Segment (Almost) Nothing: Prompt-Agnostic Adversarial Attacks on Segmentation Models

First Author
Institution1
Institution1 address
firstauthor@i1.org

Abstract—**General purpose segmentation models are able to generate (semantic) segmentation masks from a variety of prompts, including visual (points, boxed, etc.) and textual (object names) ones. In particular, input images are pre-processed by an image encoder to obtain embedding vectors which are later used for mask predictions. Existing adversarial attacks target the end-to-end tasks, i.e. aim at altering the segmentation mask predicted for a specific image-prompt pair. However, this requires running an individual attack for each new prompt for the same image. We propose instead to generate prompt-agnostic adversarial attacks by maximizing the ℓ_2-distance, in the latent space, between the embedding of the original and perturbed images. Since the encoding process only depends on the image, distorted image representations will cause perturbations in the segmentation masks for a variety of prompts. We show that even imperceptible ℓ_∞-bounded perturbations of radius $\epsilon = 1/255$ are often sufficient to drastically modify the masks predicted with point, box and text prompts by recently proposed foundation models for segmentation. Moreover, we explore the possibility of creating universal, i.e. non image-specific, attacks which can be readily applied to any input without further computational cost.**

Index Terms—**adversarial robustness, image segmentation, foundation models**

I. Introduction

Foundation models, that is large pre-trained models which can be easily adapted to downstream tasks, have been recently proposed in a variety of domains [24, 29]. In the context of segmentation, general purpose models like Segment Anything (SAM) [15] and Segment Everything Everywhere All at Once (SEEM) [35] are able to segment objects given visual, text or audio prompts, or even provide a (semantic) segmentation map for an entire image without any specific prompts. These models exhibit strong generalization performance to unseen datasets, which makes them well-suited for being readily deployed in a multitude of practical applications. This emphasizes the relevance of understanding their robustness to adversarial attacks, as their potential vulnerabilities might threaten the safety of systems integrating these models. However, this aspect has been so far only partially studied: in fact, previous works [33, 11, 23] analyze the robustness of SAM when an end-to-end task is attacked, which means that one generates an adversarial perturbation to alter the predicted mask for a given image-prompt pair. While this yield strong task-specific attacks, it requires running the attack for each new prompt (and image) independently, which might be expensive, and the attack circumvented by using different prompts.

In this work, we exploit the fact that foundation segmentation models process an input image via an encoder which does not take into account the prompts defining the segmentation tasks. In this way, the image embedding is computed once, as it is typically computationally intense, and then re-used to generate the segmentation masks corresponding to multiple prompts. We propose to generate prompt-agnostic adversarial attacks by modifying an input image to distort the embedding produced by the image encoder, i.e. maximize the ℓ_2-distance between the representations, in latent space, of clean and adversarial image (see Fig. 1). The intuition behind this approach is that, if the embedding of the adversarial image is sufficiently altered compared to the original one, it will not be useful to obtain precise segmentation masks regardless of the type and instance of prompts received. Moreover, this allows us to compute a single attack which effectively degrades the performance of the model independently of the end task.

With this approach we show that small ℓ_∞-norm bounded perturbations, which can be optimized with standard techniques for adversarial attacks like projected gradient descent [19], are able to significantly degrade the performance of SAM and SEEM on a variety of tasks, while introducing imperceptible changes to the original image (Sec. IV-A and Sec. IV-E). Moreover, we adapt our algorithm to generate universal adversarial perturbations (Sec. IV-C), that is a single perturbation is generated leveraging a limited number of training images and then can be applied to any new unseen image without additional cost. Finally, in Sec. IV-D we provide a version of our attacks to counter the use of a more sophisticated and expensive configuration of the mask generator of SAM.

II. Related Work

Robustness of SAM. [33] first studied the adversarial robustness of SAM: they generate ℓ_∞-bounded perturbations with either FGSM [7] or PGD [19] (with 10 steps) to remove or manipulate the predicted mask for a given pair of image and prompt. Similarly, [11] use 10 steps of various algorithms (BIM [16], PGD, SegPGD [8]) for ℓ_∞-bounded attacks to alter the original predicted mask for a given prompt by maximizing the training loss of SAM, i.e. a combination of Focal [17] and Dice [28] loss. [23] further conducts an evaluation similar to [33], increasing the number of iterations in PGD to 20. These works focus on the task of removing or changing the predicted mask for a specific point prompt,

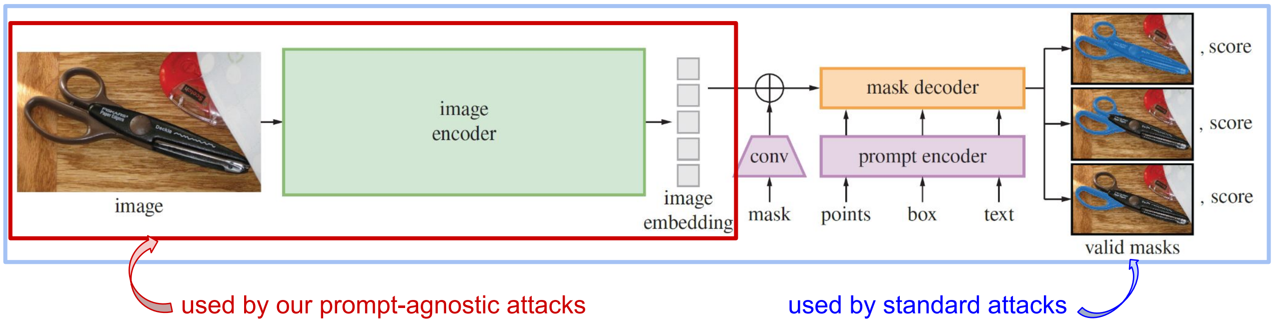

Fig. 1: **Visualization of the architecture of SAM and our proposed method.** We show the structure of the architecture of SAM (image taken from [15]) and different attack approaches. While standard attacks consider all components of the segmentation model and the predicted masks to generate their perturbations, we propose to only distort the embedding provided by the image encoder. In this way, the resulting perturbations are not specific to the prompt used for generating the attacks.

and show the vulnerability of SAM to attacks targeting it. In [33] *cross-prompt* tasks are also introduced, where an attack generated with a source prompt is evaluated with a different target prompt, and show that increasing the number of source prompts (up to 400) to optimize the adversarial perturbations might improve its generalization to unseen prompts. However, only single point prompts are used as targets. Finally, [11, 23] study the performance of SAM when the input images are changed by various common (non-adversarial) corruptions e.g. those in ImageNet-C [10] or style transfer.

Adversarial attacks for semantic segmentation. Several works have focused on developing adversarial attacks against semantic segmentation models: most research [20, 2, 21, 8, 1, 4, 9] has considered the popular ℓ_∞-bounded threat model, while some attention has been received by patch attacks [22] and unconstrained perturbations [30, 27, 13, 25]. Moreover, some of these works have introduced universal [20, 13] and data-free [21] attacks against semantic segmentation models.

Related to our approach, [21, 9] introduce techniques which involve attacking some internal representation of the target models. In particular, in recent work, [9] attack a semantic segmentation model maximizing the cosine similarity between the output of the model backbone for the adversarial image and a random vector (or a target vector for targeted attacks). Additionally they combine such objective with a loss on the predictions of the entire model (as standard for attacks in semantic segmentation), and solve the resulting optimization problem with Adam [14] together with a projection operation. Unlike [9], we do not rely on a target vector in the embedding space or additional losses including the predicted masks.

In general, distorting internal representations for adversarial attacks has been explored in the literature, even in the context of attacks on image classifiers e.g. by [26, 12]. However, compared to [9, 21], our work is to our knowledge the first one to focus on prompt-agnostic (universal) attacks: by targeting the image encoder of foundation models, we obtain adversarial perturbations which are independent of a specific downstream segmentation task and its associated performance metric.

III. Prompt-Agnostic Adversarial Attacks

The architecture of both SAM, MobileSAM [32] (a lightweight version of SAM) and SEEM consists of several components: first the input images are processed by a vision encoder, while the visual prompts (points, boxes, etc.) and text prompts, which characterize the object to segment, are processed separately by different encoder networks. The features extracted by the these various encoders are then used by a decoder to generate the predicted segmentation masks (and possibly classes, when semantic segmentation is supported as in SEEM). Fig. 1 shows an illustration of the structure and functioning of SAM.

We can formalize such models as a function f which, given an image $\boldsymbol{x}$ and set of prompts P, returns a mask $\boldsymbol{m} = f(\boldsymbol{x}, P)$ with the predicted segmentation. Previous works [33, 11, 23] have generated adversarial perturbations against SAM by solving

$$\begin{aligned} \max_{\boldsymbol{\delta}\in\mathbb{R}^{w\times h\times c}} \; & \mathcal{L}\big(f(\boldsymbol{x}+\boldsymbol{\delta}, P), f(\boldsymbol{x}, P)\big) \\ \text{s.th.} \quad & \|\boldsymbol{\delta}\|_p \leq \epsilon, \quad \boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^{w\times h\times c}, \end{aligned} \tag{1}$$

for some loss function $\mathcal{L}$, where ϵ is the largest ℓ_p-norm allowed for the perturbation. This attack aims at changing the prediction of f for specific image-prompt pairs $(\boldsymbol{x}, P)$, but need not affect the outcome for a different prompt P as has been shown in [33].

However, as part of the model f, a vision backbone $\phi : \mathbb{R}^{w\times h\times c} \longrightarrow \mathbb{R}^n$ is used as image encoder to extract a feature vector $\phi(\boldsymbol{x})$ for the input image $\boldsymbol{x}$ which is at this stage independent of the prompt P, see Fig. 1. Thus we propose to use instead the following attack objective:

$$\begin{aligned} \max_{\boldsymbol{\delta}\in\mathbb{R}^{w\times h\times c}} \; & \|\phi(\boldsymbol{x}+\boldsymbol{\delta}) - \phi(\boldsymbol{x})\|_2^2 \\ \text{s.th.} \quad & \|\boldsymbol{\delta}\|_p \leq \epsilon, \quad \boldsymbol{x}+\boldsymbol{\delta} \in [0,1]^{w\times h\times c}, \end{aligned} \tag{2}$$

where, compared to Eq. (1), we remove the dependence on P. Intuitively, maximally perturbing the features extracted by the image encoder should mislead the downstream segmentation output regardless of the prompt provided by the user.

