ASAM: Boosting Segment Anything Model with Adversarial Tuning

Bo Li   Haoke Xiao   Lv Tang[*]
vivo Mobile Communication Co., Ltd
{libra,xiaohaoke,lvtang}@vivo.com

Abstract

In the evolving landscape of computer vision, foundation models have emerged as pivotal tools, exhibiting exceptional adaptability to a myriad of tasks. Among these, the Segment Anything Model (SAM) by Meta AI has distinguished itself in image segmentation. However, SAM, like its counterparts, encounters limitations in specific niche applications, prompting a quest for enhancement strategies that do not compromise its inherent capabilities. This paper introduces ASAM, a novel methodology that amplifies SAM's performance through adversarial tuning. We harness the potential of natural adversarial examples, inspired by their successful implementation in natural language processing. By utilizing a stable diffusion model, we augment a subset (1%) of the SA-1B dataset, generating adversarial instances that are more representative of natural variations rather than conventional imperceptible perturbations. Our approach maintains the photorealism of adversarial examples and ensures alignment with original mask annotations, thereby preserving the integrity of the segmentation task. The fine-tuned ASAM demonstrates significant improvements across a diverse range of segmentation tasks without necessitating additional data or architectural modifications. The results of our extensive evaluations confirm that ASAM establishes new benchmarks in segmentation tasks, thereby contributing to the advancement of foundational models in computer vision. Our project page is in https://asam2024.github.io/.

1. Introduction

The concept of foundation models has been pivotal in advancing the fields of natural language processing (NLP) and, more recently, computer vision. Originating in NLP with influential models such as BERT [13], the GPT series [44], LLaMA [62] and PaLM [10], these models have showcased remarkable zero-shot generalization capabilities to unseen tasks. This success has spurred the development of similar paradigm-shifting models in computer vision. These visual foundation models, such as DINOv2 [45], CLIP [51], BLIP [34], SAM [31] and Stable Diffusion [54], demonstrate remarkable zero-shot capabilities and broad generalization across various tasks.

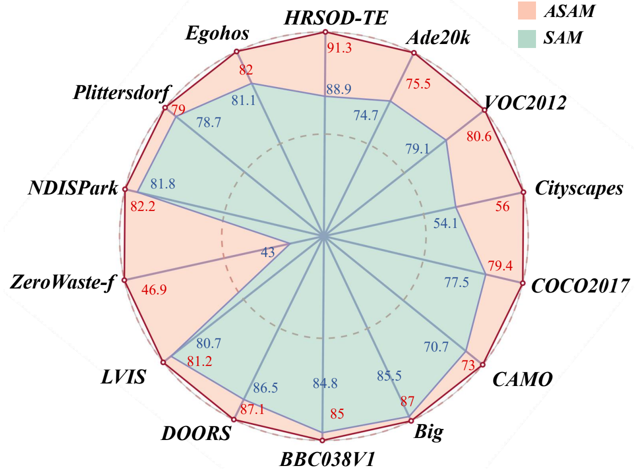

Figure 1. Performance comparison between ASAM and SAM on diverse segmentation datasets across different downstream tasks.

Among them, Segment Anything Model (SAM) stands out as a pioneering visual foundation model specializing in image segmentation. Trained on over 1 billion masks from a massive visual corpus, SAM has revolutionized the field with its ability to segment a diverse range of objects and structures across various scenarios. Despite its impressive performance, SAM, like any foundational model, has areas where it can be further enhanced [25, 26, 61].

An important research direction is identifying SAM's limitations on certain downstream tasks and developing techniques to boost its performance. Many techniques have explored like fine-tuning [3, 37, 47, 82] and adapter modules [6, 48, 69] to specialize the SAM for specific downstream tasks. While fine-tuning unlocks the potential of SAM for a specific task, it compromises the model's inherent generalization capabilities [30]. Alternative approaches preserve SAM's original parameters, adding adaptation layers or post-processing modules [30, 33]. Those methods,

[*] Lv Tang is the corresponding author of this paper

though effective, require additional parameters and annotated training data, limiting its scalability and efficiency.

The above challenges bring us to the core motivation of this work: How can we further boost the generalization ability of SAM as a foundational vision model without relying on substantial extra data, altering its base architecture, or compromising its zero-shot capabilities? So that we can unlock SAM's potential while keeping its broad applicability across vision tasks. Existing solutions, while effective in specific contexts, do not address the fundamental challenge of enhancing SAM's inherent performance across a diverse range of scenarios.

In response to this challenge, we turn to the realm of NLP for inspiration, particularly its pioneering advancements in foundational model research. The unique successes [88] observed in adversarial training (AT) within NLP offer a new vantage point. In contrast to the visual domain, where standard adversarial training often necessitates a compromise between robustness and model performance [66], AT in NLP not only strengthens model robustness but also concurrently bolsters generalization and accuracy [88]. This divergence is believed to be attributed to the closer resemblance of adversarial examples in natural language to real-world textual scenarios, such as common human spelling errors. We surmise that the triumph of adversarial training in NLP is derived from the "realness" and "naturalness" of its generated adversarial examples. This insight leads us to explore the possibility of adapting adversarial training techniques, which have been successful in NLP, to visual foundation models like SAM. This approach aims to apply cross-disciplinary insights innovatively to improve specific tasks in computer vision.

Applying the above concept to SAM, our approach aims to utilize "natural" adversarial examples akin to those found in NLP to elevate visual foundation models. Inspired by the effective tuning methodologies in NLP [46, 64, 65], we propose fine-tuning the SAM using these more "natural" adversarial examples, thereby circumventing the high costs often associated with conventional adversarial training. Traditional methods for generating visual adversarial examples typically adhere to l_p norm constraints, resulting in perturbations that are not entirely natural and exhibit a domain shift from real-world noise. This leads to a disparity between such adversarial examples and the genuinely challenging examples encountered in real-world scenarios [89].

