2.6	43.0 ± 4.9	31.6 ± 3.4
Nicole Kidman	12.7 ± 0.8	66.0 ± 4.3	21.3 ± 1.3

Celebrity ♂	Precision (%)	Recall (%)	F1 (%)
Keanu Reeves	4.0 ± 1.2	8.7 ± 2.7	5.5 ± 1.7
Orlando Bloom	11.1 ± 0.1	$\mathbf{100.0 \pm 0.0}$	20.0 ± 0.1
Prince William	32.5 ± 2.0	68.0 ± 4.8	$\mathbf{43.9 \pm 2.8}$
Abdullah Gul	25.7 ± 2.2	50.0 ± 4.4	33.9 ± 2.8
Paul Burrell	34.3 ± 1.4	80.6 ± 3.6	$\mathbf{48.1 \pm 1.9}$
Mick Jagger	11.9 ± 0.6	73.1 ± 3.9	20.4 ± 1.0
Michael Jackson	11.1 ± 0.1	$\mathbf{100.0 \pm 0.0}$	20.0 ± 0.1
Nicolas Cage	31.5 ± 2.7	52.1 ± 4.8	39.2 ± 3.4

Table 1. Precision, recall, and F1 scores for celebrity classification, accompanied by their respective standard deviations, in relation to grid cell permutations. Celebrities are divided into equally sized gender groups.

Michael Jackson lead the model to provide less conservative predictions, frequently predicting all images in a grid as containing the target celebrity.

These results indicate that the model exhibits highly variable performance for different celebrity names, deviating significantly from the behavior expected of a random classifier. This underscores the nuanced understanding and bias inherent in the model, which deviates significantly from random chance in its predictions. Example masks for different celebrity names as prompts are depicted in Figure 4.

4.3. Robustness to adversarial attacks

To evaluate SAM's robustness to adversarial attacks in a practical and realistic scenario, we base our evaluation on a case study of segmenting single individuals when prompt-

Figure 4. Example results of segmenting celebrity faces. Images are in a 3x3 grid with a small green cross in the bottom-right corner of the ground-truth image for the text prompt above the grid. Colored images with a red bounding box are the model's answer to the text prompt. In the first and last example, the model correctly segmented the person but introduced false positives. In the middle example, the model did not correctly classify the person.

ing SAM with only textual input. This setting aims at reflecting a simple use case of SAM when the user requires the model to provide accurate segmentation masks when provided with only a very general conditioning signal.

To simulate a real-life setting, we divide our approach into three distinct phases (see Fig. 1 Task III):

1. We first manually choose 100 images of single individuals from the original training dataset of SAM, which should give the model a slight advantage.
2. Using LangSAM, we simulate SAM's textual prompting. Each image is fed into GroundingDINO alongside a fixed descriptor (e.g., "human") to generate a bounding box, ideally encompassing the individual's area. Then, we input the image to the original SAM with the bounding box as a prompt. This design is due to the SAM limitation of not allowing textual prompting. To ensure that only SAM is attacked, we keep the initial bounding box fixed throughout the attack phase.
3. We perform an adversarial attack with the goal of inverting the ground truth mask in the final segmentation.

The effectiveness of each algorithm was assessed with different textual prompts: *person*, *human*, and *man*. For a

qualitative comparison, we include the mean and standard deviation of IoU between the original and attacked mask, Mean-squared Error (MSE), and the infinity norm (L_{inf}) of the difference between the original and the attacked image in Table 2 for each method and prompt.