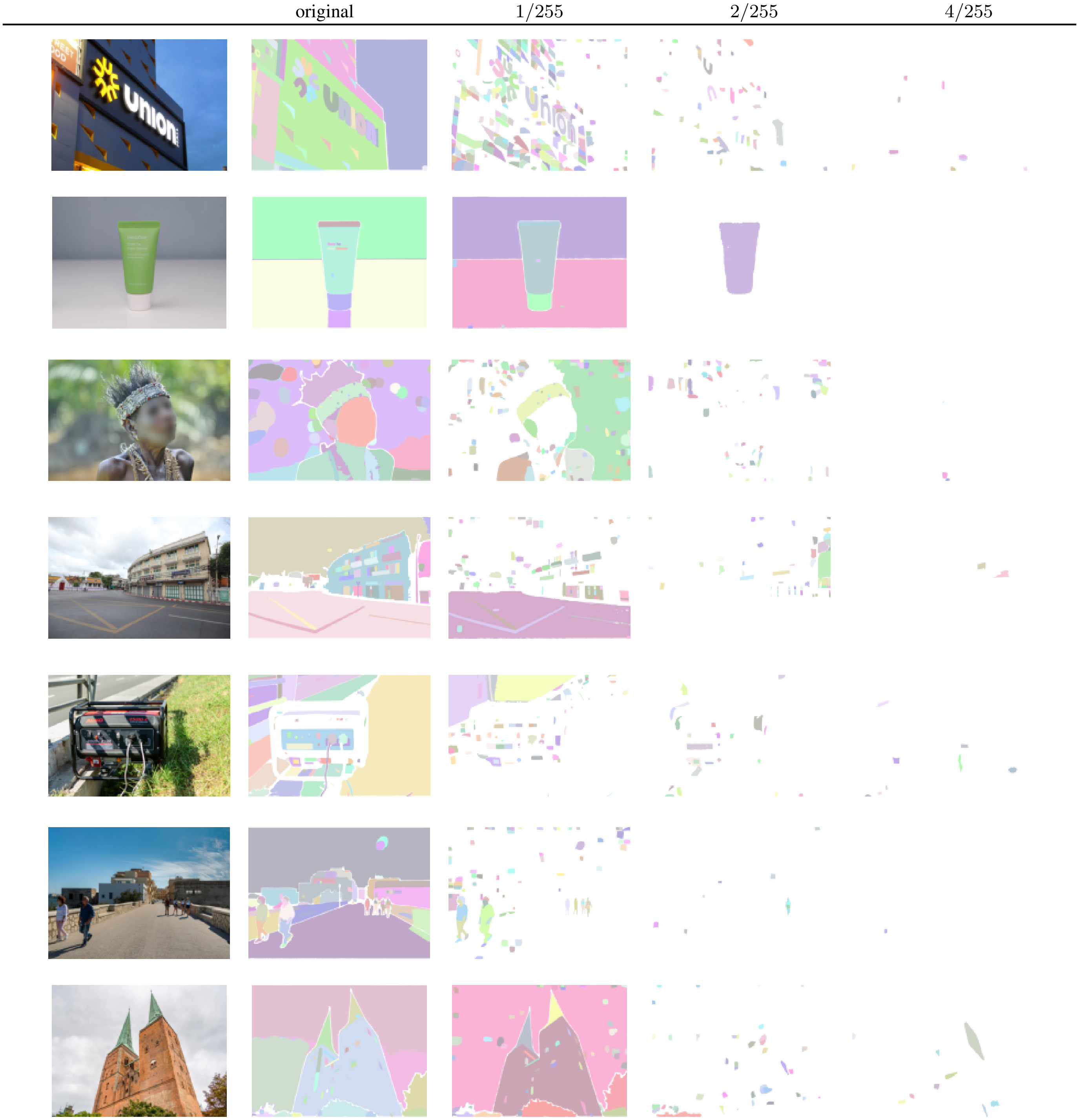

Fig. 2: **Segment Everything mode of SAM.** We show the effect of adversarial perturbations with different ℓ_∞ bounds $\epsilon = \{1/255, 2/255, 4/255\}$ on the results of Segment Everything mode of SAM. Even small perturbations are sufficient to significantly modify the predicted segmentation masks.

The optimization problems in Eq. (1) and Eq. (2) can be solved by Projected Gradient Descent (PGD) [19] and its variants [16, 3] similar to ℓ_p-bounded attacks on image classification. We use APGD [3] with an automatically adaptive schedule and step size as shown to outperform standard PGD. However, APGD can only be used if the objective is fixed as it tracks the objective, e.g. for the step size selection. Thus for varying objectives as in universal attacks we use PGD.

IV. Experiments

A. White-box attacks on SAM

In the following we test our attack in the white-box scenario (the attacker has complete access to the target model) using the publicly available[1] checkpoint for the ViT-H [6] backbone, and the default parameters for the mask generators. If not stated otherwise we use 100 steps of APGD [3] to generate ℓ_∞-bounded adversarial perturbations by solving the problem in Eq. (2). Note that this is a relatively small computational budget for adversarial attacks, which typically rely on thousands of iterations, possibly distributed over several random restarts [3]. We fix such budget since it already provides strong attacks and keeps a relatively small computational cost, given the large architectures used by the target models. As images we use random samples from the a random chunk of the SA-1B dataset [15]. We illustrate the effect of the prompt-agnostic attacks with several qualitative examples, and report a summary of their results in the quantitative evaluation in Table I.

Segment Everything mode. We first consider the mode of SAM in which it tries to segment all existing objects in an image. For this, a grid of point prompts is automatically generated, and their predicted masks are then automatically filtered and deduplicated.[2] In Fig. 2 we show, for each row, first the original input image and its predicted masks (i.e. no perturbations is added), then the predicted segmentation masks when a perturbation of size ϵ is applied, with increasing values $\epsilon \in \{1/255, 2/255, 4/255\}$. These radii are sufficiently small to introduce perturbations not visible in the large majority of cases, and in the range of what commonly used in works on adversarial robustness in image classification [5] and segmentation [4]. One can observe that already the smallest perturbations ($\epsilon = 1/255$) are able to significantly alter the model predictions in most cases. Note that the white areas indicate that no mask is predicted for those pixels. Some of the largest areas with uniform colors are still correctly segmented, but disappear with larger ϵ values.

Point prompts. Next we test the effect of the adversarial perturbations on SAM when using point prompts to segment a desired object. We explore using both positive and negative prompts, which indicate whether a point belongs or not to the target object. Fig. 3 and Fig. 4 show the predicted masks with p positive and n negative prompts, for $(p, n) \in \{(1, 0), (3, 0), (2, 2), (4, 4)\}$, for several images, objects and seeds. We select masks from the annotations of SA-1B and uniformly sample points inside or outside it as positive and negative prompts respectively. Since SAM outputs three proposals as segmentation masks (Fig. 1) with associated quality scores, we show the one with highest score. We observe that, while across prompts the original image is precisely segmented, the adversarially perturbed inputs, even with small ℓ_∞-bounds, significantly alter the predictions, and in most cases the mask fails to identify the target object. Increasing the number of prompts does not lead, in general, to recovering the correct segmentation, especially at the largest radius $\epsilon = 8/255$. We recall that the same perturbed image is used with all prompts and masks, which have not been seen by the attack when optimizing the perturbations.

TABLE I: **Prompt-agnostic white-box attacks on SAM.** We report the average IoU, over all masks of 100 images from SA-1B, of predicted and ground truth masks. We show the results when using clean and perturbed (with attack radii $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$) images and various types of prompts. We compute the results for either all masks or masks with a minimum area. Our attacks, which have been generated without using any prompt, can reduce the mIoU when using a variety of prompt modalities.

prompt type	clean	1/255	2/255	4/255	8/255
all masks					
single point	74.4	40.7	16.4	5.6	3.2
multiple points	76.8	46.5	21.2	8.2	4.8
box	92.7	76.3	56.1	42.6	36.3
single point + box	92.2	76.8	59.2	46.9	40.9
average	84.0	60.1	38.2	25.8	21.3
masks with area $\geq$ 3000 pixels					
single point	74.9	38.4	16.0	6.6	3.9
multiple points	80.3	48.5	23.0	10.4	6.7
box	94.3	73.2	48.4	34.1	28.6
single point + box	94.2	75.6	55.5	42.7	37.3
average	85.9	58.9	35.7	23.5	19.1
masks with area $\geq$ 10000 pixels					
single point	73.3	29.6	11.5	6.0	4.4
multiple points	81.9	44.5	21.3	12.0	9.3
box	95.6	68.1	41.9	29.7	25.8
single point + box	95.7	72.0	51.1	40.0	35.5
average	86.6	53.5	31.4	21.9	18.8

Box prompts. Similarly to point prompts, we select several box prompts from the dataset annotations and show the effect of using perturbed inputs (varying radii $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$) in combination with them in

[1] We use the implementation and models provided at https://github.com/facebookresearch/segment-anything.

[2] See e.g. https://github.com/facebookresearch/segment-anything/blob/main/notebooks/automatic_mask_generator_example.ipynb for more details.

