Supplementary Material for ASAM: Boosting Segment Anything Model with Adversarial Tuning

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

This supplementary material contains the following parts:

- Section 1 shows the generalization of ASAM when transfer it to other SAM-based models, containing HQ-SAM [11] and SAM-Adaptor [3].
- Section 2 shows more qualitative results of ASAM to show why ASAM can help improve the performance of SAM.
- Section 3 provides more ablation experiment of ASAM, such as using different numbers of images to train ASAM and using SAM encoders of different scale when training ASAM.
- Section 4 provides more implement details of ASAM.
- Section 5 shows more details of 14 segmentation datasets used in this paper.

We hope this supplementary material can help you get a better understanding of our work.

*Lv Tang is the corresponding author of this paper

1. Generalization of ASAM

Table 1. HQ-SAM vs. HQ-ASAM on ViT-base backbones.

Methods	Year	COIFT		HRSOD		ThinObject5k-TE		DIS5K-VD		Average	
		miou ↑	mbiou ↑	miou ↑	mbiou ↑	miou ↑	mbiou ↑	miou ↑	mbiou ↑	miou ↑	mbiou ↑
HQ-SAM	NeurIPS2023	76.2	68.2	93.3	88.2	91.6	84.0	84.0	72.1	86.3	78.1
HQ-ASAM	2023	76.6	68.7	94.2	89.2	91.6	84.1	84.9	73.6	86.8	78.9

Table 2. SAM-Adaptor vs ASAM-Adaptor on the camouflage detection dataset.

Methods	Year	CAMO			CHAMELEON		
		$S_m \uparrow$	$F_\beta^{mean} \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$F_\beta^{mean} \uparrow$	$MAE \downarrow$
SAM-Adaptor	ICCV2023 Workshop	0.764	0.663	0.119	0.798	0.682	0.082
ASAM-Adaptor	2023	0.798	0.685	0.114	0.803	0.685	0.071

Table 3. SAM-Adaptor vs ASAM-Adaptor on the shadow detection dataset.

Methods	Year	ISTD
		BER ↓
SAM-Adaptor	ICCV2023 Workshop	5.44
ASAM-Adaptor	2023	4.58

Our paper aims to utilize adversarial examples to create a version of the SAM that is more powerful than the original, without altering SAM's fundamental structure. To this end, we integrate our enhanced version of SAM, obtained through our ASAM approach, into two other SAM-based models: HQ-SAM [11] and SAM-Adaptor [3]. The results presented in Table. 1, Table. 2 and Table. 3 demonstrate that our ASAM significantly improves the performance of these two methods on their respective test samples. This success strongly indicates that the SAM enhanced by our proposed ASAM method can be directly generalized to other SAM-based approaches. Consequently, this can lead to further enhancement of these methods' capabilities in specific scenarios. This finding is particularly noteworthy as it suggests that the improvements rendered by our ASAM are not limited to specific instances or tasks but are broadly applicable to a range of models built upon the SAM architecture.

