Unsegment Anything by Simulating Deformation

Jiahao Lu Xingyi Yang Xinchao Wang*
National University of Singapore
{jiahao.lu, xyang}@u.nus.edu, xinchao@nus.edu.sg

Abstract

Foundation segmentation models, while powerful, pose a significant risk: they enable users to effortlessly extract any objects from any digital content with a single click, potentially leading to copyright infringement or malicious misuse. To mitigate this risk, we introduce a new task "Anything Unsegmentable" to grant any image "the right to be unsegmented". The ambitious pursuit of the task is to achieve highly transferable adversarial attack against all prompt-based segmentation models, regardless of model parameterizations and prompts. We highlight the non-transferable and heterogeneous nature of prompt-specific adversarial noises. Our approach focuses on disrupting image encoder features to achieve prompt-agnostic attacks. Intriguingly, targeted feature attacks exhibit better transferability compared to untargeted ones, suggesting the optimal update direction aligns with the image manifold. Based on the observations, we design a novel attack named Unsegment Anything by Simulating Deformation (UAD). Our attack optimizes a differentiable deformation function to create a target deformed image, which alters structural information while preserving achievable feature distance by adversarial example. Extensive experiments verify the effectiveness of our approach, compromising a variety of promptable segmentation models with different architectures and prompt interfaces. We release the code at https://github.com/jiahaolu97/anything-unsegmentable.

1. Introduction

The emergence of promptable segmentation models, exemplified by the Segment Anything Model (SAM) [23], has demonstrated astonishing generalization capabilities across unseen data distributions and downstream tasks. Nevertheless, while these models offer remarkable convenience, they also introduce potential risks. They enable covert and effortless content filching, allowing unauthorized users to segment and misappropriate visual content with a single click. This is particularly concerning for artworks, digital designs, or promotional images, as segmenting such content can lead to commercial disputes. On the other hand, combining promptable segmentation models with generative AI techniques empowers users to perform precise in-place image editing or even 3D generation with a high level of realism [9, 38, 54]. Segmentations taken out of their original context can be deceptive and may be misused, leading to unauthorized advertising and the generation of misleading news content, thereby posing potential societal risks.

The driving force behind this work is the need to address the above emerging risks with a proactive technical solution. We introduce an innovative task called "Anything Unsegmentable", which aimed at enhancing image resistance to any promptable segmentation model, consequently thwarting any unlawful attempts at image appropriation or manipulation. In pursuit of this ambitious objective, we propose a new adversarial attack Unsegment Anything by Simulating Deformation(UAD) emphasizing its remarkable transferability, as the adversarial perturbations remain model-agnostic and prompt-agnostic. This means that they can effectively confound segmentation foundational models, irrespective of their specific parameterizations and prompt formats.

While existing research has explored adversarial attacks targeting segmentation models [1, 4, 14, 46], our problem presents unique challenges. Firstly, it deviates from the existing approaches due to fundamental differences in input and output spaces. Semantic and panoptic segmentation models take images as input, producing pixel-level classification results. Conversely, SFMs generate binary masks in response to prompts, which can be either spatial (points, boxes, strokes) or semantic (speech, text, or exemplar references). Due to difference in input and output spaces, we need to devise novel attack objectives instead of encouraging pixel-level misclassifications. Secondly, recent studies [17, 34, 43] have demonstrated impressive robustness against various corruptions, surpassing the capabilities of ordinary segmentation models. Crafting attacks that can effectively transfer across these already robust foundation models poses a considerable challenge.

*Corresponding Author.

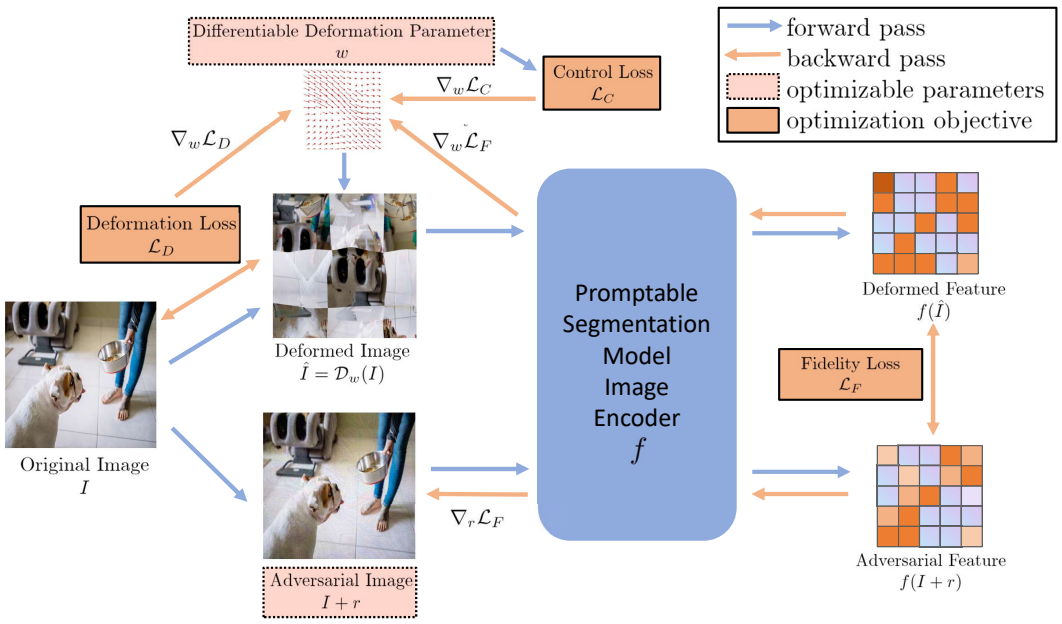

Figure 1. Pipeline of our attack. We optimize a deformation transform as well as the adversarial perturbation, to misguide the promptable segmentation model towards the deformed image.