To generate adversarial examples that are both natural and photorealistic for tuning SAM, we are inspired by recent adversarial attack [7] and hypothesize that natural images can be projected onto a low-dimensional manifold via a generative model [54]. This manifold, trained on natural images, ensures the photorealism and richness of content. By mapping an image onto this manifold and then shifting it along an adversarial direction within the manifold, we can produce adversarial examples that are both natural and photorealistic. To maintain the consistency of object shapes with the original mask labels during the back-mapping process, we incorporate an additional mask prompt branch in the generative model. This integration ensures that the adversarial examples are not only realistically aligned but also accurately correspond to their original mask labels. Ultimately, by fine-tuning a select subset of parameters in a large vision model with these naturally realistic and accurately aligned adversarial examples, we achieve significant enhancements in performance. In conclusion, our work makes several key contributions:

- Drawing inspiration from the successes in NLP, we introduce a novel framework, termed adversarial tuning, aimed at enhancing the generalization abilities of visual foundation models like SAM. This approach represents an innovative application of cross-disciplinary insights to address specific challenges in computer vision tasks.
- By projecting natural images onto a low-dimensional manifold using a generative model, we generate adversarial examples that are both natural and photorealistic. We further enhance this approach by integrating a mask prompt branch into the generative model, ensuring that the adversarial examples maintain consistency with the original mask labels in terms of object shape.
- Leveraging our approach, we fine-tune SAM with "natural" adversarial examples, derived from just 1% of the SA-1B dataset, resulting in an enhanced version termed ASAM. To validate ASAM's effectiveness, we conduct extensive quantitative and qualitative analyses. As shown in Fig. 1, ASAM has achieved significant improvements in SAM's performance across a wide range of segmentation datasets and various downstream tasks.

2. Related Works

2.1. Segment Anything Model (SAM)

Meta Research team has released the "Segment Anything" project [31]. This project develops the SAM and an extensive dataset, SA-1B, featuring over 1 billion masks on 11 million licensed and privacy-respecting images. Designed for prompt-based segmentation, SAM is capable of zero-shot adaptation to new image distributions and tasks. As a pioneering visual foundation model, its zero-shot segmentation abilities and prompt-based approach have facilitated rapid application in diverse areas, going beyond image segmentation to tasks like 3D understanding and video processing [5, 22, 60, 71, 72, 75, 76, 86].

While SAM's capability is impressive, its effectiveness in real-world scenarios, such as medical images and other challenging segmentation conditions, has been a topic of investigation. Difficulties arise when segmenting minuscule and slender objects [30], objects with obscure boundaries [25, 28], camouflaged objects [25, 26, 61], and transparent objects [20]. Just like any foundational model, SAM has areas where it can be further enhanced.

To address these challenges, researchers have introduced various methods. For instance, the work [37] proposes a straightforward fine-tuning approach to tailor the SAM for general medical image segmentation. Rigorous experimentation on both 3D and 2D segmentation tasks illustrates that MedSAM surpasses the default SAM. SAM-Adapter [6, 69] leverages domain specific information or visual prompts to enhance the segmentation network through the use of simple yet effective adapters. By combining task-specific knowledge with general knowledge learned by the large model, SAM-Adapter can notably improve the performance of SAM in challenging tasks. While fine-tuning unlocks the potential of SAM for a specific task, it compromises the model's inherent generalization capabilities [30]. Alternative approaches preserve SAM's original parameters, adding adaptation layers or post-processing modules like in SAM-HQ [30] and Semantic-SAM [33]. Those methods, though effective, require additional parameters and annotated training data, limiting its scalability and efficiency. Additionally, instead of direct modifying SAM's parameters, refining the input prompt [85] or output of SAM [16, 67] are also viable strategies.

Our approach diverges from these existing methods, aiming to further enhance SAM's generalization capabilities as a foundational vision model. We seek to achieve this without substantial reliance on extra data, alterations to its architecture, or compromising its zero-shot capabilities.

2.2. Adversarial Examples & Adversarial Training

Adversarial examples, in computer vision, are deliberately modified inputs designed to cause misclassification by a model [18, 59]. These perturbations, initially defined as imperceptible variations in image pixels within small l_1, l_2, and l_∞ norms (uniformly referred as l_p), form the basis for understanding adversarial vulnerabilities in visual models. AT, proposed as an effective defense mechanism, aims to enhance robustness by training models with these adversarial examples [38]. However, it has been observed that AT often leads to a trade-off between adversarial robustness and clean accuracy, presenting a challenge to model generalization [63, 79]. Despite great efforts [24, 50, 52] have been made for mitigating this trade-off, the bad generalization of AT still cannot be fully remedied till now.

In contrast, the NLP realm exhibits a different trend: AT has been found to enhance both the generalization and robustness of language models [9, 40, 41]. Recent studies like the work [88] demonstrate that AT can even boost the performance of transformer-based language foundational models. The work [39] wants to directly copy the success of AT in NLP to enhance the visual features, suggesting discrete representation as a key factor. Although they generate adversarial examples with more imperceptible perturbations than traditional l_p perturbations, the perturbations are still not entirely natural and exhibit domain shift from real-world noise. In this paper, we surmise that the triumph of AT in NLP is derived from the "realness" and "naturalness" of its adversarial examples.

Notably, there have been attempts to use AT for improving clean accuracy in vision tasks. The work [73] employs split batch norms to separate clean and adversarial example statistics, enhancing adversarial feature learning for generalization. However, this operation is not applicable to transformer-based modern foundation models [13, 31, 44]. Another related work to ours is [24], which although similar in name, focuses on using fine-tuning to replace adversarial training to obtain adversarial robustness at low cost. Inspired by works [7, 54] and NLP, we introduce a novel framework ASAM, fine-tuning SAM with "natural" adversarial examples. This approach paves a new path for enhancing visual foundation models, leveraging the "realness" and "naturalness" of adversarial examples to augment SAM's generalization capabilities without substantial additional data or major architectural changes.

3. Method

3.1. Overview

We aim to generate "natural" adversarial images from the SA-1B [31] dataset, and subsequently, employ these generated images along with corresponding SA-1B masks to fine-tune SAM. Note that, during fine-tuning the SAM, we do not modify the SAM structure and incorporate any extra annotated data. Therefore, our proposed ASAM framework achieves the goal of enhancing the generalizability of SAM solely based on its inherent data and structural characteristics. Our proposed ASAM framework contains three steps which are described in detail in the following.