Method	IoU (%) ↓	MSE ↓	L_{inf} ↓
	prompt: *person*		
FGSM*	1.4 ± 5.2	13 ± 11	11.6 ± 8.9
FIGA	1.7 ± 2.2	7 ± 4	254.2 ± 2.35
SIMBA	67.0 ± 23.0	2927 ± 1699	226.4 ± 45.7
EBAD	96.0 ± 4.0	29549 ± 2403	247.6 ± 3.7
	prompt: *human*		
FGSM*	5.7 ± 19.6	13 ± 11	11.6 ± 8.9
FIGA	19.0 ± 23.4	7 ± 4	254.2 ± 2.4
SIMBA	67.5 ± 23.8	2928 ± 1699	226.0 ± 45.7
EBAD	94.0 ± 4.0	27322 ± 1849	243.2 ± 3.2
	prompt: *man*		
FGSM*	3.3 ± 12.3	13 ± 11	11.6 ± 8.9
FIGA	19.0 ± 24.3	7 ± 3	254.2 ± 2.4
SIMBA	68.1 ± 22.9	2928 ± 1698	227.1 ± 46.1
EBAD	93.0 ± 3.0	28199 ± 2132	244.3 ± 2.8

Table 2. Results for white-box and black-box attacks on SAM. The mean and standard deviation for the prompts: *person*, *human*, *man* are presented.

4.3.1 White-box attacks

We began our tests by comparing the white-box methods. Results indicate that our algorithm FIGA requires less perturbative noise to achieve comparable or superior adversarial effects when measured against modified FGSM, based on MSE comparisons. Additionally, our method outpaces JSMA in execution times, offering a more efficient alternative for generating adversarial examples.

Through extensive testing utilizing the Optuna framework for hyperparameter optimization [2], we identified optimal settings for FIGA at $k = 2653$ and $\epsilon = 5$. These parameters balance the effectiveness of adversarial perturbations with minimal noise introduction and expedited computation.

Our proposed method only makes changes on a limited number of pixels, so some of the changes can be very large, as can be seen in the norm L_{inf}. In terms of this metric, our modified FGSM algorithm performs better, as it generates noise that is spread over the surface of the entire image. The differences between the original images and the generated attacks are invisible to the human eye, as shown in Figure 5.

Initially, we generated an attack for each image in our dataset and prompt "person" and then we tested the result-

Figure 5. Examples of attacks with perturbations small enough to be imperceptible by the human eye, created using an FGSM-based approach. The attack successfully destroys the original masks.

ing adversarial examples for other prompts ("human" and "man"). In this comparison the FGSM performed very well, demonstrating the robustness of this algorithm.

Thanks to the changes we have made to the FGSM, we have influence over what mask will be generated during an attack. The aim of this algorithm is to minimize the distance between the generated mask and any target, so we also performed attempts to display text on the mask. Example results can be seen in Figure 6. Unfortunately, generating this type of attack requires many iterations, so we performed these tests on a smaller scale.

Figure 6. Examples of targeted attacks, generated using an FGSM-based approach. Masks can be changed to an arbitrary text.

4.3.2 Black-box attacks

We carried out analogous experiments for black-box methods. As can be seen in Table 2, SIMBA outperformed EBAD in our tests despite the higher visibility of perturbations. However, the quantitative results clearly indicate that both methods fail to produce successful attacks, showcasing the difficulties black-box methods encounter with complex models like SAM.

Interestingly, EBAD is unable to achieve a mask different from the original (on average 96% of IoU) regardless of the prompt used. We hypothesize that the implicit assumptions of the algorithm are the main cause since the ensemble should consist of models similar to the victim model. However, models from the original experiments, which we also use, possess significantly fewer parameters than SAM (e.g. 44.5M of ResNet101 [44] vs. 636M of SAM).

4.3.3 Robustness of attacks

We have already shown that FGSM and FIGA can generate effective attacks by making only small changes that are invisible to the human eye. It is then worth seeing how robust the examples achieved at such a small cost are. To test this, we generated random Gaussian zero-centered noise with increasing standard deviation for all non-targeted attacks from our prepared dataset. One might have expected that perturbing the attacks would clearly weaken their effectiveness since the aim of both white-box methods is to find a local minimum beyond which a good model should behave normally. This conjecture coincides with our results for the FIGA algorithm, as disrupting the images caused a sudden drop in attack quality. However, the results we obtained for FGSM show that even after introducing a small noise, the attacks are effective, which can be seen in Figure 7.
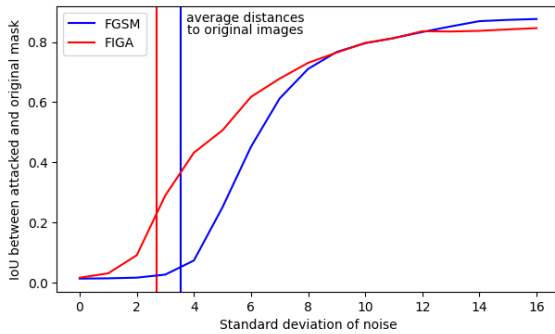
Figure 7. Performance of FGSM and FIGA attacks measured against added noise with increasing standard deviation. IoU between the attacked mask and the original mask increases when adding noise with greater standard deviation. The vertical lines show what is the average distance from the adversarial example to the original image (for both methods separately).