We intend to present several key findings in this study. Firstly, through an examination of a prompt-specific adversarial perturbation algorithm proposed by concurrent research [56], we observed that adversarial noises derived from specific prompts typically exhibit high variance and lack generalizability across each other. In simpler terms, adversarial noise optimized for a specific prompt tends to overfit to that prompt and does not readily extend to other prompts. Secondly, we reveal that it is better to perturb features along the image manifold than against it to create a transferable adversarial sample. The adversarial noise crafted to shift away from original features in source model may be meaningless to target model and cannot arouse significant feature change, thus limiting their transferability. In contrast, targeted feature attacks bring similar feature disturbance in source and target models. Lastly, we introduce a novel attack UAD, which optimizes an image deformation function as well as the adversarial perturbations. With our approach, the adversarial perturbation introduces shape misinformation, biasing segmentation results towards that particular deformation. Since the deformed image retains some natural image structure, such as textures and object parts, albeit distorted in shape, the feature distortion can be well transferred across segmentation models. This approach allows us to achieve prompt-agnostic and model-agnostic attacks. Empirical results demonstrate the superior effectiveness and transferability of our method compared to concurrent work and prior methods.

In brief, the contribution of this work is three-fold:

1. We introduce a new challenging task *Anything Unsegmentable*, which aims at prompt-agnostic and highly transferable adversarial attacks.
2. We reveal interesting findings on the robustness of promptable segmentation models, including (1) the overfitting nature of prompt-specific attacks and (2) targeted feature disruptions are more transferable than untargeted ones. The findings somehow disagree with previous observations on semantic segmentation models or classifiers, indicating the essential differences of their feature space.
3. We propose a new adversarial attack method *UAD* as a progressive attempt to address the *Anything Unsegmentable* task. We utilize differentiable deformation parameters to get an optimal target deformed image, which possess considerable structural change as well as feasible feature distance for adversarial updates. Compared with existing and contemporary works, our approach achieves state-of-the-art results, showing the effectiveness of our method.

2. On the Robustness of Foundation Segmentation Models

2.1. Objective of Anything Unsegmentable Task

For a promptable segmentation model, it typically contains an image encoder f_{θ^I}, a prompt encoder h_{θ^P} and a mask decoder g_{θ^M}. The promptable segmentation task is designed to return a valid binary mask M given an image I and a

prompt P:

$$M = g_{\theta^M}(f_{\theta^I}(I), h_{\theta^P}(P)). \quad (1)$$

The prompt P offers high flexibility, encompassing spatial prompts such as foreground/background points, rough bounding boxes as well as semantic prompts that include high-level content descriptions like free-form text or memory prompts encapsulating prior segmentation information.

Our goal is to generate quasi-imperceptible noise r to produce an adversarial image $I+r$ which significantly alters its segmentation outcome (e.g., yielding a low Intersection over Union (IoU)) regardless of the prompt applied. Adversarial perturbation (or adversarial noise) r is constrained in a feasible set, typically within an infinity norm ball with radius ϵ, i.e. $\|r\|_\infty \leq \epsilon$. The Anything Unsegmentable task demands that the optimal adversarial perturbation r^* is effective across various prompts and model parameters, formally represented as a solution to the subsequent optimization problem:

$$\begin{aligned} r^* &= \arg\min_{\|r\|_\infty \leq \epsilon} \mathop{\mathbb{E}}_{\{\theta^I, \theta^P, \theta^M\}} \mathop{\mathbb{E}}_{P} \quad IoU(M, M') \\ &= \arg\min_{\|r\|_\infty \leq \epsilon} \mathop{\mathbb{E}}_{\{\theta^I, \theta^P, \theta^M\}} \mathop{\mathbb{E}}_{P} \\ &\quad IoU\{g_{\theta^M}(f_{\theta^I}(I), h_{\theta^P}(P)), g_{\theta^M}(f_{\theta^I}(I+r), h_{\theta^P}(P))\}. \end{aligned} \quad (2)$$

2.2. Prompt-specific Attacks Transfer Poorly

As outlined in the preceding section, the attacker's aim is to significantly alter the segmentation mask in response to any prompt. A straightforward idea is to craft an attack which deteriorates the segmentation outcome for a given prompt. This approach was recently employed in Attack-SAM [56]. They introduced an innovative attack objective that minimizes the feature responses within the masked region.