Adversarial Latent Optimization. Existing methods [27, 53, 80, 81, 88] for generating adversarial images typically adhere to l_p norm constraints, resulting in perturbations that are not entirely natural and exhibit a domain shift from real-world noise. In this paper, to generate adversarial examples that are both natural and photorealistic for tuning SAM, we hypothesize that natural images can be first projected onto a low-dimensional manifold via a generative model, such as Stable Diffusion [54]. Subsequently, by optimizing the low-dimensional manifold, we are able to search for a suitable adversarial latent representation, allowing for a reprojection into the natural image domain effectively. We illustrate the process of optimizing adversarial latent representation in *Sec.* 3.2.

Controllable Adversarial Samples Generation. The above optimization process adds slight perturbations to the

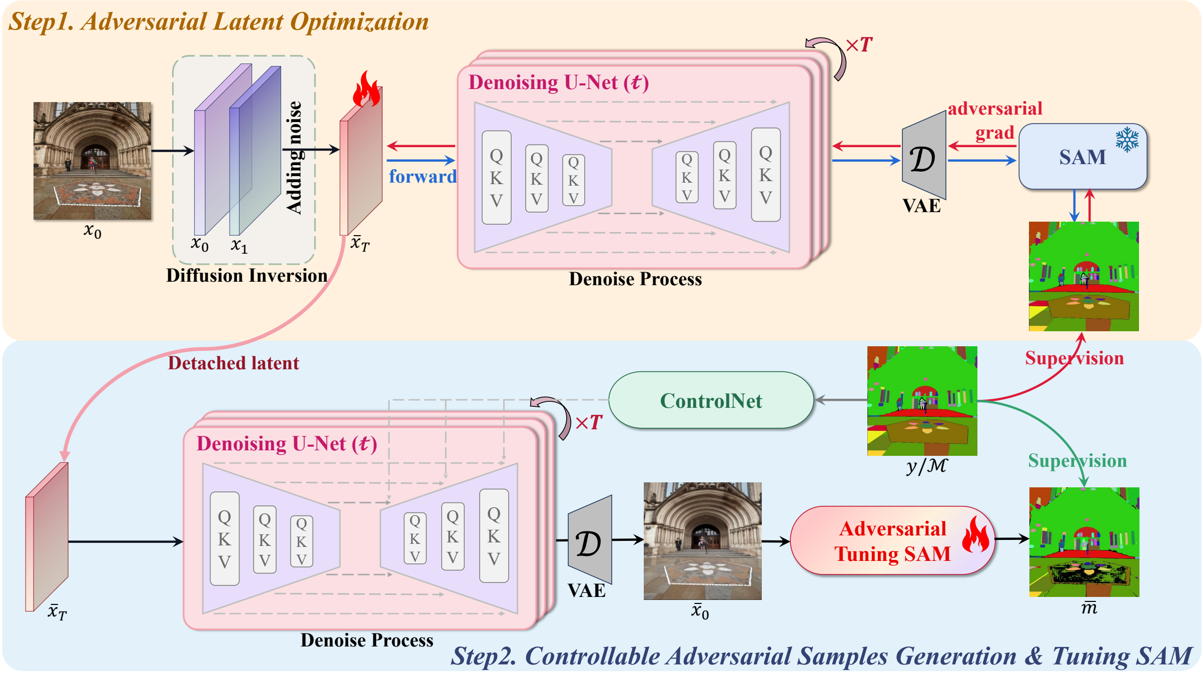

Figure 2. The architecture of our proposed ASAM framework. In the first step, we project the input image into the latent space and then optimize the latent space with adversarial technologies. In the second step, we use the optimized latent to generate adversarial samples controlled by masks. Finally, we fine-tune the SAM with the generated “natural” adversarial samples.

latent representation. Therefore, the naive re-projection may result in the generated adversarial images not aligning properly with the corresponding SA-1B masks. To address this issue, after the optimization is completed, we further design the control branch, which leverages the ControlNet [83] to guide the re-projection process. More details about this process are described in *Sec. 3.3*.

3.2. Adversarial Latent Optimization

Herein, we demonstrate the methodology for searching the adversarial latent representation of SA-1B images within the low-dimensional manifold space of the generative model. Taking into account the balance between computational expense and images quality, we opt for Stable Diffusion as our generative model to produce low-dimensional latent representations. Subsequently, we optimize the generated latent representation which enables the creation of diverse adversarial images.

3.2.1 Projecting Image to Diffusion Latent

The diffusion inversion is commonly used for projecting the image to low-dimensional latent space. In the case of the diffusion model, we employ the DDIM inversion technique [57] which utilizes the conditional embedding $C = \psi(P)$ derived from prompts P using CLIP text encoder, predicated on the premise that the ordinary differential equation procedure is reversible within a finite number of steps:

$$x_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}x_t + (\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t - 1}}) \cdot \epsilon_\theta(x_t, t, C). \tag{1}$$

Given an image x_0, we use a schedule $\{\beta_1, \dots, \beta_T\} \in (0, 1)$, with $\alpha_t = \prod_{i=1}^{t}(1 - \beta_i)$ following [57]. This approach effectively operates in the opposite direction to the denoising process (i.e., $x_0 \rightarrow x_T$ rather than $x_T \rightarrow x_0$), projecting the image x_0 into the latent space at x_T. The text description of each image is generated through BLIPv2 [35].

Text-to-image synthesis frequently emphasizes the role of the prompt, culminating in the introduction of a classifier-free guidance approach [23]. This method generates predictions with no condition and merges them with predictions that are conditioned on specific inputs. Let ω represents the guidance scale factor and $\emptyset = \psi(``")$ denotes the embedding for an empty text prompt, then the formula for classifier-free guidance is articulated as follows:

$$\tilde{\epsilon_\theta}(x_t, t, C, \emptyset) = \omega \cdot \epsilon_\theta(x_t, t, C) + (1 - \omega) \cdot \epsilon_\theta(x_t, t, \emptyset). \tag{2}$$

$\omega = 7.5$ is adopted as the standard setting for Stable Diffusion. During the reverse process of DDIM sampling, the model ϵ_θ forecasts the noise, which might introduce minor inaccuracies at each step. Given its substantial guidance scale parameter ω, the classifier-free guidance method is prone to magnifying these small errors, resulting in a build-up of inaccuracies. Thus, utilizing the reverse DDIM sampling process alongside classifier-free guidance not only disrupts the Gaussian noise distribution but also generates visual anomalies that compromise realism [42].