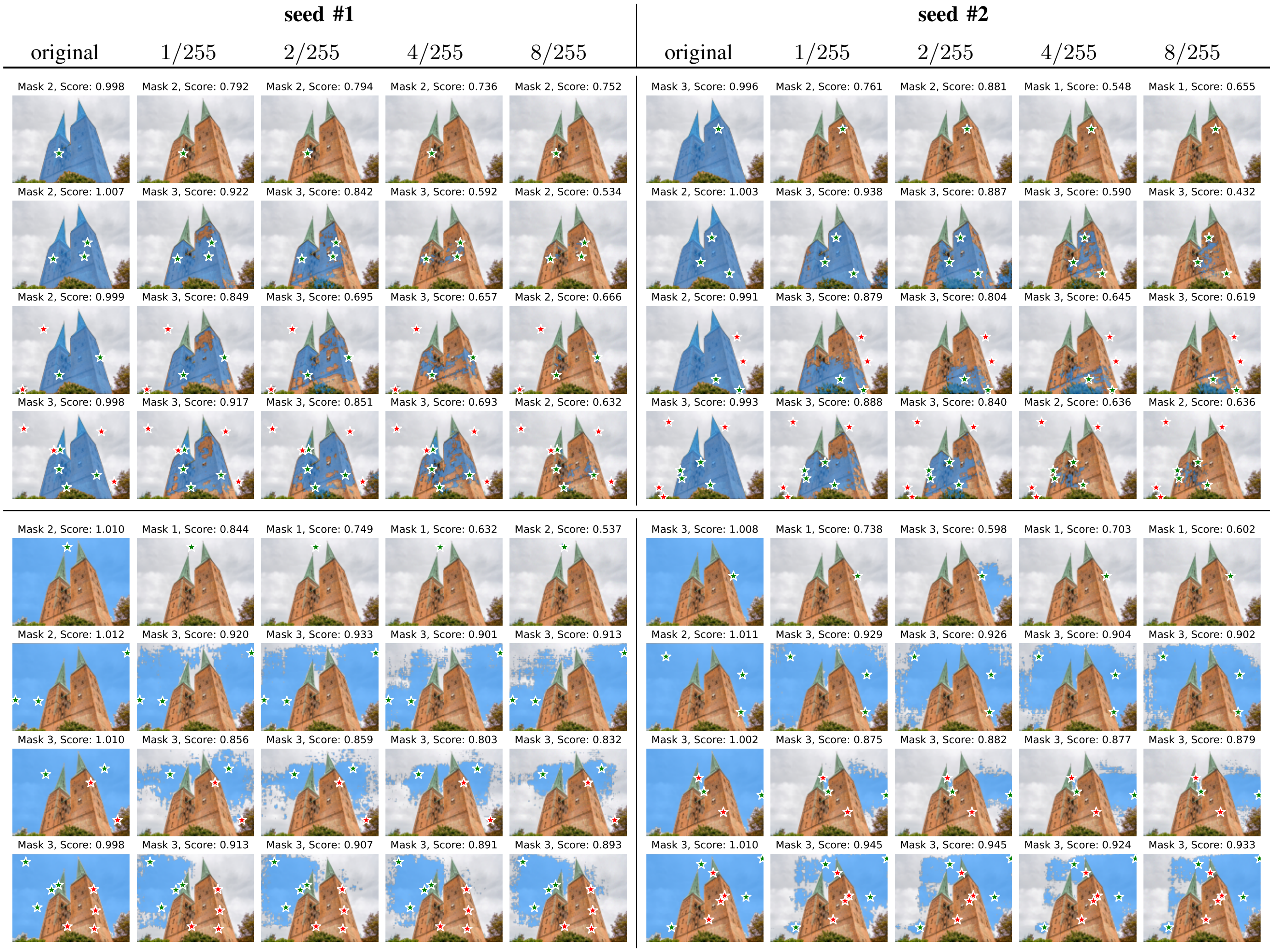

Fig. 3: **Multiple point prompts on SAM.** For each mask (different blocks) we vary the number of random positive (green stars) and negative (red stars) point prompts (different rows), and repeat with two seeds (left and right sides of the panel), i.e. sampling different prompts. Above each image we report the quality score predicted by the model for the mask (the mask with highest score is selected among the three proposals). For adversarially perturbed images (radii $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$) the quality of the masks is significantly degraded, even when increasing the number of point prompts.

Fig. 5. The adversarial images often lead to masks which are either (almost) empty, inside the box but complementary to the correct ones or of poor quality, sometimes leaving the boundaries of the box. Only when the box includes the entire image, which for unperturbed images results in the background being segmented, the attacks cannot completely erase the correct masks.

B. *Quantitative evaluation of white-box attacks on SAM*

In the following we report a quantitative evaluation of the effectiveness of our white-box (image-specific) attacks on SAM, and an analysis of how the size of the masks influences the results. While we here focus on the SA-1B dataset, we provide further results for ADE20K in App. A-A.

We randomly select 100 images from the SA-1B dataset, whose annotations contain for each image a set of masks with a single point and a box prompt. We create, for each masks, further prompts by using the single point and box prompts simultaneously (single point + box setting), and sampling multiple point prompts (we use 2 random positive and 2 random negative prompts as described in Sec. IV-A). Thus we test how four different types of prompts affect the success of our prompt-agnostic attacks. As performance metric, we compute the Intersection over Union (IoU) between the prediction of SAM and the ground truth masks, when using either clean or adversarially perturbed images. We then average IoU (mIoU) over all masks (9320 in total for the 100 image).

In Table I we report the mIoU for each setup (type of prompt) when using clean images or perturbed ones, with attack radii $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$. We recall that a single perturbation is generated for every clean image at each radius. We see that for single and multiple point prompts even

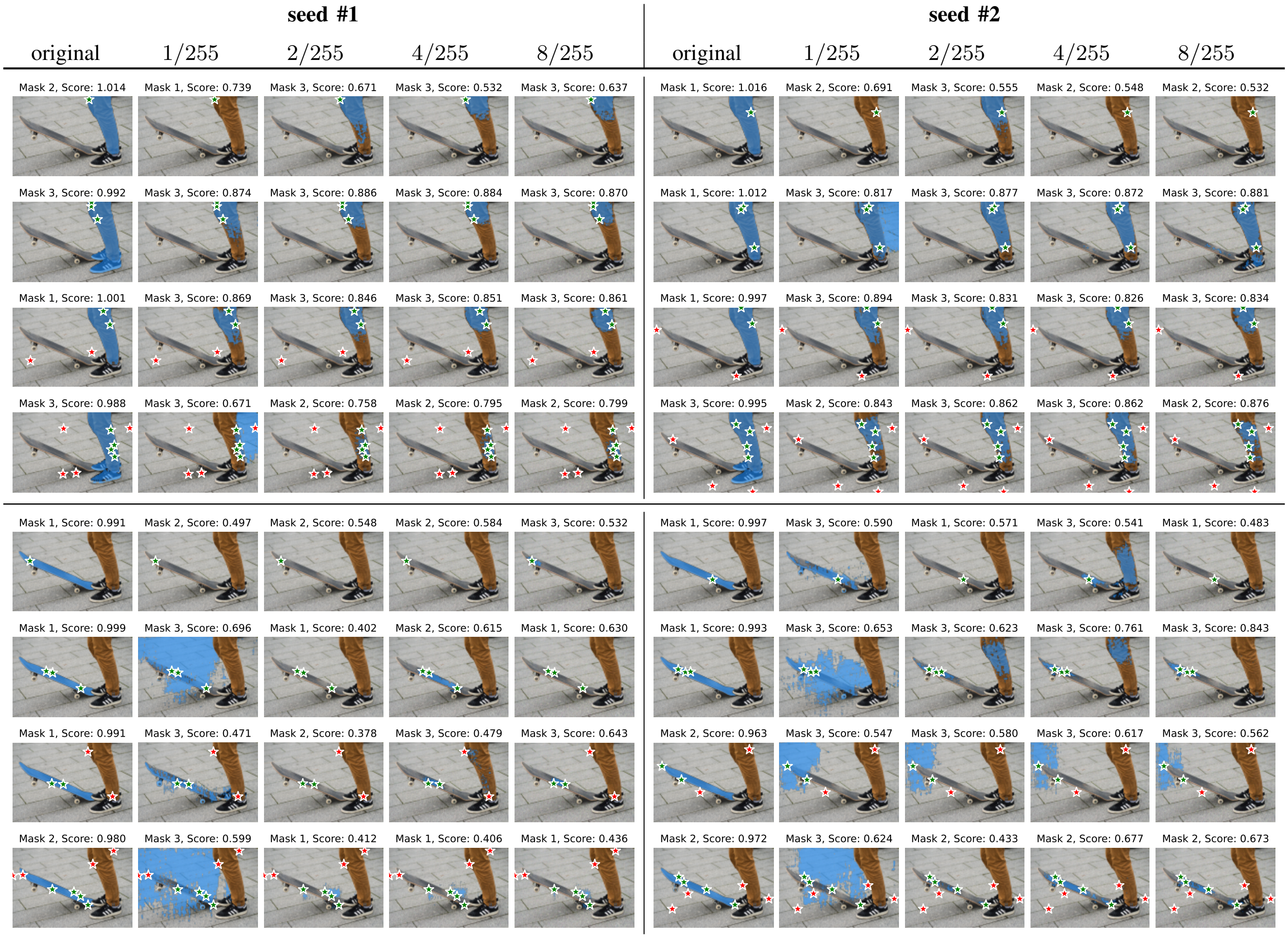

Fig. 4: **Multiple point prompts on SAM.** We repeat the experiment in the setting reported in Fig. 3 for a different input image from SA-1B, with similar observations about the effectiveness of the attacks.

small perturbations can significantly degrade the quality of the segmentation masks, e.g. mIoU at $\epsilon = 4/255$ is below 10% in both cases. Using box prompts improve clean performance as well as robustness of SAM, since those provide richer information than point prompts. However, even in this case our attacks can reduce mIoU from above 92% to 36%-41%.

Since mIoU might be largely influenced by small masks, where correctly segmenting only a few pixels is enough to achieve high values, we explore how the size of the mask changes the success of the attacks. Table I additionally reports mIoU when considering only masks with area of at least either 3000 or 10000 pixels (note that these represent a small fraction of the images). One can observe that the robust mIoU for box and single point plus box prompts is significantly reduced compared to when all masks are used, with drops up to 14.2% and 8.1% respectively at $\epsilon = 2/255$.

C. Universal attacks on SAM

We have so far tested image-specific perturbations: we now aim at finding a *single* perturbations which can be applied to any input image and prevent generating precise segmentation masks for it. This corresponds to modifying Eq. (2) to

$$\max_{\boldsymbol{\delta}\in\mathbb{R}^{w\times h\times c}} \sum_{i=1}^{n} \left\|\phi(\boldsymbol{x}_i+\boldsymbol{\delta}) - \phi(\boldsymbol{x}_i)\right\|_2^2 \qquad (3)$$
$$\text{s.th.} \quad \left\|\boldsymbol{\delta}\right\|_p \leq \epsilon, \quad \boldsymbol{x}_i + \boldsymbol{\delta} \in [0,1]^{w\times h\times c},$$

where one jointly optimizes the loss for n training images, and the same perturbation $\boldsymbol{\delta}$ is added to all of them.

However, the training images, and more importantly the test images, will not in general have the same resolution. Thus we fix the shape of the perturbation to 1024x1024 pixels (the smaller side of each image in SA-1B is 1500 pixels), and interpolate it to match the size of the target image via a function[3] g. This changes the optimization problem in Eq. (3)

[3] In practice we use the function available in `torch` with default “nearest” mode, see details at https://pytorch.org/docs/stable/generated/torch.nn.functional.interpolate.html.