Their technique proved to be potent for the given prompt, however we found it challenging to generalize to alternative, unseen prompts. We offer visual evidence in Fig.7 and qualitative results in Tab. 1 as evidence. Zhang et al. [56] admitted similar findings and they introduced an improvement to enhance the transferability: instead of using only a single prompt, they randomly sample numerous point prompts, then execute the attack to invalidate the ensembled prompts. While this improvement partially alleviates the issue, their subsequent work [59] still demonstrates a noticeable performance gap between attacked prompts and unseen prompts. This observation highlights the challenge of prompt-based attacks being prone to overfit and lack generalizability. We further discuss the heterogenous and overfitting nature of prompt-specific attacks in Appendix Sec.1.

2.3. Perturbations Pointing Inside Image Manifold Transfer Better

To avoid the overfitting behavior of prompt-specific attacks, we, therefore, endeavor to find an alternative approach for prompt-agnostic attacks. Considering image encoders have a more standardized and common functionality compared to prompt encoders or mask decoders [15, 24, 51, 53], we opt to launch the attack from feature space of the image encoder.

There existed many adversarial attack works launching attacks from feature space [8, 16, 19, 26, 44, 45, 60]. Most of them fall into the category of *untargeted feature disruption*, which maximize the distance between adversarial features and original features. The remaining fall under *targeted feature perturbation* which brings adversarial sample closer to a specified input within the feature space.

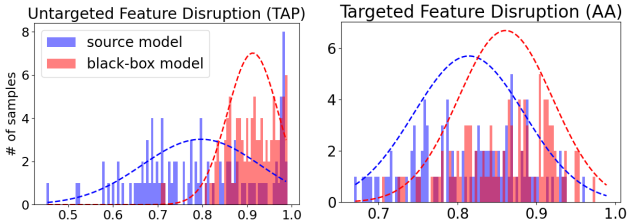

Figure 2. The histogram of feature similarities between adversarial and clean image, induced by *untargeted feature disruption* (left) and *targeted feature disruption* (right) attacks on source (blue) and target (red) model.

Previous literature has not provided clear evidence regarding the superiority of targeted or untargeted feature disruption in the context of classification or semantic segmentation. To our surprise, our investigation reveals that these two approaches exhibit distinct behaviors within the context of foundational segmentation models. We show in Fig.2 that the feature perturbations resulted by untargeted attacks across models are significantly ineffective compared to targeted ones, even when the source and target models share similar architectural designs and training data. We selected TAP [60] as the untargeted feature disruption attack and AA [19] as a targeted attack. By evaluating feature similarity (cosine similarity of vectorized features of clean samples and adversarial samples) on the first 100 images in SAM-1B dataset, we found that targeted feature attacks can arouse similar level of feature disturbance on both models, while untargeted feature attacks can hardly arouse feature shifting on target model, suggesting the guidance from inside the image manifold is indispensable.

We attribute this surprising phenomenon to the inherent differences in tasks and their consequent effects on feature spaces. Classification or semantic segmentation models driven by class-discriminative objectives, tend to yield feature spaces rich in class-sensitive features. Consequently, any deviation from the original features introduces class-sensitive features from other classes, resulting in noticeable feature changes that lead to misclassifications. However in our case, the task doesn't involve category-specific information. Directions shifting away from the original fea-

Figure 3. Adversarial examples crafted by our approach. In each row, we present, from left to right: the adversarial example, the adversarial perturbation, the optimized deformation target, attacked segmentation results and the original segmentation results on SAM-B (using a box prompt), SAM-H (using point prompts), and FastSAM (using a text prompt) respectively. The results demonstrate the high effectiveness of our approach against unseen models and versatile prompts.

tures, may resemble meaningless random noise to target model and neglected by its robustness to image corruptions [18]. In contrast, for in-distribution target images, adversarial perturbation towards them may effectively present a recognized and consistent update direction across all models. Briefly put, the adversarial perturbations pointing inside image manifold transfer better than perturbations pointing outside the image manifold for foundation segmentation models [27].

3. Unsegment Anything by Simulating Deformation

Building upon the insights from the preceding sections, it becomes evident that the target for feature perturbation should be close to natural images. Nevertheless, selecting a random target image from a population of images does not guarantee a sufficient level of structural dissimilarity with the original image, thus offering room for improvement.

We propose that optimization can be not limited to just the adversarial example, it should also encompass the target image. Our approach involves the optimization of a **differentiable image deformation function** applied to the original image to create the target for feature disruption. In essence, it drives the target to exhibit maximal shape deformation through optimization process. Our two-stage pipeline starts by identifying an optimal deformed image that balances high structural dissimilarity with closeness to natural image manifold and feasible set of adversarial samples. In the second stage, we align the features of adversarial sample to features of target deformed image.

3.1. Stage One: Deformation

We aim to make an ideal target deformed image through:

$$\hat{I} = \mathcal{D}_w(I). \tag{3}$$

where $\mathcal{D}_w$ is the deformation function to be optimized and w is the differentiable deformation parameters which controls the deformation function. In practice, the design of $\mathcal{D}_w$ could be highly flexible. We can adopt any form of image transformation (e.g. rotation, translation, scaling, warping) to apply deformation as long as it has parameters to optimize.