The average distance (in L_2 norm) of the adversarial example generated using FGSM from the original image is only 3.6. When the Gaussian noise of this standard deviation is added, the average attack is still successful, which means that the random noise is not an effective defense. Only the introduction of large noise restores proper segmentation which also shows that SAM can robustly segment even noisy images. An example of this can be seen in Figure 8.

5. Defense strategies

We hypothesize that the performance of SAM on images of roads modified with style transfer could be improved by fine-tuning SAM with specially augmented photos [38]. Undesired memorization of celebrities' faces could be due to features appearing in large datasets and be potentially resolved by better filtering of the datasets before training,
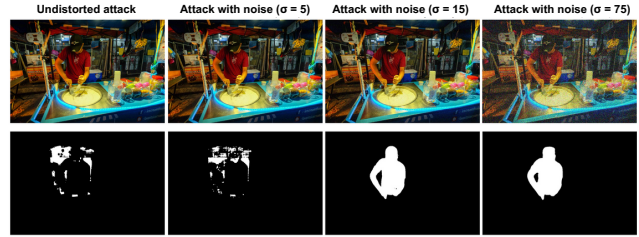

Figure 8. Behaviour of the FGSM attack when adding noise with increasing standard deviation. Big values of standard deviation prevent the attack from destroying the mask.

identifying neurons that hold most of such information [10] and resetting them or other unlearning techniques [42]. Finally, we believe that robustness against white-box attacks could be largely increased by adversarial training [23] or fine-tuning the model in a special adversarial setup suited for segmentation tasks [7]. Future research can focus on analyzing and applying such defense techniques to make SAM more robust.

6. Conclusion

Segment Anything is a useful foundation model used as a component in many complex pipelines solving real-world problems. Throughout this paper, we argue that one has to be careful when using SAM and rely on explainability analysis to assess what are the strengths and weaknesses of this model in specific tasks. For example, we demonstrated that SAM is robust to changing weather conditions unless they get very extreme and feature big distortions in images caused by the presence of water drops.

Moreover, we show that the model's embeddings created by LangSAM hold specific details about certain celebrities. It becomes very concerning due to the potential for misuse and invasion of peoples' privacy. Our study highlights the effectiveness of white-box attack methods in identifying vulnerabilities in SAM. It underscores the challenges in attacking large-scale models without access to their internal parameters, pointing to the need for robust defenses of machine learning systems. We believe the analysis shown in this paper could be repeated on subsequent segmentation models to evaluate their robustness.

7. Acknowledgments

References

[1] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. 4

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *KDD*, 2019. 7

[3] Anna Arias-Duart, Ferran Parés, Dario Garcia-Gasulla, and Victor Gimenez-Abalos. Focus! rating xai methods and finding biases. In *FUZZ-IEEE*, 2022. 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[5] Zikui Cai, Yaoteng Tan, and M. Salman Asif. Ensemble-based blackbox attacks on dense prediction. In *CVPR*, 2023. 3, 4

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3

[7] Francesco Croce, Naman D Singh, and Matthias Hein. Robust semantic segmentation: Strong adversarial attacks and fast training of robust models, 2023. 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[9] Irena Gao, Gabriel Ilharco, Scott Lundberg, and Marco Tulio Ribeiro. Adaptive testing of computer vision models. In *ICCV*, 2023. 1