2. More Qualitative Results of ASAM

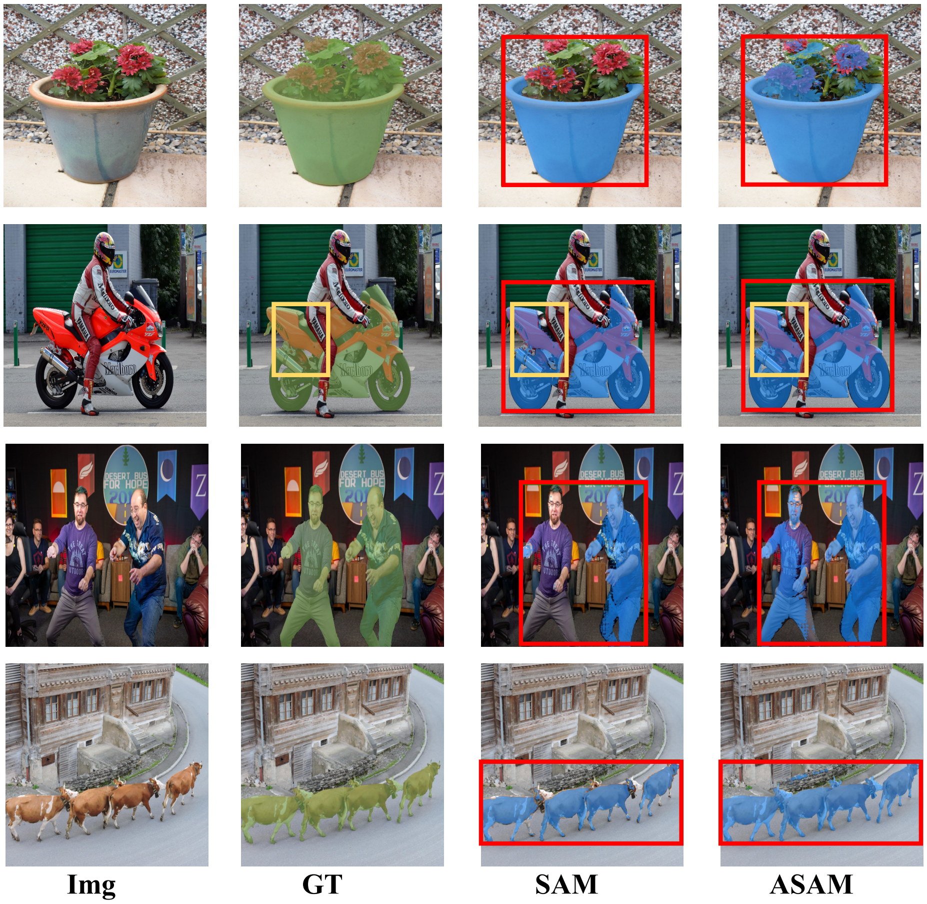

Figure 1. Some qualitative results where our proposed ASAM is better than SAM.

In Fig. 1, we additionally show the zero-shot segmentation results of ASAM and SAM. The red box represents the input box prompt. As shown in Fig. 1, our method either significantly outperforms SAM or exhibits a clear difference in accuracy within the yellow regions.

In Fig. 2, we present additional visualization results to demonstrate how our ASAM effectively enhances the original performance of SAM. We also attempt to use the generated adversarial examples to illustrate why our ASAM can enhance the performance of SAM itself. From first three rows of Fig. 2, it is evident that our generated samples make the foreground and background more similar, which is beneficial for improving SAM's performance in camouflaged scenes. Moreover, some of the generated samples in the last two rows of Fig. 2 exhibit more intricate detail structures than the original samples. The increased complexity and richness in these samples can further aid in enhancing SAM's performance by training it to recognize and process more nuanced and subtle features within an image.

These visualizations not only highlight the effectiveness of ASAM in boosting SAM's capabilities but also provide insights into the kind of adversarial scenarios that are particularly useful for improving model performance in detecting and segmenting objects in complex visual environments.