In our implementation, we use **flow fields** as w to enable refined image deformation. The flow field w_{ff} indicates a motion vector for each pixel position in the original image I. More specifically, for each pixel position $I^{(i,j)}$ in the original image, the direction of its motion is indicated by flow vector in corresponding position $w_{ff}^{(i,j)} = (\Delta u^i, \Delta v^j)$, and its destination position on deformed image is $\hat{I}^{(i+\Delta u^i, j+\Delta v^j)}$. As the flow vector $(\Delta u^i, \Delta v^j)$ could be fractional numbers and not necessarily integer, we use the

differentiable bilinear interpolation [20] to transform input image with flow field.

Deformation Loss
The primary objective of deformation stage is to optimize a target image with maximal structural dissimilarity with original image, to misguide segmentation results. Any loss which encourages appearance deformations can serve the purpose, for example structural similarity index measure (SSIM) as the loss:

$$\mathcal{L}_D = SSIM(\hat{I}, I). \tag{4}$$

In practice we found that SSIM is easy to achieve zero loss, causing deformation to stall. To encourage greater deformation, we devised a strategy that combines patterns from various parts of the image. We concurrently optimize multiple flow fields, generating several image distortion results, and then combine them using pre-defined filter masks. Only a certain part of each deformation result can pass the filter mask and contribute to the final outcome, resulting a deformation that incorporates patches. This approach has proven more effective than just optimizing a single flow field, since the resulting patch contrastive patterns have even more significant structural differences than a single distorted target.

Control Loss
One advantage of using a flow field lies in its flexibility, allowing us ample room to customize the regulation of the deformation function. For instance, we can apply Total Variation Loss to the flow field to promote locally smooth spatial transformations. Additionally, we can limit the variance of flow vectors to encourage globally uniform deformation like shifting. We employ a combination of variance loss and total variation loss to preserve local image patterns, creating the effect of assembling warped and shifted images:

$$\begin{aligned}\mathcal{L}_C &= \lambda_1 \mathcal{L}_{TV} + \lambda_2 \mathcal{L}_{var} \\ &= \lambda_1 \sum_{p}^{\text{all pixels}} \sum_{q \in \mathcal{N}(p)} \sqrt{\|\Delta u(p) - \Delta u(q)\|_2^2 + \|\Delta v(p) - \Delta v(q)\|_2^2} \\ &+ \lambda_2 \sum_{p}^{\text{all pixels}} \sqrt{\|\Delta u(p) - \sum_{q}^{\text{all pixels}} \frac{\Delta u(q)}{|q|}\|_2^2 + \|\Delta v(p) - \sum_{q}^{\text{all pixels}} \frac{\Delta v(q)}{|q|}\|_2^2}.\end{aligned} \tag{5}$$

Fidelity Loss

As previously noted, not all solutions with low structural resemblance to original image are equally viable targets for adversarial noise to simulate. Intuitively, we believe even though there are infinitely many $\hat{I}$ with low enough deformation loss $\mathcal{L}_D$, their feature distances to the feasible set of I are not evenly distributed. Some targets among others are easier for the adversarial perturbation to approach, which results in better disturbance effect.

We take into the explicit consideration the difficulty for an adversarial example to approach the target deformed image in the feature space. To reduce the feature distance from the deformed target to the feasible set of adversarial examples, we introduce a 'proxy' adversarial sample optimization process. This proxy sample $I_{proxy} = I + r'$ should be close to the boundary of feasible set, so that the distance from $\hat{I}$ to feasible set $\mathcal{N}^{adv}(I)$ can be approximated by the distance from $\hat{I}$ to $I + r'$. We impose a feature fidelity loss as negative cosine similarity between features to regulate the deformation:

$$\begin{aligned}&\mathcal{L}_F(\hat{I}, \mathcal{N}^{adv}(I)) \\ &\approx \mathcal{L}_F(\hat{I}, I + r'^*) \\ &= 1 - \frac{f_{\theta^I}(\hat{I}) \cdot f_{\theta^I}(I + r'^*)}{\|f_{\theta^I}(\hat{I})\|^2 \cdot \|f_{\theta^I}(I + r'^*)\|^2} \\ &\text{where } r'^* = \arg\min_{r'} \mathcal{L}_F(\hat{I}, I + r'^*).\end{aligned} \tag{6}$$

As a conclusion, our desired target deformed image is an optimal solution to the following optimization problem:

$$\hat{I}^* = \arg\min_{\hat{I}} \mathcal{L}_D + \lambda_C \mathcal{L}_C + \lambda_F \mathcal{L}_F, \tag{7}$$

where λ_C and λ_F are coefficients of each loss term. The optimization can be practically solved by gradient descent on differentiable deformation parameters w.