To mitigate the accumulation of errors, our approach is inspired by the strategy outlined in [42], where we optimize a distinct null text embedding $\emptyset_t$ for each timestep t. Initially, executing the DDIM inverse sampling process with $\omega = 1$ yields a series of successive latent representations $\{x_0^*, ..., x_T^*\}$, starting with $x_0^* = x_0$. Subsequently, we embark on an optimization process for the timesteps $t = \{T, ..., 1\}$, employing $\omega = 7.5$ and setting $\bar{x}_T = x_T^*$:

$$\underset{\emptyset_t}{min} ||x_{t-1}^* - x_{t-1}(\bar{x}_t, t, C, \emptyset_t)||_2^2. \tag{3}$$

For ease of understanding, let $x_{t-1}(\bar{x}_t, t, C, \emptyset_t)$ denote the DDIM sampling step, where $\bar{x}_t$ serves as the input latent, $\emptyset_t$ as the null text embedding, and C is the text embedding. Upon finishing each step, $\bar{x}_{t-1}$ is updated in accordance with the equation:

$$\bar{x}_{t-1} = x_{t-1}(\bar{x}_t, t, C, \emptyset_t). \tag{4}$$

Finally, we can achieve the latent representation $\bar{x}_T = x_T^*$ with the optimized null text embedding $\{\emptyset_t\}_1^T$ generated by the diffusion model. We exploit this latent in the low-dimensional manifold to generate adversarial images.

3.2.2 Adversarial Optimization of Latent

In this section, we undertake an optimization of the latent representation to enhance the generation of natural adversarial images. Within the latent space established by Sec. 3.2.1, the null text embedding $\emptyset_t$ ensures the quality of the reconstructed image, whereas the text embedding C retains the semantic content of the image. Consequently, optimizing both embeddings simultaneously may not lead to optimal outcomes. Considering that the noise $\bar{x}_T$ significantly encapsulates the image's details in the latent space, we opt to focus our optimization efforts on it.

Building upon the latent representation generated in Sec. 3.2.1, we characterize the denoising procedure of the diffusion model as $\Omega(\cdot)$, implemented via the DDIM sampling step. This process encompasses T iterations:

$$\begin{aligned} &\Omega(x_t, T, C, \{\emptyset_t\}_1^T) = \\ &x_0(x_1(..., (x_T, T, C, \emptyset_T), ..., 1, C, \emptyset_1), 0, C, \emptyset_0). \end{aligned} \tag{5}$$

Here, x_t denotes the latent variable at iteration t, with T being the total number of iterations, C standing for the additional conditioning variables, and $\{\emptyset_t\}_1^T$ signifying the sequence of null text embeddings applied at each iteration. The process concludes with the reconstructed image, represented by $\bar{x}_0 = \Omega(\bar{x}_T, T, C, \{\emptyset_t\}_1^T)$. The operations of the Variational Autoencoder (VAE) are not elaborated upon in this manuscript, given its differentiable nature. We frame our adversarial objective optimization as follows:

$$\max_{\delta} \mathcal{L}(\mathcal{S}_\theta(\bar{x}_0), y), \text{ s.t. } ||\delta||_\infty \leq \kappa, \tag{6}$$

In this equation, δ signifies the adversarial perturbation within the latent space, y represents the mask label obtained from the SA-1B dataset, and $\mathcal{S}_\theta$ denotes the SAM with a fixed parameter set θ. The loss function, $\mathcal{L}$, is an amalgamation of mean square error, binary cross-entropy loss, and dice loss, articulated as $\mathcal{L} = \mathcal{L}_{mse} + \mathcal{L}_{bce} + \mathcal{L}_{dice}$. To preserve the consistency between the original image x_0 and its reconstructed counterpart $\bar{x}_0$, we posit that the perturbation δ exerts a minimal impact on this consistency, provided its magnitude is exceedingly slight, namely $||\delta||_\infty \leq \kappa$. The principal challenge is to pinpoint the optimal δ that escalates the segmentation loss. Echoing the approach of traditional adversarial strategies, we utilize gradient-based methods to approximate δ with the formula: $\delta \approx \eta \nabla_{x_T} \mathcal{L}(\mathcal{S}_\theta(\bar{x}_0), y)$, where η is the scale of perturbations aligned with the gradient's direction. By applying the chain rule to unfold $\nabla_{\bar{x}_T} \mathcal{L}(\mathcal{S}_\theta(\bar{x}_0), y)$, we delineate each derivative component:

$$\nabla_{\bar{x}_T} \mathcal{L}(\mathcal{S}_\theta(\bar{x}_T), y) = \frac{\partial \mathcal{L}}{\partial \bar{x}_0} \cdot \frac{\partial \bar{x}_0}{\partial \bar{x}_1} \cdot \frac{\partial \bar{x}_1}{\partial \bar{x}_2} \cdots \frac{\partial \bar{x}_{T-1}}{\partial \bar{x}_T}. \tag{7}$$

3.3. Controllable Adversarial Samples Generation

After obtaining an adversarial latent representation, a reverse diffusion process can be employed to generate the final adversarial examples. However, the optimization process in Stable Diffusion space would introduce minor disturbances to the adversarial latent variables, which would result in misalignment between the generated image's shape and its corresponding label. Intuitively, this issue could potentially be addressed by using more precise prompts in the diffusion model. Nonetheless, the capability of text prompts to control the spatial shape of images is limited, as it's challenging to describe the exact shape of objects through text alone. To overcome this limitation, we additionally train a mask-to-image ControlNet inserted into the reverse process, which offers enhanced spatial shaping capabilities.

ControlNet adjusts the task-specific conditions within the denoising U-Net architecture, aiming to steer the overall behavior of the diffusion model more precisely. The core architecture of the Stable Diffusion model is a U-Net, consisting of an encoder, a middle block, and a decoder that utilizes skip connections. Both the encoder and decoder feature 12 blocks each, culminating in a total of 25 blocks when including the middle block. ControlNet is employed to generate a trainable duplicate of the 12 encoder blocks and the single middle block from the Stable Diffusion model. These

Table 1. Zero-shot segmentation result mIOU comparison on 14 datasets using box prompt.