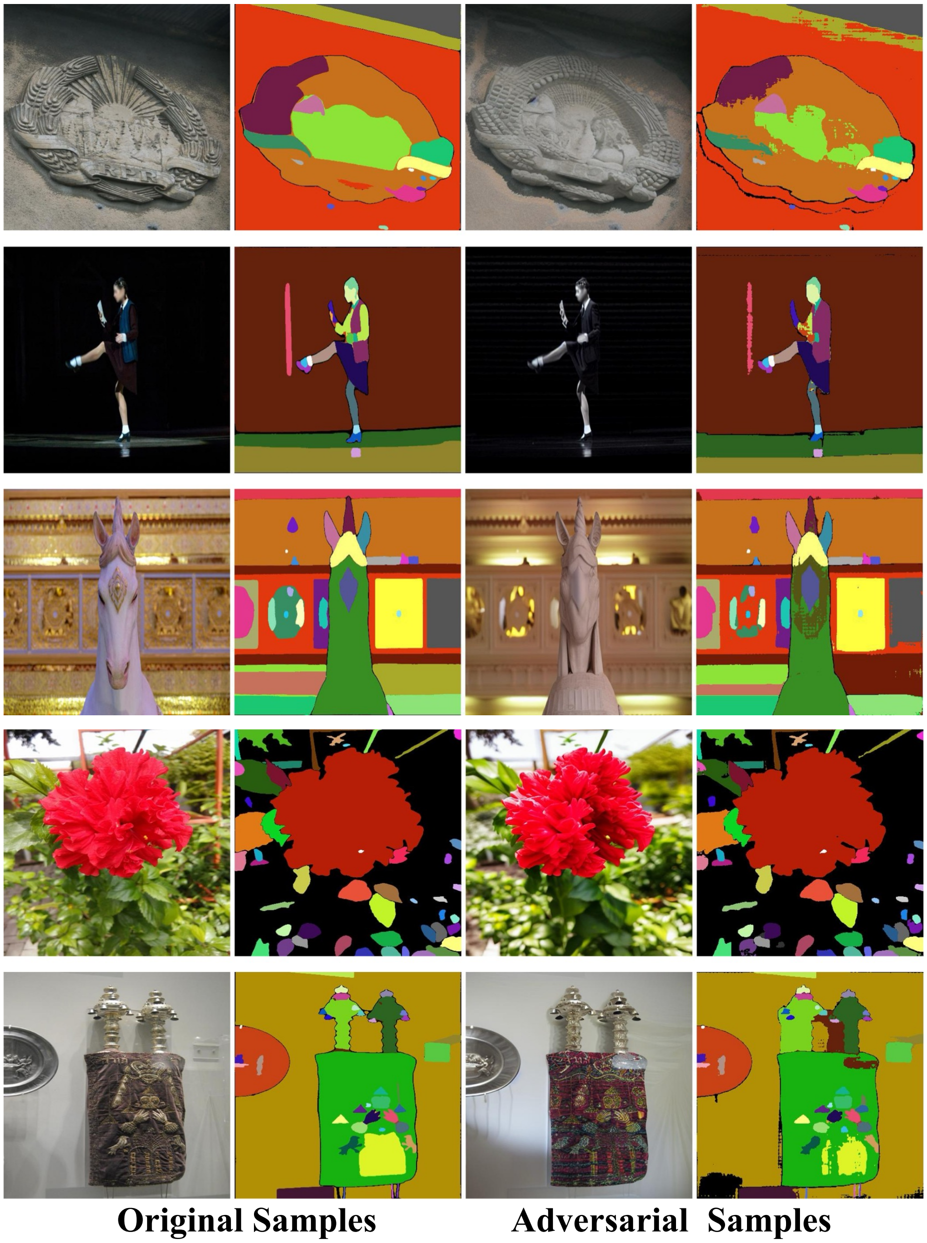

Figure 2. Adversarial examples with low contrast and detailed structural information.

3. More Ablation Experiments of ASAM

3.1. Training ASAM with Different Numbers of Samples.

Table 4. Performance comparison of ASAM using different scale data

Methods	Ade20k	Voc2012	Cityscapes	COCO2017	LVIS	Average
SAM	74.7	79.1	54.1	77.5	80.7	73.2
ASAM(1k)	74.9	79.3	56.3	78.0	80.7	73.8
ASAM(3k)	74.7	79.3	55.6	78.0	80.4	73.6
ASAM(5k)	74.8	79.6	55.5	78.4	80.6	73.8
ASAM(8k)	75.4	80.1	55.9	79.1	81.0	74.3
ASAM(10k)	75.5	80.6	56.0	79.4	81.2	74.5
ASAM(13k)	75.3	80.6	55.9	79.3	81.3	74.5
ASAM(15k)	75.6	80.5	56.1	79.2	81.4	74.6

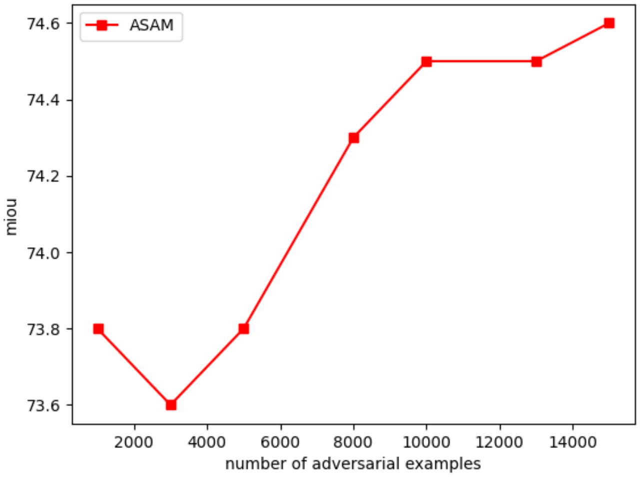

Figure 3. Average miou of ASAM using different scale data.

Herein, we experiment with using varying quantities of adversarial samples to train ASAM, demonstrating its robustness across different sample sizes. As indicated in Table. 4, even with just 1K adversarial samples, there is a notable enhancement in SAM's. Additionally, as the number of samples increases, the performance of ASAM continues to improve. Our primary focus is on validating the effectiveness of our proposed ASAM method, hence we did not solely concentrate on increasing the number of training samples.

3.2. Training ASAM with Different backbone.

Table 5. SAM vs. ASAM on various ViT backbones.