3.2. Stage Two: Feature Simulation

Once we have acquired the optimal target deformed image, the subsequent step aligns with those previous feature perturbation works, to encourage the adversarial perturbation close to a target image. We use the same feature distance measure as Eq. 6. In order to accelerate the feature simulation effect, we encourage minimizing the feature distance between adversarial and target images, meanwhile maximizing the feature distance from original image:

$$r^* = \arg\min_{r} \mathcal{L}_F(\hat{I}, I + r) - \mathcal{L}_F(I, I + r). \tag{8}$$

4. Experiment

4.1. Experiment Settings

Evaluation metrics

We use three metrics to describe the effects of adversarial attack, which are mean Intersection over Union(mIoU), attack success rate at IoU$< 50\%$ (ASR@50) and attack success rate at IoU$< 10\%$(ASR@10). The first metric captures the average attack performance, and latter two capture how many output masks are significantly destroyed by adversarial noise, which serve as worst case measurements.

For robust evaluation to test the cross-prompt generalization, for each adversarial samples, we take randomness over

Attacks \ Models	SAM-B(white-box)			SAM-L			SAM-H			FastSAM		
Evaluation metric	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑
Attack-SAM-K [56]	68.07 ± 28.65	24.10	6.87	77.14 ± 25.05	14.46	4.13	78.71 ± 24.02	12.93	3.51	38.13 ± 40.66	59.90	48.43
TAP [60]	63.49 ± 32.58	29.69	13.12	75.12 ± 27.83	16.96	6.68	77.36 ± 26.21	14.55	5.31	37.67 ± 40.89	60.53	49.45
ILPD [26]	63.21 ± 32.54	30.15	13.09	75.17 ± 27.75	16.82	6.59	77.52 ± 26.02	14.37	5.18	37.84 ± 40.84	60.30	49.02
AA [19]	61.06 ± 32.33	32.48	13.11	70.70 ± 29.74	21.49	8.67	72.87 ± 28.61	19.13	7.39	32.64 ± 39.58	65.86	55.10
PATA [59]	61.36 ± 32.31	32.23	13.04	70.81 ± 29.73	21.27	8.62	73.07 ± 28.49	18.66	7.30	32.74 ± 39.56	65.66	54.97
PATA++ [59]	61.54 ± 32.22	32.00	12.94	71.02 ± 29.55	21.16	8.38	73.16 ± 28.44	18.84	7.22	32.85 ± 39.60	65.69	54.65
UAD (ours)	$\mathbf{51.53 \pm 34.00}$	**43.89**	**20.79**	$\mathbf{66.07 \pm 32.04}$	**26.44**	**12.27**	$\mathbf{68.96 \pm 30.87}$	**23.42**	**10.23**	$\mathbf{28.83 \pm 38.36}$	**69.95**	**59.63**

Table 1. Results of our methods in comparison with prior and contemporary works. Our proposed significantly outperforms other methods in both terms of average mask destruction (low mIoU) and number of drastically affected masks (high Attack Success Rate).

prompts and report their mean and standard deviation of IoUs with annotated masks. For evaluating point prompts, we randomly sample 5 times for each grount-truth mask; for box prompts, we vary bounding box sizes for 3 times, resizing them to 80% or 120% of their original size.

Compared baselines

We carried out the experiments of our proposed attack in comparison with several prior or contemporary works:

1. Attack-SAM-K [56] lowers the feature response globally given K (usually large, e.g. 400) point prompts over the whole image;
2. *Transferable Adversarial Perturbations* (TAP) [60] drives adversarial features away from original features in Minkowski distance;
3. *Intermediate-level perturbation decay* (ILPD) [26] is a refined version of TAP, keeping an effective adversarial direction while possessing a greater magnitude;
4. *Activation attack* (AA) [19] minimizes the distance between the adversarial feature and a target image feature;
5. *Prompt-Agnostic target attack* (PATA) [59] introduces a regularization term to boost the feature dominance of adversarial image over a random clean competition image, on the basis of AA [19].
6. PATA++ [59] is an enhanced version of PATA that addresses the conflict between increasing feature similarity and reducing feature dominance. PATA++ alleviates the issue by randomly pick one new competition image in every adversarial update iteration.

Attack settings

In all the experiments, we set adversarial update steps of final adversarial example to be 40. The results shown in Fig. 3 are adversarial examples crafted using a mild $\epsilon = 12/255$ noise to highlight the attack results. In other experiments, if not explicitly stated, the perturbation range $\epsilon$is set to $8/255$ and perturbation step size α is set to $2/255$. For our attack, we set the proxy perturbation iterationsT_f to be 4, allowing the proxy example to approximately reach the feasible set boundary. The deformation iteration is set to 40.

4.2. Adversarial Examples

We show the effectiveness of our method through some adversarial examples in Fig.3. We have chosen art paintings[1], sculpture designs[2], and personal photos[3] to underscore the practical usage of the *Anything Unsegmentable* task to safeguard digital assets, art copyrights, and portrait rights. All the advresarial samples are crafted on SAM-B model, but they indeed transfer to SAM-H and FastSAM models, regardless of their changes in parameters or architectures.