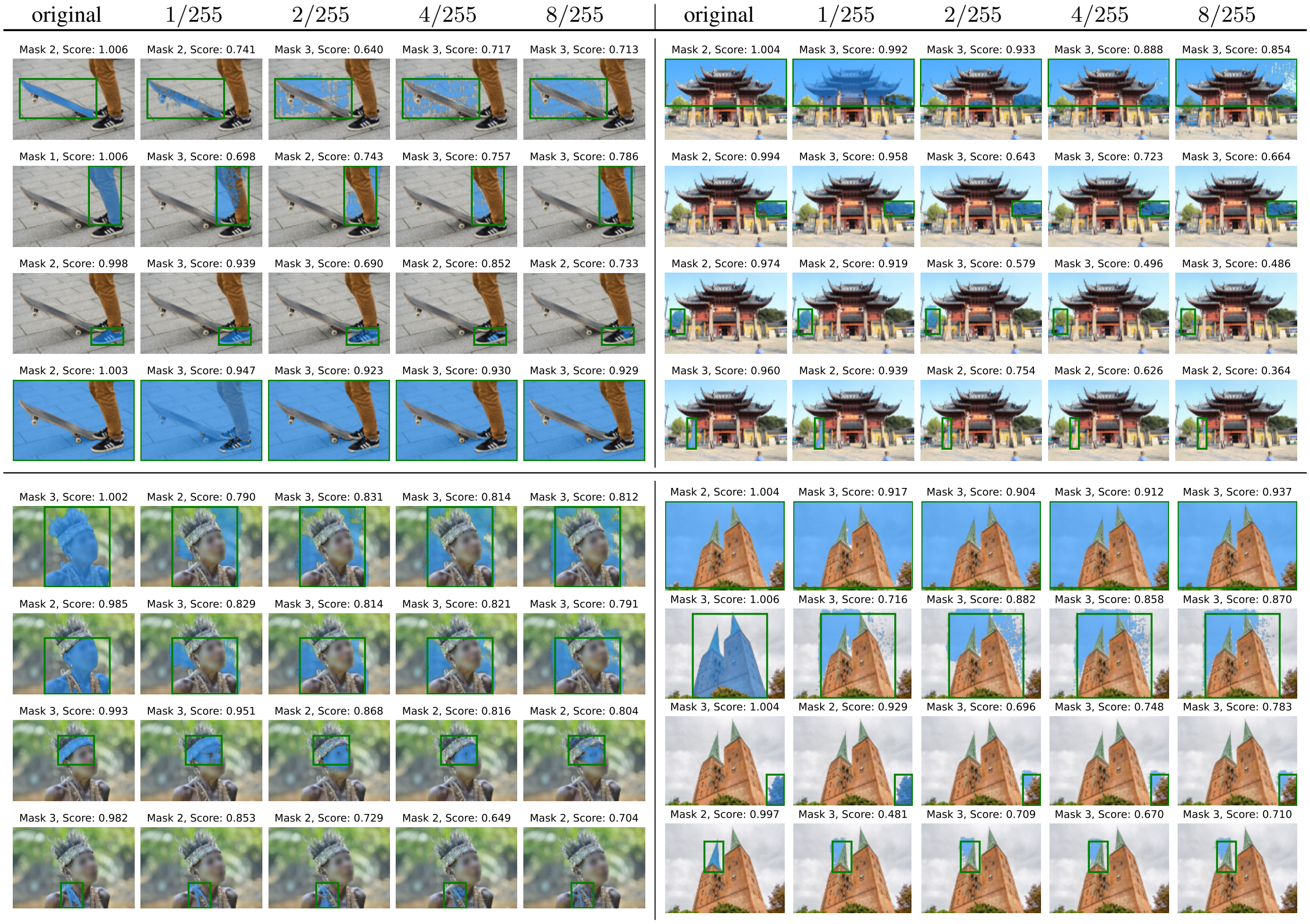

Fig. 5: **Box prompts on SAM.** We show the segmentation masks obtained with several box prompts and its quality score as in Fig. 3. We use either the original image or those perturbed with perturbations of ℓ_∞-norm $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$. Small perturbations effectively degrade the mask quality, especially for small and medium size objects.

to

$$\max_{\boldsymbol{\delta} \in \mathbb{R}^{w \times h \times c}} \sum_{i=1}^{n} \left\| \phi\big(\boldsymbol{x}_i + g(\boldsymbol{\delta})\big) - \phi(\boldsymbol{x}_i) \right\|_2^2 \quad \text{s.th.} \quad \left\|\boldsymbol{\delta}\right\|_p \leq \epsilon, \quad \boldsymbol{x}_i + g(\boldsymbol{\delta}) \in [0,1]^{w_i \times h_i \times c}, \tag{4}$$

where $\boldsymbol{x}_i$ has resolution $w_i \times h_i$. When optimizing the attack, we compute the gradient of the target loss wrt each input image, normalize it wrt its ℓ_2-norm (so that all images have comparable influence on the updates), and finally sum them.

We select a random set of 100 images for generating the attack. To avoid overfitting to such training images, at each iteration we randomly sample a batch of 10 out of the 100 training images and make a gradient step to optimize the sum of their losses. This procedure is also meant to foster generalization to unseen images. We use PGD as optimizer with $\epsilon = 8/255$, step size $1/255$ and 500 iterations, which amounts to 5000 total gradient computations (since this is a more challenging setup we use larger perturbation size and computational budget).

In Fig. 6 we show the effect of the found universal attacks on the results of the Segment Everything mode of SAM (see App. A-A for an evaluation of universal attacks with single point prompts). For both training (seen) and unseen images, adding the perturbation significantly deteriorates the predicted segmentation masks. For the majority of cases, when the adversarial attack is applied, almost no mask is produced, and sometimes many small segmentation masks (not corresponding to any object in the image) appear. We further show the results in the same setup with either a different set of randomly selected training images (Fig. 10) or the smaller perturbation budget $\epsilon = 4/255$ (Fig. 11) in App. A. Although these universal attacks produce slightly worse degradation than the image-specific ones (see Fig. 2, where even smaller radii up to $4/255$ are used), they are realized with a single perturbation which can be applied to any input image. Finally, our goal in this context was to show that prompt-agnostic universal attacks can be achieved even from a small set of images, and we expect that allocating more computational budget for the algorithms, e.g. using more training images, iterations or larger

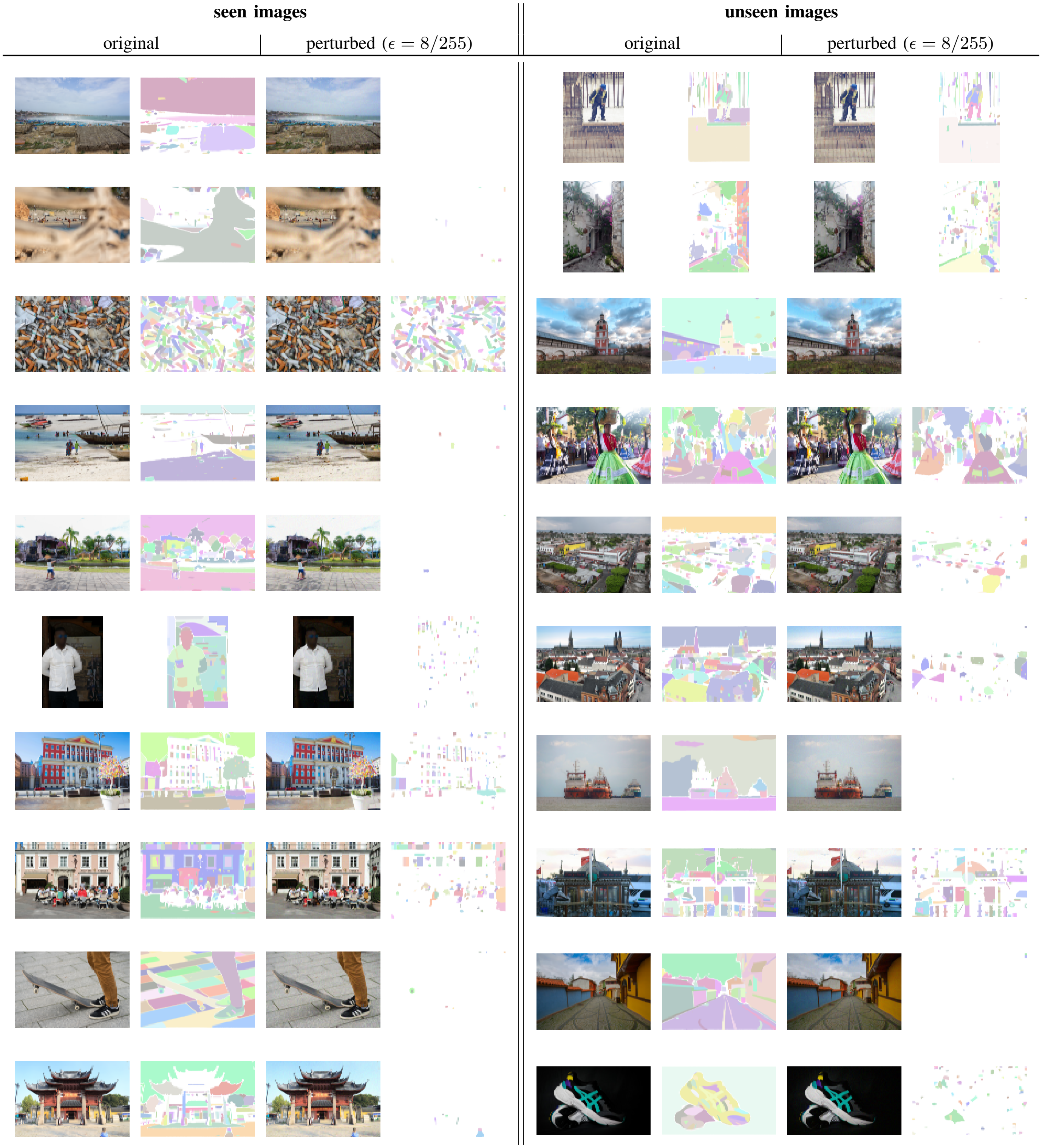
seen images
unseen images
original
perturbed ($\epsilon = 8/255$)
original
perturbed ($\epsilon = 8/255$)

Fig. 6: **Universal attacks on SAM.** In the case of universal attacks, the same perturbation (with ℓ_∞-norm of $8/255$) is applied to every image, and we study its effect on the Segment Everything mode of SAM. On the left we show a subset of the images used for generating the universal attack (first column), together with their predicted segmentation masks (second), the images obtained after adding the universal perturbation (third) and their corresponding predictions (fourth). On the right we follow the same procedure, this time using images not seen by the attack. While some areas are still correctly segmented, for most images the adversarial attacks either prevents any precise segmentation or introduce many small masks.