Methods	Ade20k	Voc2012	Cityscapes	COCO2017	LVIS	Average
SAM(vit-large)	76.4	82.7	54.9	80.7	83.1	75.6
ASAM(vit-large)	77.0	82.8	56.8	81.0	83.2	76.2
SAM(vit-huge)	76.9	82.6	57.0	80.9	83.5	76.2
ASAM(vit-huge)	77.3	82.9	57.2	81.2	83.9	76.5

In the main manuscript, we validate the performance of ASAM on a ViT-base SAM. To further assess the robustness of ASAM across different backbone architectures of SAM, we train ASAM on various backbones. The results, as shown in Table. 5, demonstrate that ASAM effectively enhances performance across these different backbone configurations. This consistency in performance improvement across various backbone architectures is a strong indication of ASAM's adaptability and robustness.

3.3. Comparison with model fine-tuning different parameters.

Table 6. Comparison with model fine-tuning different parameters.

Methods	Ade20k	Voc2012	Cityscapes	COCO2017	LVIS	Average
Finetune SAM's Output token	75.7	80.6	56.0	79.4	81.2	74.5
Finetune SAM's Decoder	71.9	79.4	45.4	74.8	76.6	69.9
Fintune the entire SAM	65.6	77.2	47.5	71.4	75.6	67.5

In Table. 6, we explore the impact of different tuning mechanisms within ASAM on its performance. This exploration aimed to understand how various adjustments and refinements in the tuning process could affect the overall effectiveness of ASAM. The results presented Table. 6 indicate that fine-tuning only the output token of SAM yields the best balance between performance and efficiency.

4. More Details of Adversarial Optimization of Latent.

In this section, we delve into two key aspects of Adversarial Optimization of Latent, specifically focusing on 'skip grad' and 'differentiable boundary processing'. As discussed in Section3.2.2 of main manuscript, we propose an optimization approach for adversarial latent variables, which can be expressed as follows:

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

Adversarial Gradient Approximation: Upon examining the elements involved, we discovered that while each item is differentiable, deriving the entire calculation graph is not feasible. Firstly, we analyze the term $\frac{\nabla \mathcal{L}}{\nabla \bar{x}_0}$, which represents the derivative of the SAM with respect to the reconstructed image $\bar{x}_0$ and provides the direction for adversarial gradients. Then, each derivative calculation of $\frac{\nabla \bar{x}_t}{\nabla \bar{x}_{t+1}}$ corresponds to a backpropagation calculation. However, a complete denoising process generates a cumulative number of T calculation graphs, which can lead to memory overflow (similar phenomena are also found in [16]). Consequently, it becomes impractical to directly obtain the gradient for the denoising process. To address this challenge, we introduce the gradient approximation of $\frac{\partial \bar{x}_0}{\partial \bar{x}_T}$. Drawing upon the diffusion process, the denoising step aims to eliminate the Gaussian noise introduced during DDIM sampling [7, 15, 17]. DDIM samples x_t at any given time step t using a closed-form reparameterization trick, as shown:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, \epsilon \in \mathcal{N}(0, I). \tag{2}$$

By rearranging Equation 2, we obtain the following manipulation: $x_0 = \frac{1}{\sqrt{\alpha_t}}x_t - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\epsilon$. Consequently, we have $\frac{\partial x_0}{\partial x_t} = \frac{1}{\sqrt{\alpha_t}}$. Considering that in Stable Diffusion, the timestep t is at most 1000, we can approximate $\lim_{t \to 1000} \frac{\partial x_0}{\partial x_t} = \lim_{t \to 1000} \frac{1}{\sqrt{\alpha_t}} \approx 14.58$. Therefore, we can treat $\frac{\partial \bar{x}_0}{\partial \bar{x}_t}$ as a constant ρ, and Equation 1 can be re-expressed as $\nabla_{\bar{x}_T}\mathcal{L}(\mathcal{F}_\theta(\bar{x}_T), y) = \rho \frac{\partial \mathcal{L}}{\partial \bar{x}_0}$. In summary, skip gradient provide an approximation of the gradients for the denoising process while reducing computational complexity and memory usage.

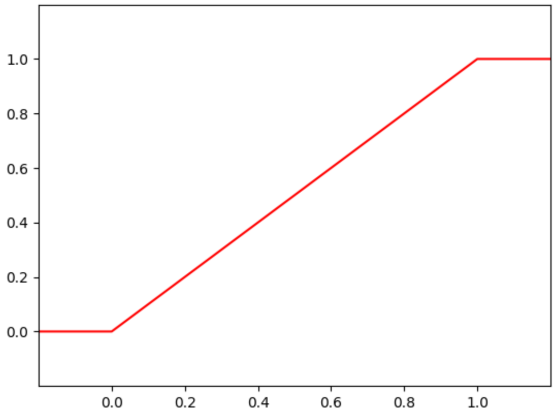

Figure 4. Effectiveness of Differentiable Range Constrain.