Upon examining the results presented in Figure 3, it is evident that our attack has a profound impact on the destruction of masks, regardless of whether the prompts are spatial or semantic. In white-box scenarios where the target model is SAM-B, the alterations are most noticeable, with the bounding box prompt highlighting little more than meaningless ripples on the image. Even for black-box models SAM-H and FastSAM, the segmentation masks are significantly distorted. The point prompt and text prompt, which originally could highlight the entire foreground object, now, due to the influence of the deformation target, highlights disjointed parts that cannot form a valid whole object. Importantly, we can observe clear evidence that segmentation results are influenced by the deformed target. For example, in the rightmost figure of the second row, the text prompt "Marble" highlights an area that was originally a background in the clean image. Notably, this wrong segmentation aligns with the deformed target image, where a part of a marble statue is displayed.

4.3. Quantitative Evaluation

We conducted a comprehensive comparison of our approach with all prior works in Tab. 1. All adversarial examples were generated from SAM-B, which has the smallest parameter size among the SAM family of models. Consequently, the other three models (SAM-L, SAM-H, and FastSAM) are considered as black-box models. We selected the SAM-1B dataset for evaluation [23], which includes a wide range of diverse images close to real-world scenarios. We conducted our study on a subset of the SAM-1B, specifi-

[1] Kaggle Artist dataset
[2] Art Images dataset
[3] FLICKR30K dataset

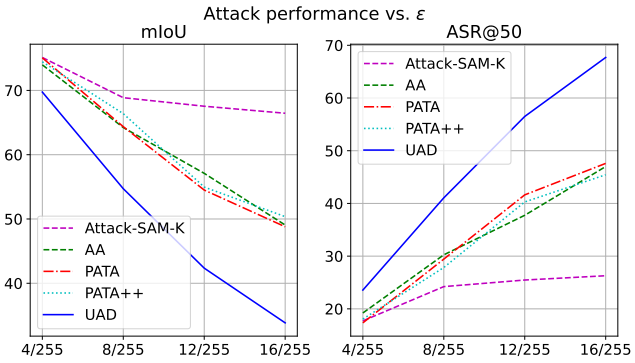

Figure 4. Our attack show consistent superiority under different settings of attack perturbation range ϵ.

cally the first 1000 images (sa_1.jpg to sa_1000.jpg in subset sa_000000.tar). This subset encompasses a total of 98875 masks, which is already a large quantity of masks and has statistical significance.

The results in Table 1 reveal several intriguing facts. Firstly, our proposed attack achieves state-of-the-art results and outperforms other methods by a large margin. Secondly, upon comparing different methods, we observe that targeted feature disruption attacks (AA [19] and PATA [59]) perform significantly better than untargeted feature attacks (TAP [60] and ILPD [26]). This aligns with our earlier analysis, which indicated that optimizing within the feature manifold yields better results than moving away from it. Attack-SAM-K [56] exhibits the weakest performance, further supporting the notion that prompt-specific attacks are less effective than feature perturbation attacks.

Interestingly, we observed that FastSAM performs notably poor, exhibiting a very low mean Intersection over Union (mIoU) and a high portion of masks with drastic changes. This subpar performance might be attributed to the fact that the authors of FastSAM [57] trained their model using only 2% of the SAM-1B dataset. The limited training data likely led to a lack of robustness in their model , making it sensitive to out-of-distribution samples.

4.4. Ablation Studies

4.4.1 On Perturbation Budget ϵ

We conducted experiments using different values of ϵ to assess the effectiveness of our proposed attack across varying perturbation ranges. Remarkably, even when operating within a small perturbation range ($\epsilon = 4$), UAD significantly outperforms other perturbation methods. This demonstrates the superiority and versatility of our approach across a wide range of scenarios.

4.4.2 On Functionality of Loss Terms

We show all loss components in our optimization process $\mathcal{L} = \mathcal{L}_D + \lambda_C\mathcal{L}_C + \lambda_F\mathcal{L}_F$ are necessary and complementary. High-level speaking, first two terms controls deformation to be structural dissimilar($\mathcal{L}_D$) and locally smooth($\mathcal{L}_C$), while last term $\mathcal{L}_F$ constrains adversarial feature distance.

(a) UAD (b) no $\mathcal{L}_D$ (c) no $\mathcal{L}_C$ (d) no $\mathcal{L}_F$

Figure 5. Deformed targets without each individual loss terms.

We visualize deformed targets in Fig. 5 for cases without individual loss terms and illustrate the functionality of each loss term vividly. Without $\mathcal{L}_D$, there would be almost no structural change to misguide segmentation models. Without $\mathcal{L}_C$, the deformation inside each patch goes too wild to contain valid natural shape. This will cause a target image away from the image manifold, resulting in sub-optimal attacks. Without $\mathcal{L}_F$ the attack is less effective, which we will further illustrate in the next subsection.

4.4.3 On Proxy Adversarial Update Steps T_f

In each deformation step, we calculate the feature distance between the deformed target and a proxy adversarial sample as an estimation of the feature distance from deformed target to the feasible set of adversarial images. We have observed that increasing the number of iterations leads to a more reachable target from the feasible set, allowing for better control over the deformation's development. However, it's essential to strike a balance between achieving finer adversarial attack performance and incurring a higher computational overhead.