Methods	DOORS[49]	LVIS[19]	ZeroWaste-f[1]	NDISPark[11]	Egohos[84]	Plittersdorf[21]	BBC038V1[2]	Average
SAM	86.5	80.7	43.0	81.8	81.1	78.7	84.8	76.7
SAM + DAT Tuning	65.0	48.1	35.2	53.2	45.7	31.4	54.6	47.6
SAM + PGD Tuning	68.3	52.5	38.9	60.0	51.8	40.3	65.8	53.9
SAM + DatasetDM	86.0	62.2	29.2	69.5	53.5	70.3	84.2	65.0
ASAM	**87.1**	**81.2**	**46.9**	**82.2**	**82.0**	**79.0**	**85.0**	**77.6**
Methods	**Ade20k[87]**	**VOC2012[15]**	**Cityscapes[12]**	**COCO2017[36]**	**HRSOD-TE[78]**	**CAMO[32]**	**Big[8]**	**Average**
SAM	74.7	79.1	54.1	77.5	88.9	70.7	85.5	75.8
SAM + DAT Tuning	54.6	53.3	31.2	53.7	55.0	52.5	57.6	51.1
SAM + PGD Tuning	58.7	60.3	33.5	58.8	63.4	55.1	63.9	56.2
SAM + DatasetDM	57.4	43.7	44.9	54.1	45.4	26.0	36.4	44.0
ASAM	**75.5**	**80.6**	**56.0**	**79.4**	**91.3**	**73.0**	**87.0**	**77.5**

12 blocks are distributed across four different resolutions (64 × 64, 32 × 32, 16 × 16, 8 × 8), with each resolution comprising three blocks. The generated outputs from these blocks are then integrated into the 12 skip connections and the middle block of the Diffusion U-Net, enhancing its capability to manipulate image characteristics with greater finesse. The operation of ControlNet is denoted as $Z(\cdot;\cdot)$, and it allows for a reconfiguration of the denoising U-Net:

$$n = Dec(Enc(x_t, T, C, \emptyset_t), Z(x_t, T, \mathcal{M}, C, \emptyset_t)), \quad (8)$$

where $\mathcal{M}$ is the mask prompt. Based on the denoising U-Net, we represent the adversarial examples reconstruction:

$$\Omega(\bar{x_t}, T, \mathcal{M}, C, \{\emptyset_t\}_1^T) = \quad (9)$$
$$x_0(x_1(..., (\bar{x_T}, T, \mathcal{M}, C, \emptyset_T), ..., 1, \mathcal{M}, C, \emptyset_1), 0, \mathcal{M}, C, \emptyset_0).$$

3.4. Fine-tuning SAM with Adversarial Samples

Different from previous methods [6, 30, 33, 69] which alter the structure of SAM, our aim is to enhance the overall capabilities of the SAM without any structural modifications. The selection of appropriate parameters for fine-tuning necessitates careful consideration, taking into account factors such as efficiency and the risk of over-fitting. In this regard, we specifically choose to fine-tune the output tokens and mask token of SAM, which accounts for only approximately 0.001% of the total parameters in the SAM. Additionally, to ensure fast convergence on adversarial samples while maintaining generalization, we adopt the learning rate schedule strategy "slow start fast decay", as described in the work [24]. Furthermore, our proposed ASAM indicates that employing only 1% samples from the SA-1B dataset already yields significant performance improvements.

4. Experiment

4.1. Experimental Setting

Implementation Details. We use stable-diffusion-v1-5 [54] pre-trained on the LAION5B [55] dataset. The description of each training image is automatically generated using BLIPv2 [35]. We use ControlNet v1.0 to control the generation process. We use SAM with vit-base backbone. The training dataset used in this paper is *sa_000000* subset from SA-1B dataset. For the adversarial example generation process, we set DDIM steps T to 50, the number of optimization steps of null text embedding to 10, the number of attacks on adversarial samples to 10, and the attacks size κ to 0.02. We fine-tune the SAM with 10 epochs using Adam optimizer. The learning rate first increases linearly from 0.01 to 0.05, then decay exponentially. We adopt 8 NVIDIA 48G A6000 GPUs for training.

Evaluation Datasets. Following SAM [31], we evaluate ASAM on datasets and tasks that are not seen during training. The evaluation datasets may include novel image distributions, such as underwater or ego-centric images that, to our knowledge, do not appear in SA-1B. We use a newly compiled suite of 14 datasets with diverse image distributions under mIoU evaluation, as shown in Table. 1.

4.2. Quantitative and Qualitative Comparison

To thoroughly evaluate the effectiveness of our proposed ASAM, we compare it with four different approaches: the original SAM, SAM fine-tuned with PGD Tuning [53], SAM fine-tuned with DAT Tuning [39], and SAM fine-tuned with new data generated through DatasetDM [70]. As shown in Table. 1, ASAM clearly outperforms the other tuning methods. ASAM, compared to the original SAM, achieves performance improvements across all 14 test datasets, with an average performance increase of ***1.3*** mIoU. This consistent enhancement across a diverse range of datasets underscores the robustness and effectiveness of our approach, demonstrating its capacity to significantly boost the model's capabilities in various contexts. A key reason for this superiority is that SAM has already been trained on a large-scale dataset. Therefore, simply adding noise perturbations to some samples or generating new samples for tuning SAM does not introduce a significantly different data distribution to SAM. In fact, re-tuning might disrupt the originally well-trained parameters of SAM. Different from existing methods like PGD and DAT, our adversar-