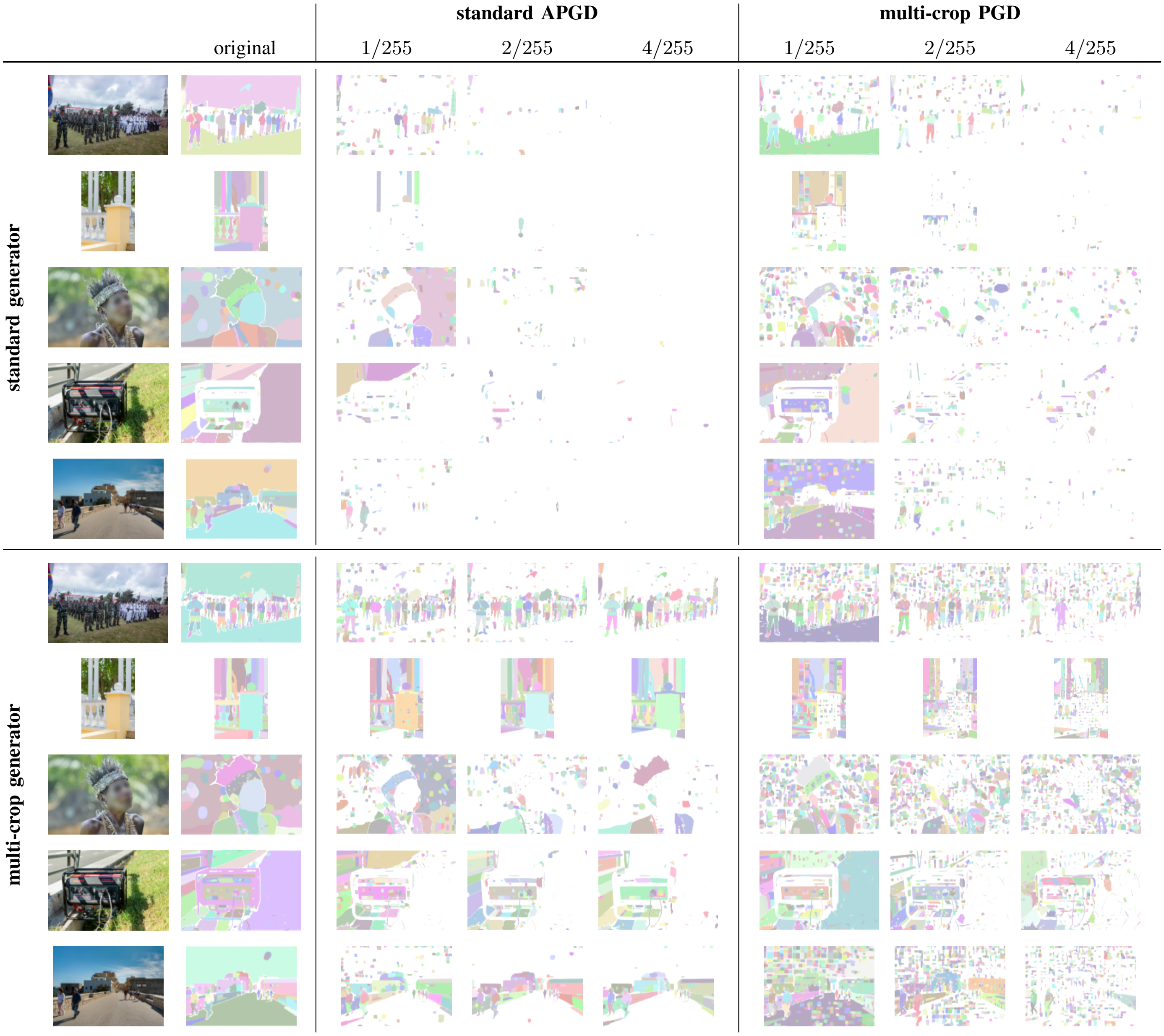

Fig. 7: **Effect of different mask generators and attack algorithms on SAM.** We show the mask predicted by SAM (Segment Everything mode) with standard (top part) and multi-crop (bottom) mask generators, for the clean images (left column) and the adversarially perturbed image produced by either standard APGD (middle) and multi-crop PGD (right). We use each attack with three perturbation bounds $\epsilon \in \{1/255, 2/255, 4/255\}$. The multi-crop PGD is also effective when the multi-crop generator is used, with better results than APGD.

batch size, can improve the generalization of universal attacks.

D. Adaptation to different mask generators in SAM

To control some properties of the predicted masks in the Segment Everything mode, one might tune several parameters of the mask generator in SAM. A more sophisticated mask generator, which computes the segmentation masks for multiple crops of the image and then combines them, is suggested in the original code.[4] While this multi-crop generator is more computationally expensive than the default one, it can produce more refined results, as shown in Fig. 7. Note that this is independent of the image encoder, then different configurations of the generator do not affect the optimization of prompt-agnostic attacks, and can be used to test them.

In Fig. 7 we first show the performance of both standard

[4] The details of the parameters used by the mask generators are reported in https://github.com/facebookresearch/segment-anything/blob/main/notebooks/automatic_mask_generator_example.ipynb.

(the one also used in the previous experiments) and multi-crop generators on clean images (left column) and on the adversarially perturbed inputs given by APGD (middle column) with radii $\epsilon \in \{1/255, 2/255, 4/255\}$ (the same attacks used for Fig. 2). We see that while the attacks are very effective with the standard generator, the multi-crop configuration is, in some cases, still able to segment several objects, especially small ones. We hypothesize that this is due to the fact that in the standard APGD all pixels of the perturbation contribute to the distortion of the features given by the image encoder. However, when only a smaller portion of the image is used, the cropped perturbation might not be as effective.

To counter this, we design a new algorithm, named multi-crop PGD, where, at each iteration, with probability $p_{\text{crop}} = 0.8$ we use a randomly cropped version of the image to compute the objective loss and update the current perturbation. In particular, we first select a random rectangular subset of the current iterate (the perturbed image at the current iteration) whose width and height are uniformly (and independently one from another) sampled between 30% and 90% of the original width and height respectively. Then we make an update step to maximize the feature distortion for this cropped image. Note that this updates the adversarial perturbation only in the area corresponding to the sampled crop. We use 100 iterations of PGD with step size $\epsilon/8$ (we do not use APGD since the objective function is not the same for all iterations).

In the right part of Fig. 7 we show the results of multi-crop PGD: when using the standard mask generator its attacks are still effective, although slightly less than those of standard APGD. At the same time, it is able to deteriorate the predicted masks of the multi-crop generator more significantly than APGD. In particular, with both mask generators, multi-crop PGD leads to a large number of very small masks, with some similarities to what happens for the universal attacks (Sec. IV-C) and unlike the standard attacks.

E. White-box attacks on SEEM

Another recently proposed promptable segmentation model is SEEM [35], whose structure at high level resembles that of SAM with an image encoder that extracts, from an input image, features which can be then combined with various types of visual, text and audio prompts to solve various tasks. In particular, [35] uses different backbones for the image encoder, and in the following we consider the Focal-T and Focal-L [31] which are publicly available.[5] As above, we solve Eq. (2) with 100 steps of APGD [3] wrt ℓ_∞. As for SAM, the attacks only aim at perturbing the features generated by the image encoder (only the visual backbone is used) and do not consider the segmentation head or prompt encoders. We use images from ADE20K and SA-1B: when using the larger backbone Focal-L, we resize the high resolution images from SA-1B to be able to run the attacks with batch size equal 1 in memory of a single GPU. In particular, we resize the images to have smallest edge of size 512 pixels, as suggested in the original code.[6]

Semantic segmentation. Unlike SAM, SEEM provides semantic segmentation maps for the input image, relying by default on the 133 classes (plus a void class) from COCO panoptic segmentation dataset [18]. In Fig. 8 we show the semantic segmentation masks predicted by SEEM with Focal-L backbone for images from ADE20K and SA-1B and the corresponding adversarially perturbed versions given by APGD with ℓ_∞-bounds of $\epsilon \in \{1/255, 2/255, 4/255\}$. We see that the even the small perturbations of size $1/255$ are sufficient to change the predicted classes. While some of the original shapes are still recognizable at the smallest thresholds, these disappear when increasing the budget of the attacks. A similar experiment for the Focal-T backbone is shown in Fig. 12 in App. A, where the attacks achieve similar results to those on the larger image encoder.

Text prompts. A prompt modality which is accepted by SEEM is text, where one can use class names to segment the corresponding objects. We compare the segmentation masks associated by SEEM with Focal-L backbone to different prompts for both clean (from ADE20K) and adversarial inputs in Fig. 9. While the model is able to precisely find the objects from the prompts for clean images, even small perturbations with $\epsilon = 1/255$ are sufficient to notably degrade its performance for all prompts, independently from the size of the target object. We highlight that the same adversarially perturbed image is used with all prompts.

V. CONCLUSION

Discussion. We have shown that it is possible to generate adversarial attacks on foundation models for segmentation in a prompt-agnostic fashion, at low computational cost. In fact, we demonstrated how attacking a single component of their complex architecture may suffice to significantly degrade their performance. This shows the vulnerability of these models, and potentially of the systems integrating them, to imperceptible perturbations. The existence of universal attacks might be particularly interesting, and possibly beneficial. In fact, we envision that such perturbations, especially if they could be found in a black-box setup (we provide an initial study in this direction in App. A-D, analyzing the transferability of universal attacks generated on SAM to SEEM), might be used to prevent the automatic processing of publicly shared images for segmentation tasks (and thus subsequent downstream tasks) by foundation models, resulting in more privacy. Further analyses of these aspects would allow the community to better understand the functioning of foundation models for segmentation and lead to a safer deployment of those in real world applications.

[5] We use the implementation and models provided at https://github.com/UX-Decoder/Segment-Everything-Everywhere-All-At-Once/tree/main.

[6] See https://github.com/UX-Decoder/Segment-Everything-Everywhere-All-At-Once/blob/main/demo_code/tasks/interactive.py. We use bilinear interpolation to preserve the elements of the images in [0, 1].