The ablation study on T_f presented in Fig. 6 underscore the importance of incorporating fidelity loss. We illustrate the changes in relative feature similarity, which is defined as the feature similarity of $I+r$ to $\hat{I}$ minus the feature similarity between $I+r$ and I. Our primary focus here lies in the last column of the plotted figure, when the deformation target has been optimized and fixed, and the final adversarial example is obtained by T steps to simulate the deformation. The variations in values within the last column indicate how closely the adversarial image can approach the target by the end of the optimization process.

When not using any proxy adversarial samples when updating deformation, the relative feature similarity is obviously lower (black mark in the plot). With more proxy adversarial updates integrated into the process, the final relative feature similarity increases. We found that increasing proxy adversarial iterations has diminishing returns. Compared to using the actual attack step count ($T_f = T = 40$) to estimate the distance at each step, using fewer steps only marginally reduces relative feature similarity. In practice,

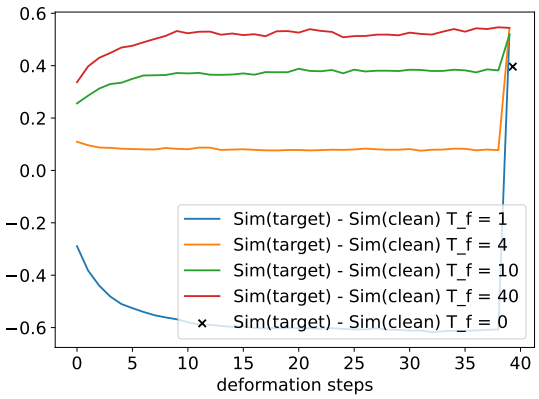

Figure 6. The development of relative feature similarity during deformation steps. More proxy adversarial iterations result in a closer feature distance between deformed target and adversarial example, coming at the cost of heavier computation burden.

to avoid introducing excessive computational burden, we choose $T_f = 4$ as a balanced point that maintains simulation effectiveness while ensuring efficiency.

We put more ablation studies in the appendix.

5. Related Work

Promptable Segmentation Models. In recent developments of segmentation techniques, there has been a shift from closed-set, non-interactive segmentations [2, 13, 22, 36, 37, 55] to more open-vocabulary and interactive settings. This evolution allows for a wide range of interaction forms, including clicks [3, 31, 50], bounding boxes [23], scribbles [61], text [5, 10, 23, 48, 49], or contextual information [42, 61]. Notably, Segment Anything Model (SAM) [23] has the most remarkable zero-shot ability due to its massive training dataset containing 11 million images and 1.1 billion masks. Scaling up training data consequently result in a highly generalized and robust model [18, 23, 35], that nowadays people call them *Vision Foundation Models*. Subsequent research improved SAM in terms of quality [21], latency [57] and semantic awareness [25].

Adversarial Attacks for Segmentation Models. While most adversarial attack research focuses on classifiers, some work extends these attacks to segmentation models. [1] evaluated adversarial attacks on semantic segmentation models, finding that these models are more robust than classifiers due to their multi-scale processing. [46] introduced Dense Adversary Generation, encouraging incorrect recognition on multiple targets simultaneously. [14] proposed generating universal perturbations guiding networks to create desired target segmentations. [4] discussed stealthy attacks on segmentation models, altering targeted labels while keeping non-targeted labels intact. [11] suggested an efficient attack reducing the number of iterations. These methods mainly focus on closed-vocabulary, non-promptable segmentation models for per-pixel classification. Very recently, we notice there are also some work discussing the adversarial attacks on SAM [12, 18, 35, 56, 59]. Zhang et al. [56] designed the first adversarial attack towards SAM which optimizes the input image to have negative feature responses in mask area. SAM-UAP [12] introduces universal adversarial perturbations towards SAM from a contrastive learning perspective. [59] propose Prompt-Agnostic Targetted Adversarial Attacks(PATA) to generate more transferable samples. In our paper, we selected Attack-SAM and PATA as two important baselines and show that our method exhibits higher effectiveness and transferability.

Transferability of Adversarial Attacks. Adversarial examples often struggle to transfer successfully across various neural networks, yielding low success rates in black-box settings. Previous research has explored methods to craft more transferable adversarial samples. These approaches encompass techniques like applying gradient momentum [6, 29, 41], input augmentation [7, 33, 39, 40, 47], feature disturbance [8, 16, 19, 26, 44, 45, 60], and model ensembling [28, 32]. For a comprehensive and up-to-date survey, readers are referred to [58]. These research directions offer various strategies for boosting the transferability of adversarial attacks [52].

6. Conclusion

In this study, we present a novel challenge termed *Anything Unsegmentable*, aimed at generating highly transferable, prompt-agnostic adversarial examples. These examples are designed to shield individuals from the potential risks of copyright and privacy violations posed by foundational promptable segmentation models. Our method *Unsegment Anything by Simulating Deformation* (UAD), marks a progressive step towards addressing this challenge, outperforming existing and concurrent approaches. Our analysis of the robustness of foundational segmentation models uncovers two compelling insights: (1) prompt-specific attacks struggle with transferability, and (2) targeted feature perturbations towards natural-image-like samples, are significantly more effective than untargeted perturbations that drive features away from their original location. We hope that our work will provide valuable perspectives on the resilience of these powerful vision models and inspire future research to mitigate the societal issues they may engender.