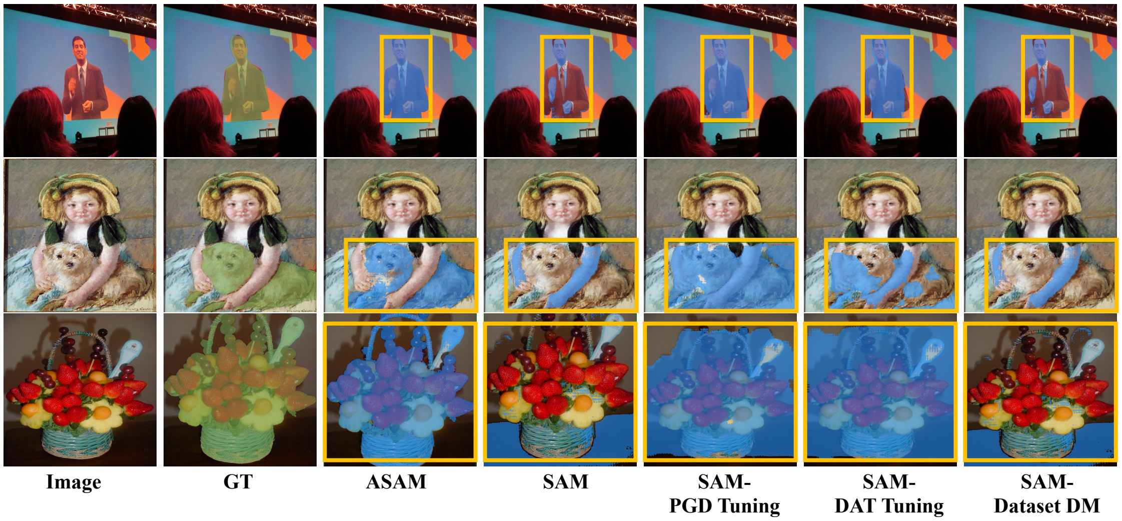

Figure 3. Qualitative comparison of the proposed ASAM and other methods. Yellow boxes represent the box prompts.

Table 2. Ablation studies of main components in ASAM.

Latent Projection	Latent Optimization	Controllable Generation	mIoU
✓			59.1
✓	✓		54.3
✓		✓	69.3
✓	✓	✓	77.6

Table 3. Image quality assessment.

Method	NIMA-AVA↑	HyperIQA ↑	MUSIQ -AVA ↑	TReS ↑
SA-1B	5.18	0.72	4.07	78.46
Inversion	5.19	0.72	4.11	78.59
PGD	5.04	0.69	3.85	76.17
DAT	4.76	0.63	3.83	66.13
Ours	5.20	0.71	4.12	76.73

ial samples are reconstructed from a well-optimized, low-dimensional manifold guided by the gradients of SAM. This approach allows us to more effectively address the shortcomings in SAM's original training. It provides a refined input that is better aligned with SAM's learning paradigm, enabling it to generalize more effectively to new or challenging scenarios. From a visual comparison in Fig. 3, it is evident that our proposed ASAM enhances the performance on samples where the original SAM fell short.

4.3. Ablation Studies

Herein, we conduct ablation studies on the 14 datasets mentioned above to indicate the effectiveness of ASAM.

Main components. As shown in Table. 2, if we solely rely on Latent Projection (Sec. 3.2.1) without employing Latent Optimization (Sec. 3.2.2), performance diminishes as it lacks the guidance from SAM's gradient. This approach misses out on the crucial step of refining the latent representation based on the model's feedback, which is essential for aligning the projection with the model's learned patterns and intricacies. Furthermore, if we use only Latent Projection followed by reconstruction with ControlNet but still omit Latent Optimization, performance again falls short. This combination, while slightly more sophisticated, still fails to leverage the model-specific insights that Latent Optimization provides, thus not fully capitalizing on the potential improvements in the projection process. Finally, when Latent Optimization is combined with ControlNet, we achieve the best segmentation result.

Adversarial Samples Visualization. To validate the utility of the adversarial samples produced in this study for the fine-tuning of SAM, we adopt a quantitative approach to image quality assessment, in line with previous research [56, 77]. Specifically, we employ non-reference perceptual image quality metrics for this purpose. The metrics selected include NIMA [14], HyperIQA [58], MUSIQ [29], and TReS [17]. Both NIMA-AVA and MUSIQ-AVA have been trained on the AVA dataset [43], utilizing the PyIQA framework [4]. As depicted in Table. 3, the inversion images produced in our work maintain comparable image quality to their clean counterparts. Notably, ASAM outperforms other methods in terms of image quality assessment. We further illustrate this with adversarial samples showcased in Fig. 4. It's important to highlight that the perturbations introduced via ASAM are designed to be natural, in contrast to the more artificial alterations typical of other techniques, such as DAT or PGD tuning methods. This approach to generating natural perturbations aims to create authentically challenging examples akin to those encountered in real-world scenarios, thereby potentially improving the model's generalization capabilities.

Framework Transferability. To further assess the transferability of our ASAM framework, we conduct experiments

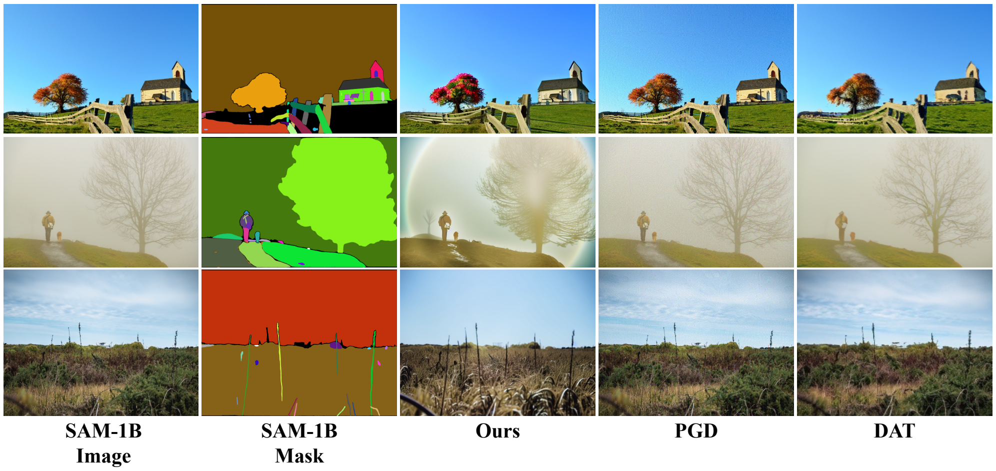

Figure 4. Adversarial examples comparison of ASAM and other attack methods.

Table 4. ESAM vs AESAM on vit-tiny backbone.

Method	Ade20k	VOC2012	Cityscapes	COCO2017	LVIS
ESAM	75.0	81.2	48.7	81.7	81.0
AESAM	**75.6**	**81.5**	**49.8**	**82.0**	**81.4**

on another large vision foundation model, EfficientSAM (ESAM) [74], which is the novel large vision foundation model proposed by Meta in CVPR2024. The results in Table. 4 corroborate the framework's capability to significantly boost ESAM's performance as well. These findings validate our framework's efficacy across different large models, paving the way for boosting the capabilities of large vision foundation models.