[10] Amirata Ghorbani and James Y Zou. Neuron shapley: Discovering the responsible neurons. *Advances in neural information processing systems*, 33:5922–5932, 2020. 8

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2015. 3, 4

[13] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *ICML*, 2019. 3, 4

[14] Dongshen Han, Sheng Zheng, and Chaoning Zhang. Segment anything meets universal adversarial perturbation. *arXiv preprint arXiv:2310.12431*, 2023. 3

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 1

[16] Yihao Huang, Yue Cao, Tianlin Li, Felix Juefei-Xu, Di Lin, Ivor W. Tsang, Yang Liu, and Qing Guo. On the robustness of segment anything. *arXiv preprint arXiv:2305.16220*, 2023. 3

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023. 1, 2

[18] Percy Liang and Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2022. 1

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 4

[21] 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

[22] Zihan Ma, Yongshang Li, Ronggui Ma, and Chen Liang. Unsupervised semantic segmentation of high-resolution UAV imagery for road scene parsing. *arXiv preprint arXiv:2402.02985*, 2024. 2

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 3, 5, 8

[24] Luca Medeiros. Language segment-anything. *GitHub repository*, 2023. https://github.com/luca-medeiros/lang-segment-anything. 4

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 3

[26] Valentina Mușat, Ivan Fursa, Paul Newman, Fabio Cuzzolin, and Andrew Bradley. Multi-weather city: Adverse weather stacking for autonomous driving. In *ICCV*, 2021. 3, 5

[27] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *EMNLP*, 2022. 1

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

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[31] Sukrut Rao, Moritz Böhle, and Bernt Schiele. Towards better understanding attribution methods. In *CVPR*, 2022. 3

[32] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2

[34] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *USENIX Security*, 2023. 1

[35] Xinru Shan and Chaoning Zhang. Robustness of segment anything model (SAM) for autonomous driving in adverse weather conditions. *arXiv preprint arXiv:2306.13290*, 2023. 3

[36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 3

[37] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[38] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. A survey on the robustness of computer vision models against common corruptions, 2024. 8

[39] Yuqing Wang, Yun Zhao, and Linda Petzold. An empirical study on the robustness of the segment anything model (SAM). *arXiv preprint arXiv:2305.06422*, 2023. 2, 3

[40] Rey Wiyatno and Anqi Xu. Maximal jacobian-based saliency map attack. *arXiv preprint arXiv:1808.07945*, 2018. 3, 4

[41] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical SAM adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 1, 2

[42] Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *arXiv preprint arXiv:2308.07061*, 2023. 8

[43] Weihao Yan, Yeqiang Qian, Xingyuan Chen, Hanyang Zhuang, Chunxiang Wang, and Ming Yang. SAM4UDASS: When SAM meets unsupervised domain adaptive semantic segmentation in intelligent vehicles. *arXiv preprint arXiv:2401.08604*, 2023. 2, 5

[44] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 7

[45] Chaoning Zhang, Yu Qiao, Shehbaz Tariq, Sheng Zheng, Chenshuang Zhang, Chenghao Li, Hyundong Shin, and Choong Seon Hong. Understanding segment anything model: SAM is biased towards texture rather than shape. *arXiv preprint arXiv:2311.11465*, 2023. 3, 5

[46] Chenshuang Zhang, Chaoning Zhang, Taegoo Kang, Donghun Kim, Sung-Ho Bae, and In So Kweon. Attack-SAM: Towards attacking segment anything model with adversarial examples. *arXiv preprint arXiv:2305.00866*, 2023. 3

[47] Ruoyu Zhao, Yushu Zhang, Tao Wang, Wenying Wen, Yong Xiang, and Xiaochun Cao. Visual content privacy protection: A survey. *arXiv preprint arXiv:2303.16552*, 2023. 3

[48] Sheng Zheng, Chaoning Zhang, and Xinhong Hao. Black-box targeted adversarial attack on segment anything (SAM). *arXiv preprint arXiv:2310.10010*, 2024. 3

[49] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. 1

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3