7. Acknowledgement

This project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-023), and the Singapore Ministry of Education Academic Research Fund Tier 1 (WBS: A-0009440-01-00).

References

[1] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018. 1, 8

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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, 2017. 8

[3] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 8

[4] Zhenhua Chen, Chuhua Wang, and David Crandall. Semantically stealthy adversarial attacks against segmentation models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4080–4089, 2022. 1, 8

[5] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 8

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 8, 1, 2

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

[8] Aditya Ganeshan, Vivek BS, and R Venkatesh Babu. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8069–8079, 2019. 3, 8

[9] Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Editanything: Empowering unparalleled flexibility in image editing and generation. In *Proceedings of the 31st ACM International Conference on Multimedia, Demo track*, 2023. 1

[10] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 8

[11] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer, 2022. 8

[12] Dongshen Han, Sheng Zheng, and Chaoning Zhang. Segment anything meets universal adversarial perturbation. *arXiv preprint arXiv:2310.12431*, 2023. 8

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 8

[14] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 2755–2764, 2017. 1, 8

[15] Szu-Yeu Hu, Andrew Beers, Ken Chang, Kathi Höbel, J Peter Campbell, Deniz Erdogumus, Stratis Ioannidis, Jennifer Dy, Michael F Chiang, Jayashree Kalpathy-Cramer, et al. Deep feature transfer between localization and segmentation tasks. *arXiv preprint arXiv:1811.02539*, 2018. 3

[16] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2019. 3, 8

[17] 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. 1

[18] 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. 4, 8

[19] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 6, 7, 8

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 5

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

[22] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 8

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 6, 8

[24] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 3

[25] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 8

[26] Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Improving adversarial transferability by intermediate-level perturbation decay. In *NeurIPS*, 2023. 3, 6, 7, 8

[27] Qian Li, Yuxiao Hu, Ye Liu, Dongxiao Zhang, Xin Jin, and Yuntian Chen. Discrete point-wise attack is not enough: Generalized manifold adversarial attack for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20575–20584, 2023. 4

[28] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 11458–11465, 2020. 8

[29] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020. 8

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[31] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023. 8

[32] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *International Conference on Learning Representations,International Conference on Learning Representations*, 2016. 8

[33] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. *ECCV 2022 Oral*, 2022. 8

[34] Yu Qiao, Chaoning Zhang, Taegoo Kang, Donghun Kim, Shehbaz Tariq, Chenshuang Zhang, and Choong Seon Hong. Robustness of sam: Segment anything under corruptions and beyond. *arXiv preprint arXiv:2306.07713*, 2023. 1

[35] Yu Qiao, Chaoning Zhang, Taegoo Kang, Donghun Kim, Shehbaz Tariq, Chenshuang Zhang, and Choong Seon Hong. Robustness of sam: Segment anything under corruptions and beyond. *arXiv preprint arXiv:2306.07713*, 2023. 8

[36] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 8

[37] Sucheng Ren, Xingyi Yang, Songhua Liu, and Xinchao Wang. Sg-former: Self-guided transformer with evolving token reallocation. In *IEEE/CVF International Conference on Computer Vision*, 2023. 8

[38] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023. 1

[39] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021. 8

[40] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021. 8

[41] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. *arXiv preprint arXiv:2103.10609*, 2021. 8

[42] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 8

[43] 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. 1

[44] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7639–7648, 2021. 3, 8

[45] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 8

[46] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 1, 8

[47] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 8, 1, 2

[48] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 8

[49] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 8

[50] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep interactive object selection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 8

[51] Xingyi Yang, Daquan Zhou, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. *Advances in neural information processing systems*, 35:25739–25753, 2022. 3

[52] Jingwen Ye, Ruonan Yu, Songhua Liu, and Xinchao Wang. Mutual-modality adversarial attack with semantic perturbation. In *AAAI Conference on Artificial Intelligence*, 2024. 8

[53] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 3

[54] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything:

Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 1

[55] Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 8

[56] Chenshuang Zhang, Chaoning Zhang, Taegoo Kang, Donghun Kim, Sung-Ho Bae, and In So Kweon. Attack-sam: Towards evaluating adversarial robustness of segment anything model. *arXiv preprint arXiv:2305.00866*, 2023. 2, 3, 6, 7, 8

[57] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 7, 8

[58] Zhengyu Zhao, Hanwei Zhang, Renjue Li, Ronan Sicre, Laurent Amsaleg, Michael Backes, Qi Li, and Chao Shen. Revisiting transferable adversarial image examples: Attack categorization, evaluation guidelines, and new insights. *arXiv preprint arXiv:2310.11850*, 2023. 8, 1

[59] Sheng Zheng and Chaoning Zhang. Black-box targeted adversarial attack on segment anything (sam). *arXiv preprint arXiv:2310.10010*, 2023. 3, 6, 7, 8

[60] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018. 3, 6, 7, 8

[61] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 8, 7