5. Discussion & Future work

Although we have demonstrated the effectiveness of our method through extensive empirical experiments, it seems that in addition to the direct inspiration from NLP research, the theoretical underpinnings specific to our method remain an area for further exploration. Fortunately, we have found some existing theoretical work that, although not directly applicable to our task, can provide some theoretical evidence. Specifically, we find that our approach in ASAM aligns with the theoretical framework proposed by Wong and Kolter [68], which emphasizes bridging the gap between real-world perturbations and adversarial defenses. This paper underlines the value of learning perturbation sets directly from data, mirroring our method of using the Stable diffusion model to generate natural adversarial examples. Furthermore, the use of Conditional Variational Autoencoders (CVAEs) for perturbation learning in the paper supports our methodology of manipulating latent space representations. These theoretical insights reinforce the effectiveness of using generative models to create adversarial examples that are not just challenging for the model but also reflect real-world complexities and variations. Although this paper cannot serve as direct theoretical proof for our work, this theoretical backing complements our empirical findings, highlighting the effectiveness of using realistic adversarial examples for enhancing SAM's performance in different real-world scenarios.

This connection, however, is just the beginning of a broader theoretical exploration. Our future work aims to delve deeper into the theoretical aspects of adversarial fine-tuning, specifically in the context of foundation models. We plan to investigate and formalize the principles underlying the efficacy of our method, which could potentially lead to a more generalized theory for enhancing model performance with adversarial examples in the field of computer vision. By establishing a solid theoretical framework, we can further legitimize the use of such techniques and possibly uncover new avenues for improving foundation models' capabilities in diverse real-world applications.

6. Conclusion

ASAM, introduced in this study, represents a significant advancement in the SAM through the innovative use of adversarial tuning. Employing a stable diffusion model to augment a segment of the SA-1B dataset, we generated natural, photorealistic adversarial images, leading to substantial improvements in SAM's segmentation capabilities across various tasks. This method, inspired by adversarial training techniques in NLP, maintains the original architecture and zero-shot strengths of SAM while enhancing its performance. Our findings demonstrate that ASAM not only sets new benchmarks in segmentation tasks but also contributes to the broader application and understanding of adversarial examples in the field of computer vision, offering a novel and effective approach to boosting the capabilities of large vision foundation models.

References

[1] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi M. Alladkani, Ping Hu, Vitaly Ablavsky, Berk Çalli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 21115–21125. IEEE, 2022. 6

[2] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nature methods*, 16(12):1247–1253, 2019. 6

[3] Shurong Chai, Rahul Kumar Jain, Shiyu Teng, Jiaqing Liu, Yinhao Li, Tomoko Tateyama, and Yen-Wei Chen. Ladder fine-tuning approach for SAM integrating complementary network. *CoRR*, abs/2306.12737, 2023. 1

[4] Chaofeng Chen and Jiadi Mo. Iqa-pytorch: Pytorch toolbox for image quality assessment. *Available: https://github.com/chaofengc/IQA-PyTorch*, 2020. 7

[5] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-Lun Chao. Segment anything model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation. *CoRR*, abs/2305.05803, 2023. 2

[6] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. SAM fails to segment anything? - sam-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, and more. *CoRR*, abs/2304.09148, 2023. 1, 3, 6

[7] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. In *NeurIPS*, 2023. 2, 3

[8] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8887–8896. Computer Vision Foundation / IEEE, 2020. 6

[9] Yong Cheng, Lu Jiang, and Wolfgang Macherey. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4324–4333. Association for Computational Linguistics, 2019. 3

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023. 1

[11] Luca Ciampi, Carlos Santiago, João Paulo Costeira, Claudio Gennaro, and Giuseppe Amato. Domain adaptation for traffic density estimation. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2021, Volume 5: VISAPP, Online Streaming, February 8-10, 2021*, pages 185–195. SCITEPRESS, 2021. 6

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society, 2016. 6

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1, 3

[14] Hossein Talebi Esfandarani and Peyman Milanfar. NIMA: neural image assessment. *IEEE Trans. Image Process.*, 27 (8):3998–4011, 2018. 7

[15] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007(1-45):5, 2012. 6

[16] Iraklis Giannakis, Anshuman Bhardwaj, Lydia Sam, and Georgios Leontidis. Deep learning universal crater detection using segment anything model (SAM). *CoRR*, abs/2304.07764, 2023. 3

[17] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *WACV*, pages 3989–3999. IEEE, 2022. 7

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[19] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019. 6

[20] Dongsheng Han, Chaoning Zhang, Yu Qiao, Maryam Qamar, Yuna Jung, SeungKyu Lee, Sung-Ho Bae, and Choong Seon Hong. Segment anything model (SAM) meets glass: Mirror and transparent objects cannot be easily detected. *CoRR*, abs/2305.00278, 2023. 3

[21] Timm Haucke, Hjalmar S. Kühl, and Volker Steinhage. SOCRATES: introducing depth in visual wildlife monitoring using stereo vision. *Sensors*, 22(23):9082, 2022. 6

[22] Haibin He, Jing Zhang, Mengyang Xu, Juhua Liu, Bo Du, and Dacheng Tao. Scalable mask annotation for video text spotting. *CoRR*, abs/2305.01443, 2023. 2

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. 4

[24] Ahmadreza Jeddi, Mohammad Javad Shafiee, and Alexander Wong. A simple fine-tuning is all you need: Towards robust deep learning via adversarial fine-tuning. *CoRR*, abs/2012.13628, 2020. 3, 6

[25] Ge-Peng Ji, Deng-Ping Fan, Peng Xu, Ming-Ming Cheng, Bowen Zhou, and Luc Van Gool. SAM struggles in concealed scenes - empirical study on ”segment anything”. *CoRR*, abs/2304.06022, 2023. 1, 3

[26] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of SAM on different real-world applications. *CoRR*, abs/2304.05750, 2023. 1, 3

[27] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. LAS-AT: adversarial training with learnable attack strategy. In *CVPR*, pages 13388–13398. IEEE, 2022. 3