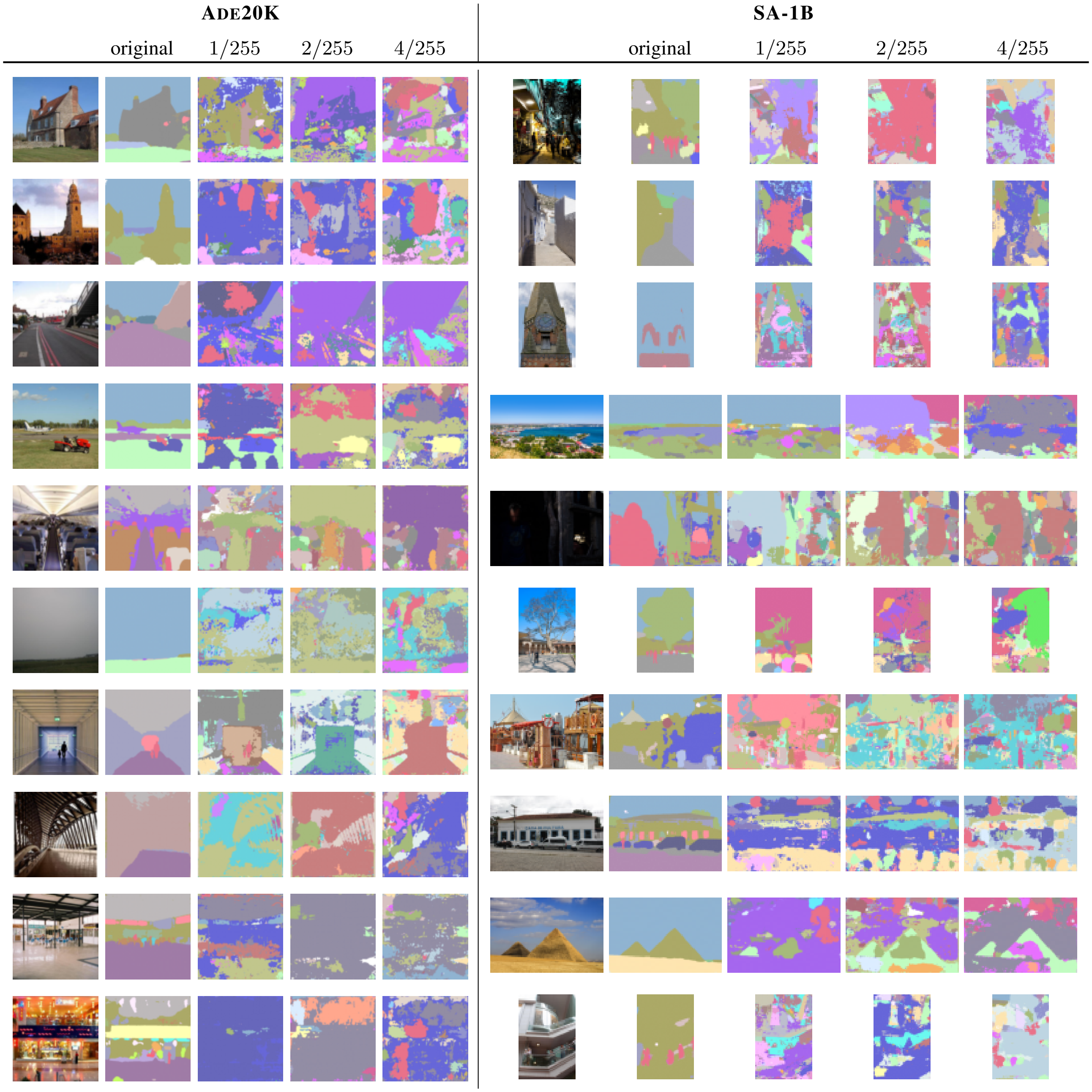

Fig. 8: **Semantic segmentation with SEEM (Focal-L).** For random images from the ADE20K (left) and SA-1B (right) dataset, we show the predicted semantic segmentation maps (each color corresponds to a predicted class) for the original input and the adversarially perturbed, with increasing radii, ones. Small perturbations are sufficient to drastically change the semantic segmentation maps predicted by SEEM.

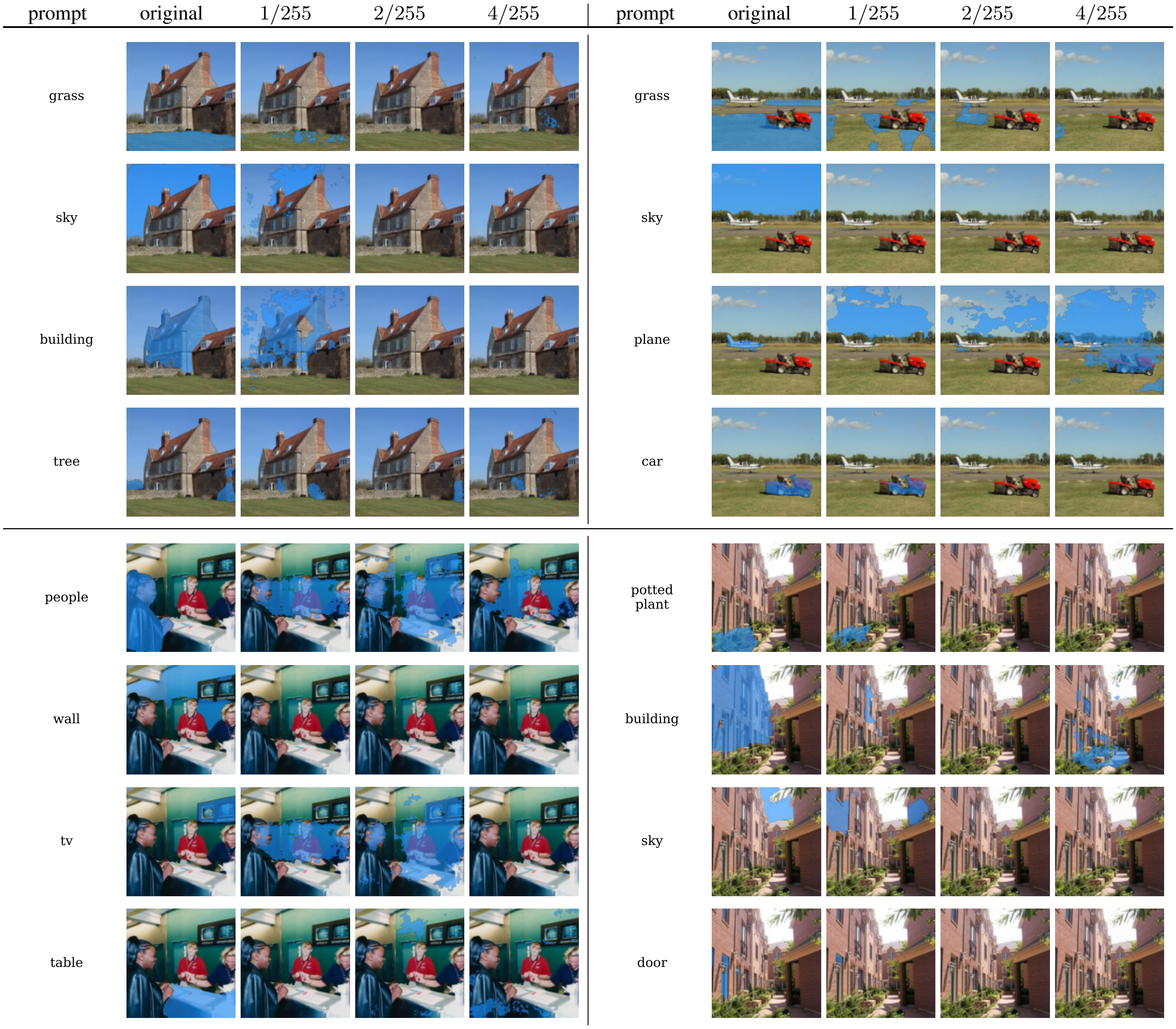

Fig. 9: **Text prompts on SEEM (Focal-L).** We show the predicted segmentation masks given several text prompts for both clean and adversarially perturbed (perturbation size $\epsilon \in \{1/255, 2/255, 4/255\}$) images from ADE20K. While the original masks can precisely identify the objects for each class, small perturbations significantly alter the predicted masks.

Limitations. While our main goal was to demonstrate the feasibility of prompt-agnostic attacks, and their effectiveness in a variety of segmentation tasks, we limited our empirical evaluation to *1)* the most popular models (SAM and SEEM), and a subset of all modalities they allow for, *2)* the white-box scenario, which might not be practical in some applications, *3)* exploring only a few aspects of the attacks. For example, we did not test all possible types of prompts, as well as tuning relevant components of the attacks e.g. the loss used by APGD (targeted attacks might even be possible by designing specific objective functions).

Future work. Since our approach is simple but effective and we use a small computational budget for our experiments, we foresee that it can be an important starting point for future research to develop more sophisticated attacks, possibly in the black-box scenario. Furthermore, one could aim at a single universal perturbation optimized to work across different segmentation models, including those not used when creating the attacks. Finally, it will be an interesting (and challenging) direction to explore how to make foundation models for segmentation adversarially robust without losing their flexibility and generalization properties.

References

[1] S. Agnihotri and M. Keuper. Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks. *arXiv preprint arXiv:2302.02213*, 2023. 2

[2] A. Arnab, O. Miksik, and P. H. Torr. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*, 2018. 2

[3] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 4, 10

[4] F. Croce, N. D. Singh, and M. Hein. Robust semantic segmentation: Strong adversarial attacks and fast training of robust models. *arXiv:2306.12941*, 2023. 2, 4

[5] E. Debenedetti, V. Sehwag, and P. Mittal. A light recipe to train robust vision transformers. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023. 4

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1

[8] J. Gu, H. Zhao, V. Tresp, and P. H. Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *ECCV*, 2022. 1, 2

[9] L. Halmosi and M. Jelasity. On evaluating the adversarial robustness of semantic segmentation models. *arXiv preprint arXiv:2306.14217*, 2023. 2

[10] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 2

[11] Y. Huang, Y. Cao, T. Li, F. Juefei-Xu, D. Lin, I. W. Tsang, Y. Liu, and Q. Guo. On the robustness of segment anything. *arXiv preprint arXiv:2305.16220*, 2023. 1, 2

[12] N. Inkawhich, W. Wen, H. H. Li, and Y. Chen. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, 2019. 2

[13] X. Kang, B. Song, X. Du, and M. Guizani. Adversarial attacks for image segmentation on multiple lightweight models. *IEEE Access*, 8:31359–31370, 2020. 2

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *preprint, arXiv:1412.6980*, 2014. 2

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1, 2, 4, 14

[16] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *ICLR Workshop*, 2017. 1, 4

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017. 1

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 10

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1, 4

[20] J.-H. Metzen, M. Chaithanya Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, 2017. 2

[21] K. R. Mopuri, A. Ganeshan, and R. V. Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2452–2465, 2018. 2

[22] F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 2

[23] Y. Qiao, C. Zhang, T. Kang, D. Kim, S. Tariq, C. Zhang, and C. S. Hong. Robustness of sam: Segment anything under corruptions and beyond. *arXiv preprint arXiv:2306.07713*, 2023. 1, 2

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[25] J. Rony, J.-C. Pesquet, and I. Ben Ayed. Proximal splitting adversarial attacks for semantic segmentation. In *CVPR*, 2023. 2

[26] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet. Adversarial manipulation of deep representations. In *ICLR*, 2016. 2

[27] G. Shen, C. Mao, J. Yang, and B. Ray. Advspade: Realistic unrestricted attacks for semantic segmentation. *arXiv preprint arXiv:1910.02354*, 2019. 2

[28] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017. 1

[29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[30] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017. 2

[31] J. Yang, C. Li, X. Dai, and J. Gao. Focal modulation networks. In *NeurIPS*, 2022. 10

[32] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 2

[33] C. Zhang, C. Zhang, T. Kang, D. Kim, S.-H. Bae, and I. S. Kweon. Attack-sam: Towards evaluating adversarial robustness of segment anything model. *arXiv preprint arXiv:2305.00866*, 2023. 1, 2

[34] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 14

[35] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 1, 10

Appendix A
Additional Results

A. *Evaluation on* ADE20K *masks with single point prompts*

In the following, we provide a quantitative evaluation of the performance degradation of SAM due to our attacks on the ADE20K dataset.