Differentiable Range Constrain: The diffusion model lacks an explicit constraint on the value range of $\bar{x}_0$, which can result in exceeding the permissible range. To address this issue, we introduce the technique of differentiable range constrain (DRC) $\tau(\cdot)$. The $\tau(\cdot)$ is to ensure that values outside the range of $[0, 1]$ are constrained within the range of $[0, 1]$ by employing a specially designed differentiable $tanh$ function. The mathematical expression for DRC is as follows:

$$\tau(\cdot) = \begin{cases} tanh\left(\frac{1000x}{10000}\right), & \text{if } x < 0 \\ x, & \text{if } 0 \leq x \leq 1 \\ tanh\left(\frac{1000(x-1)}{10001}\right), & \text{if } x > 1 \end{cases} \tag{3}$$

The image of the function DRC is illustrated in Fig. 4. Our proposed DRC effectively limits the range of values in the image and guarantees differentiability of the gradient.

5. More Details of 14 Datasets

Table 7. Segmentation datasets used to evaluate zero-shot segmentation with point and box prompts. The 14 datasets cover a broad range of domains; see column "image type".

dataset	abbreviation & link	image type	description	mask type	source split	#image nums	#mask nums
Ade20k [19]	ADE20k	Scenes	Object and part segmentation masks for images from SUN and Places datasets.	Instance	Validation	2000	45576
VOC2012 [8]	VOC2012	Scenes	A commonly used dataset for object detection and image segmentation tasks.	Instance	Validation	1449	3488
Cityscapes [6]	Cityscapes	Driving	Stereo video of street scenes with segmentation masks.	Panoptic	Validation	500	17656
COCO2017 [13]	COCO	Scenes	A large-scale object detection, segmentation, key-point detection, and captioning dataset.	Instance	Validation	5000	36480
HRSOD-TE [18]	HRSOD	Saliency	It is specifically designed for high-resolution salient object detection.	Instance	Test	400	400
CAMO [12]	CAMO	Camouflage	A dataset specifically designed for the task of camouflaged object segmentation.	Instance	Test	250	250
Big [4]	Big	Ultral-high resolution	A high-resolution semantic segmentation dataset, every image in the dataset has been carefully labeled by a professional.	Instance	Test, validation	300	300
DOORS [14]	DOORS	Boulders	Segmentation masks of single boulders positioned on the surface of a spherical mesh.	Instance	DS1	10000	10000
LVIS [9]	LVIS	Scenes	Additional annotations for the COCO [66] dataset to enable the study of long-tailed object detection and segmentation	Instance	Validation (v1.0)	5000	52160
ZeroWaste-f [1]	ZeroWaste-f	Recycling	Segmentation masks in cluttered scenes of deformed recycling waste.	Instance	Train	2947	6155
NDISPark [5]	NDISPark	Parking lots	Images of parking lots from video footage taken at day and night during different weather conditions and camera angles for vehicle segmentation.	Instance	Train	111	2577
Egohos [2]	EgoHOS	Egocentric	Fine-grained egocentric hand-object segmentation dataset. Dataset contains mask annotations for existing datasets.	Instance	Validation	1124	3792
Plittersdorf [10]	Plittersdorf	Stereo images	Segmentation masks of wildlife in images taken with the SOCRATES stereo camera trap.	Instance	Train, validation, test	187	546
BBC038V1 [2]	BBBC038v1	Microscopy	Biological images of cells in a variety of settings testing robustness in nuclei segmentation.	Instance	Train	670	34064

In Table. 7, we show more details of 14 segmentation datasets used in this paper.

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. 9

[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. 9

[3] 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, 2

[4] 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. 9

[5] 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. 9

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *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. 9

[7] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 7

[8] 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. 9

[9] 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. 9

[10] Timm Haucke, Hjalmar S. Kühl, and Volker Steinhage. SOCRATES: introducing depth in visual wildlife monitoring using stereo vision. *Sensors*, 22(23):9082, 2022. 9

[11] 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

[12] 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. 9

[13] 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. 9

[14] Mattia Pugliatti and Francesco Topputo. DOORS: dataset for boulders segmentation. statistical properties and blender setup. *CoRR*, abs/2210.16253, 2022. 9

[15] 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. 7

[16] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *ICML*, pages 29894–29918. PMLR, 2023. 7

[17] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 7

[18] 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. 9

[19] 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. 9