[28] Leiping Jie and Hui Zhang. When SAM meets shadow detection. *CoRR*, abs/2305.11513, 2023. 3

[29] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. In *ICCV*, pages 5128–5137. IEEE, 2021. 7

[30] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *CoRR*, abs/2306.01567, 2023. 1, 2, 3, 6

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 1, 2, 3, 6

[32] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 184:45–56, 2019. 6

[33] 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. *CoRR*, abs/2307.04767, 2023. 1, 3, 6

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 1

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 4, 6

[36] 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. 6

[37] Jun Ma and Bo Wang. Segment anything in medical images. *CoRR*, abs/2304.12306, 2023. 1, 3

[38] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 3

[39] Xiaofeng Mao, Yuefeng Chen, Ranjie Duan, Yao Zhu, Gege Qi, Shaokai Ye, Xiaodan Li, Rong Zhang, and Hui Xue. Enhance the visual representation via discrete adversarial training. In *NeurIPS*, 2022. 3, 6

[40] Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 3

[41] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1979–1993, 2019. 3

[42] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047. IEEE, 2023. 5

[43] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415. IEEE Computer Society, 2012. 7

[44] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 1, 3

[45] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *CoRR*, abs/2304.07193, 2023. 1

[46] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 2

[47] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Xiaokang Yang, and Wei Shen. SAM-PARSER: fine-tuning SAM efficiently

by parameter space reconstruction. *CoRR*, abs/2308.14604, 2023. 1

[48] Xinyang Pu, Hecheng Jia, Linghao Zheng, Feng Wang, and Feng Xu. Classwise-sam-adapter: Parameter efficient fine-tuning adapts segment anything to SAR domain for semantic segmentation. *CoRR*, abs/2401.02326, 2024. 1

[49] Mattia Pugliatti and Francesco Topputo. DOORS: dataset for boulders segmentation. statistical properties and blender setup. *CoRR*, abs/2210.16253, 2022. 6

[50] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1

[52] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 7909–7919. PMLR, 2020. 3

[53] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, pages 8093–8104. PMLR, 2020. 3, 6

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 1, 2, 3, 6

[55] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 6

[56] Ali Shahin Shamsabadi, Ricardo Sánchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *CVPR*, pages 1148–1157. Computer Vision Foundation / IEEE, 2020. 7

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 4

[58] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, pages 3664–3673. Computer Vision Foundation / IEEE, 2020. 7

[59] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 3

[60] Lv Tang, Peng-Tao Jiang, Haoke Xiao, and Bo Li. Towards training-free open-world segmentation via image prompting foundation models. *CoRR*, abs/2310.10912, 2023. 2

[61] Lv Tang, Haoke Xiao, and Bo Li. Can SAM segment anything? when SAM meets camouflaged object detection. *CoRR*, abs/2304.04709, 2023. 1, 3

[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 1

[63] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 3

[64] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Hannaneh Hajishirzi, Noah A. Smith, and Daniel Khashabi. Benchmarking generalization via in-context instructions on 1, 600+ language tasks. *CoRR*, abs/2204.07705, 2022. 2

[65] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics, 2023. 2

[66] Yuxin Wen, Shuai Li, and Kui Jia. Towards understanding the regularization of adversarial robustness on neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 10225–10235. PMLR, 2020. 2

[67] Dominic Williams, Fraser Macfarlane, and Avril Britten. Leaf only SAM: A segment anything pipeline for zero-shot automated leaf segmentation. *CoRR*, abs/2305.09418, 2023. 3

[68] Eric Wong and J. Zico Kolter. Learning perturbation sets for robust machine learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 8

[69] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical

SAM adapter: Adapting segment anything model for medical image segmentation. *CoRR*, abs/2304.12620, 2023. 1, 3, 6

[70] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *CoRR*, abs/2308.06160, 2023. 6

[71] Haoke Xiao, Lv Tang, Bo Li, Zhiming Luo, and Shaozi Li. Zero-shot co-salient object detection framework. *CoRR*, abs/2309.05499, 2023. 2

[72] Zeyu Xiao, Jiawang Bai, Zhihe Lu, and Zhiwei Xiong. A dive into SAM prior in image restoration. *CoRR*, abs/2305.13620, 2023. 2

[73] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L. Yuille, and Quoc V. Le. Adversarial examples improve image recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 816–825. Computer Vision Foundation / IEEE, 2020. 3

[74] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest N. Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *CoRR*, abs/2312.00863, 2023. 8

[75] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. SAM3D: segment anything in 3d scenes. *CoRR*, abs/2306.03908, 2023. 2

[76] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. Matte anything: Interactive natural image matting with segment anything models. *CoRR*, abs/2306.04121, 2023. 2

[77] Shengming Yuan, Qilong Zhang, Lianli Gao, Yaya Cheng, and Jingkuan Song. Natural color fool: Towards boosting black-box unrestricted attacks. In *NeurIPS*, 2022. 7

[78] Yi Zeng, Pingping Zhang, Zhe L. Lin, Jianming Zhang, and Huchuan Lu. Towards high-resolution salient object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7233–7242. IEEE, 2019. 6

[79] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 7472–7482. PMLR, 2019. 3

[80] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan S. Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *ICML*, pages 11278–11287. PMLR, 2020. 3

[81] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan S. Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *ICLR*. OpenReview.net, 2021. 3

[82] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *CoRR*, abs/2304.13785, 2023. 1

[83] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *CoRR*, abs/2302.05543, 2023. 4

[84] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIX*, pages 127–145. Springer, 2022. 6

[85] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *CoRR*, abs/2305.03048, 2023. 3

[86] Zhenghao Zhang, Zhichao Wei, Shengfan Zhang, Zuozhuo Dai, and Siyu Zhu. UVOSAM: A mask-free paradigm for unsupervised video object segmentation via segment anything model. *CoRR*, abs/2305.12659, 2023. 2

[87] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019. 6

[88] Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2, 3

[89] Yao Zhu, Yuefeng Chen, Xiaodan Li, Kejiang Chen, Yuan He, Xiang Tian, Bolun Zheng, Yaowu Chen, and Qingming Huang. Toward understanding and boosting adversarial transferability from a distribution perspective. *IEEE Trans. Image Process.*, 31:6487–6501, 2022. 2