Experimental setup. First, we create a set of ground truth masks and corresponding prompts from the ADE20K dataset [34], since it contains precise annotations and its images are typically evaluated at relatively low resolution (512x512), which allows us to scale the evaluation to a larger number of images and masks. Given an image from the validation set, for each class present in the ground truth semantic segmentation map (except for the background) we select the largest connected component belonging to such class. For each of these masks, after filtering out those with area smaller than 900 pixels, we compute its pixel with largest ℓ_2-distance to its border, following [15]. In this way we collect 100 images and 589 pairs of masks and point prompts.

Then, we can use SAM to predict segmentation masks from each of the point prompts found in the step described above and either the original image or its adversarially perturbed counterparts. Comparing the predicted masks with the ground truth ones derived from ADE20K allows us to measure the performance of both SAM on the unperturbed image and the effectiveness of the attacks. In practice, for each mask-prompt pair we compute Intersection over Union of the predicted (the one with highest quality score among the three provided by SAM) and correct masks, and report its average over all masks and images (mIoU) in Table II.

Results. First, we run our prompt-agnostic image-specific attack with radii $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$, optimized with 100 iterations APGD, on the 100 images used to create the evaluation set. Second, we generate universal perturbations as described in Sec. IV-C, with 100 iterations of PGD (instead of the 500 iterations used in Sec. IV-C to reduce computational cost), for the same radii as image-specific attacks. In this case, as training images, we select 100 samples from ADE20K which do not overlap with those used for evaluation (which means the universal attacks are tested on unseen images), and the universal perturbation has resolution 512x512 as the images of ADE20K.

In Table II we see that, for image-specific attacks, even perturbations of size $1/255$ are able to significantly reduce mIoU from 59.86% to 16.71%. Increasing the attack budget to $8/255$ further degrades the performance of SAM to 7.33%. Conversely, for universal attacks one needs to use the larger radii for effective attacks: notably, at $\epsilon = 8/255$, the mIou for adversarial images is 11.24%, not far from what attained with standard attacks (we recall that the universal perturbations are computed only once and then applied to unseen images without additional cost).

TABLE II: **Single point prompt evaluation on SAM.** We study the performance of SAM when predicting masks derived from ADE20K with single point prompts (we measure average IoU over images and masks). We report the effect of using either image-specific or universal (trained on images of ADE20K not used for computing mIoU) attacks of various sizes. Both types of adversarially perturbed images lead to significant performance drops.

attack	clean	1/255	2/255	4/255	8/255
image-specific	59.86	16.71	11.63	9.42	7.33
universal		58.98	45.09	20.08	11.24

B. *Universal attacks on SAM*

In Fig. 10 we report the results of universal attacks generated as described in Sec. IV-C but using a different set of randomly sampled training images. We see that even in this case we obtain similar results to those shown in Fig. 6 above, with the universal attacks being able to generalize to unseen images.

Moreover, we repeat the experiment shown in Fig. 6 except for the bound on the ℓ_∞-norm of the universal perturbation, which is reduced to $4/255$ (instead of $8/255$). Fig. 11 shows that even with the smaller budget universal perturbations are able to noticeably reduce the quality of the predicted masks of the Segment Everything mode of SAM, although to a lower degree than with the standard radius $8/255$.

C. *White-box attacks on SEEM*

In Fig. 12 we show the effect of our attacks (with radii $\epsilon \in \{1/255, 2/255, 4/255\}$) on the semantic segmentation masks provided by SEEM with the smaller Focal-T backbone, similarly to what done for the Focal-L backbone in Fig. 8. Note that unlike what done for the larger backbone, the images from SA-1B are in this case used at the original resolution. As observed above, the performance of SEEM in semantic segmentation is significantly reduced on the perturbed images.

D. *Transferring universal attacks from SAM to SEEM*

We further test the transferability of our universal attacks generated on SAM for SA-1B images (Sec. IV-C) to other models, i.e. SEEM with different backbones. In Fig. 13 we show the semantic segmentation predicted by SEEM on the images perturbed with the universal attack of size $\epsilon = 8/255$ presented in Fig. 6. We observe that the predicted maps are stable for most input images, especially for the larger Focal-L backbone, but adding the universal perturbations can anyway lead to some mild quality degradation. Transfer attacks across models are known to work better as the architecture of source and target models become more similar. In this case, SAM and SEEM encoders have quite different features, then it is not surprising to observe the limited success of transfer attacks. We leave to future work designing specific techniques (or adapting some from the rich literature on transfer attacks for image classifiers) to improve these results.

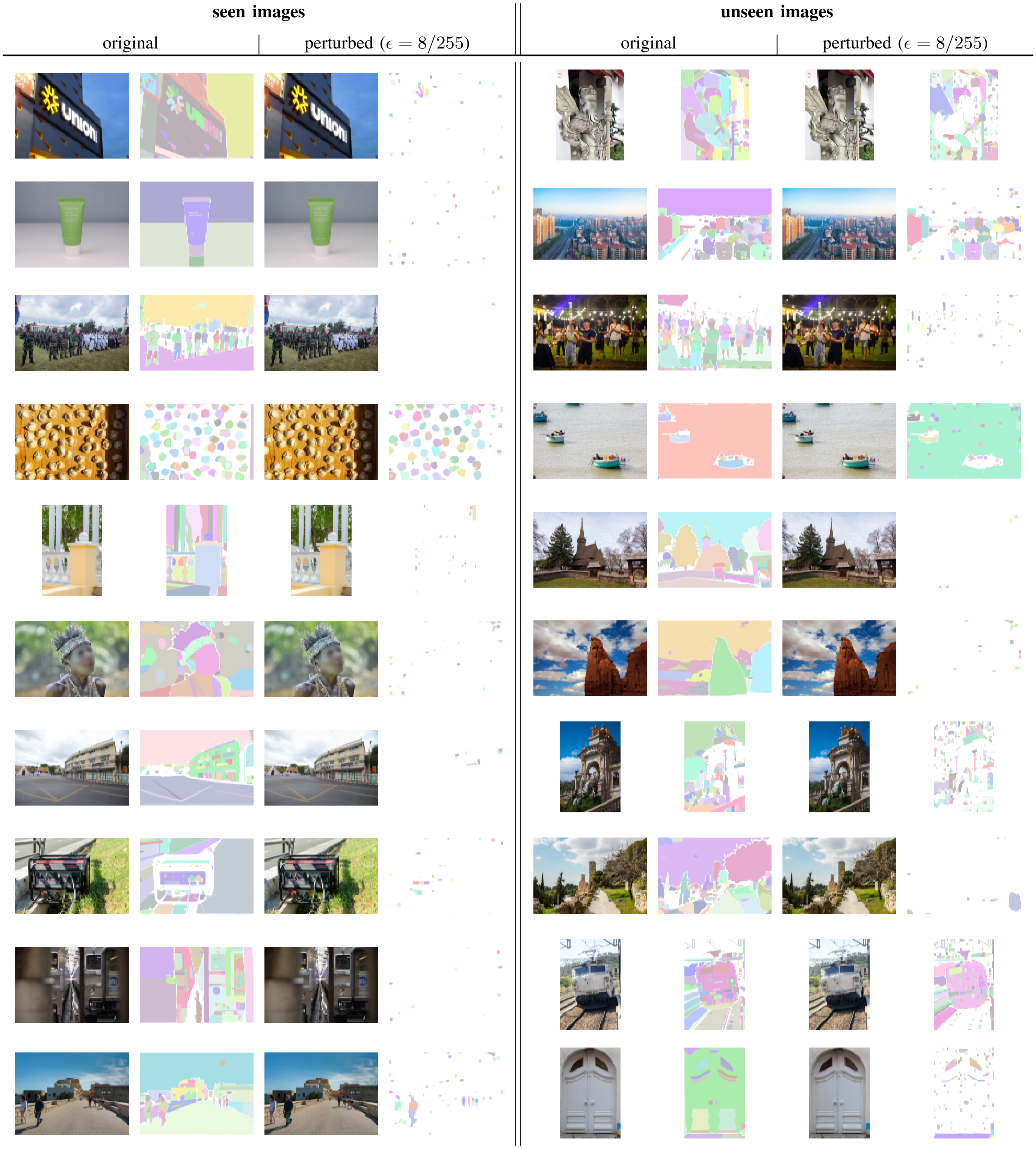
seen images
unseen images
original
perturbed ($\epsilon = 8/255$)
original
perturbed ($\epsilon = 8/255$)

Fig. 10: **Universal attacks on SAM.** We repeat the experiment shown in Fig. 6 for a different set of training images.

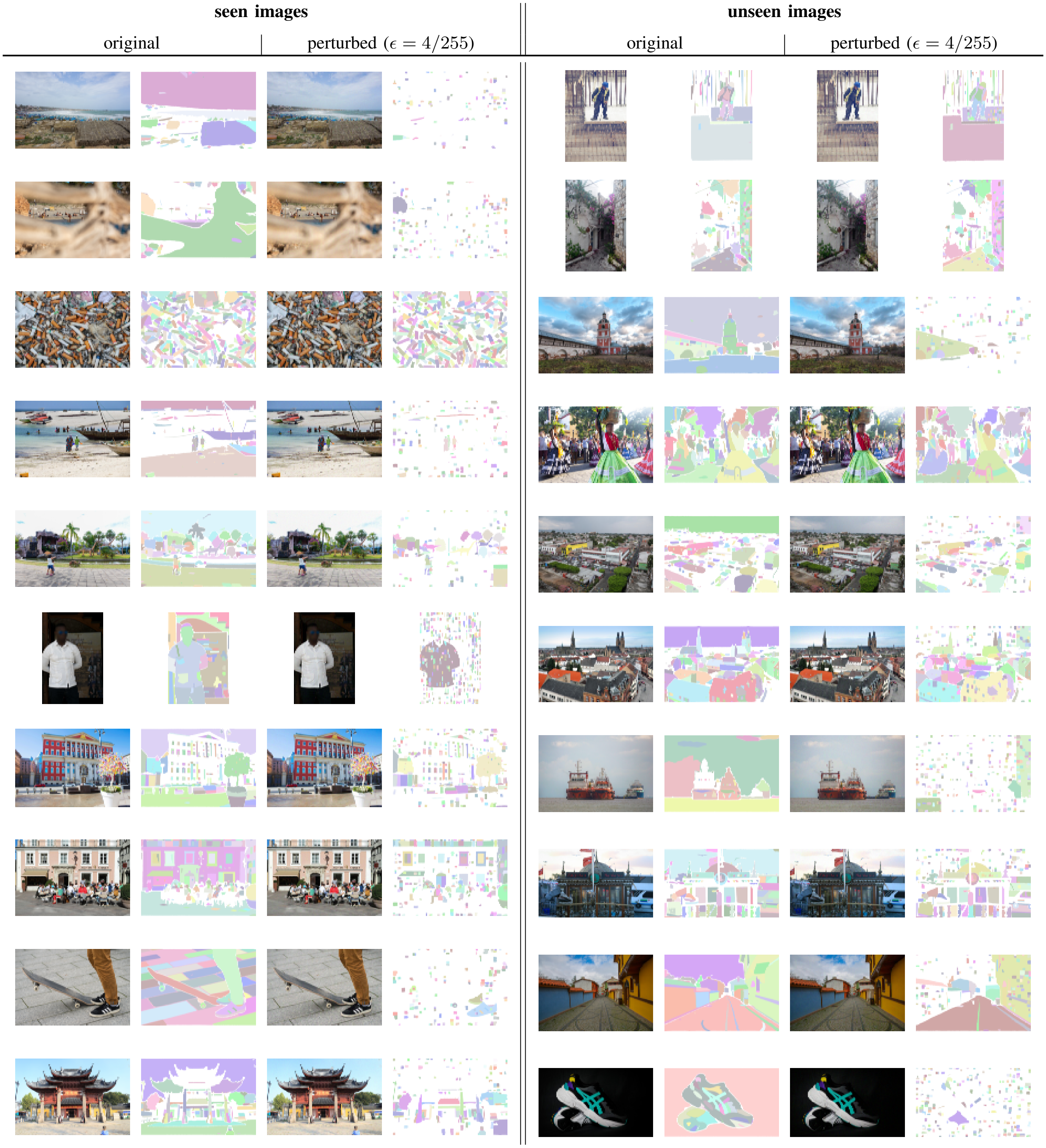
seen images
unseen images
original
perturbed ($\epsilon = 4/255$)
original
perturbed ($\epsilon = 4/255$)

Fig. 11: **Universal attacks on SAM with smaller radius.** We repeat the experiment shown in Fig. 6 with bound of $4/255$ instead of $8/255$ on the ℓ_∞-norm of the universal perturbation.

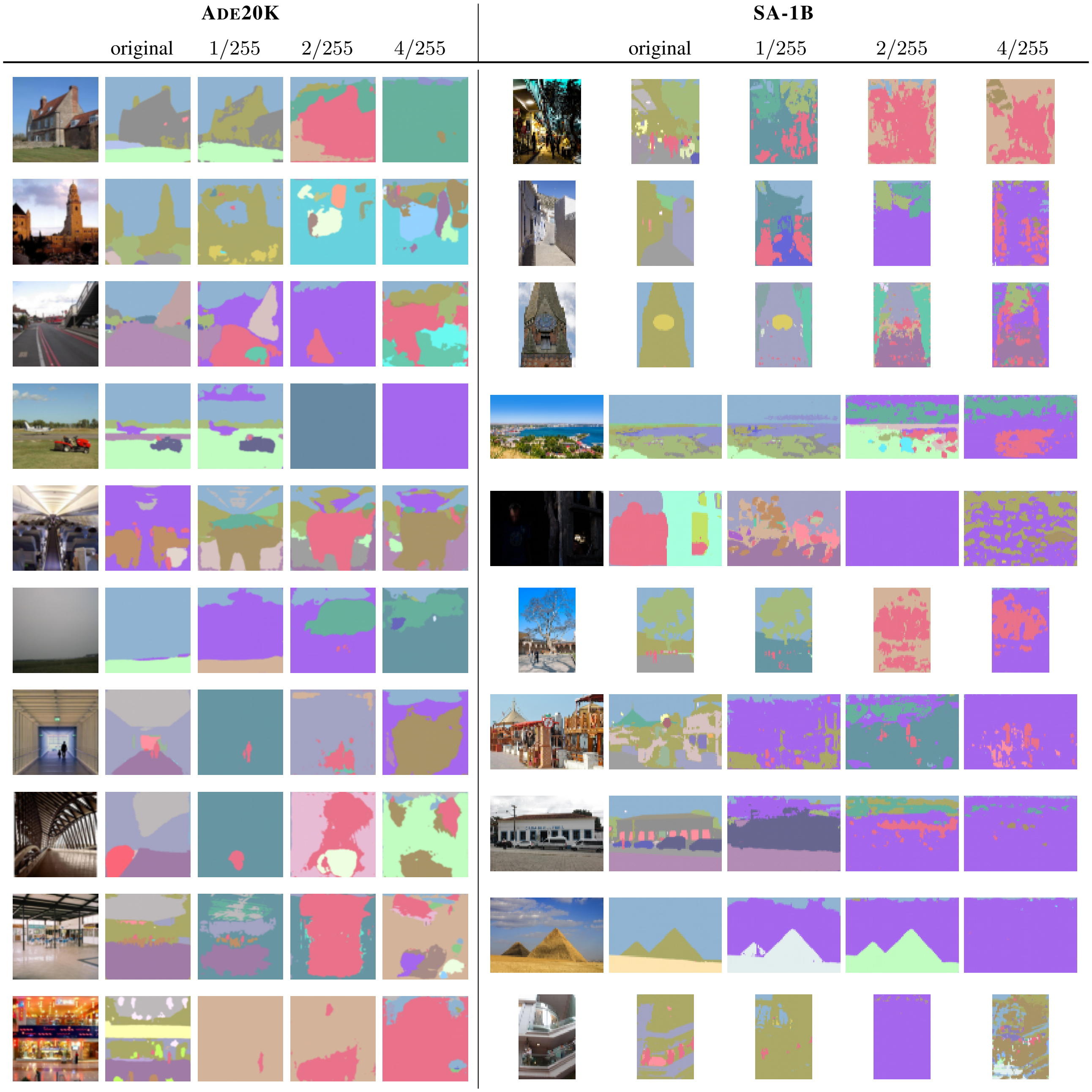

Fig. 12: **Semantic segmentation with SEEM (Focal-T).** For random images from the ADE20K (left) and SA-1B (right) dataset, we show the predicted segmentation masks (each color corresponds to a predicted class) for the original input and with adversarial perturbations of increasing strength. Small perturbations are sufficient to drastically change the semantic segmentation maps predicted by SEEM with Focal-T backbone.

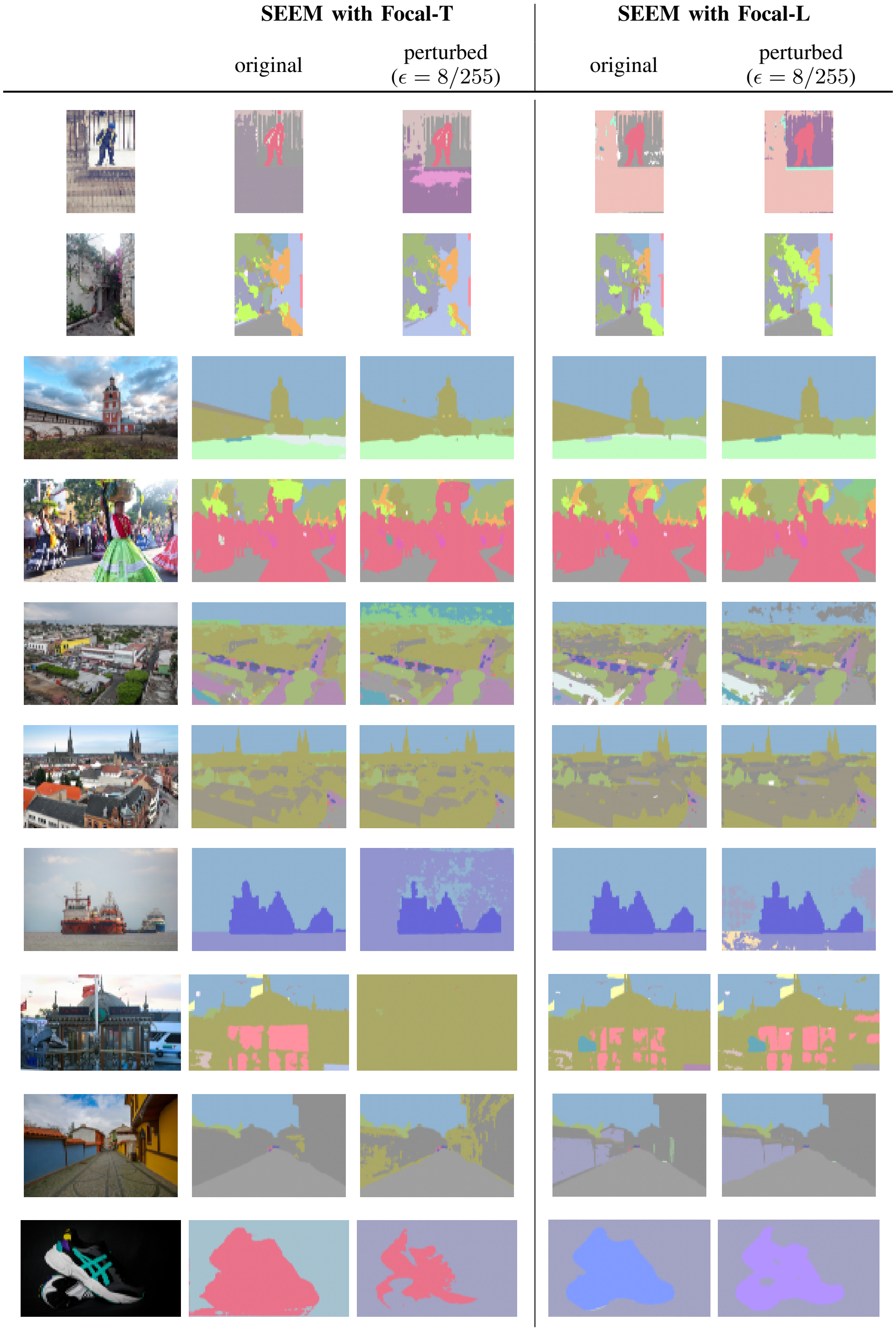
SEEM with Focal-T
SEEM with Focal-L
original
perturbed ($\epsilon = 8/255$)
original
perturbed ($\epsilon = 8/255$)

Fig. 13: **Semantic segmentation by SEEM when transferring universal attacks from SAM.** We test how well the universal attacks generated on SAM with $\epsilon = 8/255$, shown in Fig. 6, transfer to SEEM. For both Focal-T (left) and Focal-L (right) backbones, we show the predicted segmentation masks (each color corresponds to a predicted class) for the clean and perturbed images. While the predictions are mostly stable, especially for the larger Focal-L, the transferred universal perturbation introduce some mild degradation in the output of SEEM.