Unsegment Anything by Simulating Deformation

Supplementary Material

The supplementary material is organized as follows: In Sec. 1, we present our findings regarding the challenges of transferring prompt-specific attacks to unseen prompts. We offer additional ablation studies in Sec. 2, covering topics including (1) the combination with other transferability methods, and (2) the impact of source models. These sections aim to enhance the reader's comprehension of our approach's underlying mechanisms. In Sec. 5, we offer visualizations of our attacks and baseline attacks from a panoramic perspective to provide a more straightforward comparison. In Sec. 4, we provide a table listing all the notations used throughout this paper. Finally, in Sec. 6, we discuss the limitations of our work and the potential societal impact that may arise from our research.

1. Prompt-Specific Attack Fail to Transfer

We present visualizations of adversarial noises alongside their corresponding segmentation results for various prompts in Fig. 7. These visualizations underscore the heterogeneous nature of prompt-specific adversarial noises.

Notably, the adversarial noises induced by spatial prompts (the first three rows) exhibit distinct characteristics compared to those induced by semantic prompts (the last row). The adversarial noises from the spatial prompts share similarities, characterized by shattered and random-noise-like patterns. Conversely, the adversarial noise stemming from the semantic prompt aligns closely with essential image features.

Furthermore, we observe that the generated adversarial examples tend to exhibit overfitting to the specific prompts used during their generation. Consequently, these adversarial examples struggle to generalize to unseen prompts. Attacks generated using spatial prompts have minimal impact on segmentation results driven by text prompts. Similarly, the adversarial sample generated from a text prompt has little effect on box prompts.

Results. We present the histogram of feature similarities between TAP and AA attacks in Fig. 2. The findings reveal that both targeted and untargeted feature disruption attacks effectively alter features in the source model. However, untargeted attacks are notably less effective on the target model. We hypothesize this is due to the high-dimensional nature of the data, which often leads these attacks to stray from the image distribution. As a result, in the target model, adversarial features appear similar to normal features, indicating lower transferability for untargeted attacks.

2. Ablation Studies

2.1. Study 1: Combining Transferability Methods

Previous research on the transferability of adversarial examples has highlighted four distinct technical approaches, as discussed in Section 5. These approaches encompass feature disturbance, gradient momentum, input augmentation, and model ensembling. Notably, our UAD method falls under the category of feature disturbance. In the precondition of not introducing external model information (we will investigate the impact of model ensembling in the next subsection), we concentrate on exploring how gradient momentum and input augmentation could potentially help us to reach our objective.

As recently demonstrated in a comprehensive benchmark study [58], under a fair and rigorous comparison, the most effective gradient momentum and input augmentation methods are, in fact, the most classic ones, specifically MI [6] and DI [47], respectively. In Table 1, we have already presented results indicating that the inclusion of both of these techniques does not significantly enhance attack performance. Now we will separately integrate each of these techniques into our method, given that they should operate independently of each other, and examine their combined effects in conjunction with our proposed approach in Tab. 2.

Contrary to our expectations, our method, when used alone without the inclusion of MI or DI tricks, yielded the best results. The addition of gradient momentum proved to be more effective than data augmentation. However, combining both techniques resulted in a drop in performance.

2.2. Study 2: Source Model Selection

Previous research findings provide valuable insights: (1) Adversarial examples generated by high-capacity (more over-parameterized) models exhibit higher transferability to low-capacity networks, in contrast to adversarial samples crafted by low-capacity networks, which have limited success when transferred to high-capacity networks; (2) Employing an ensemble of networks proves to be more effective in generating transferable adversarial samples.

To evaluate the upper limits of our approach in tackling the *anything unsegmentable* task, we conducted an ablation study to assess the enhanced transferability of adversarial examples generated from a more capable model.

In Table 1, we performed all experiments using SAM-B as the source model, which has one of the smallest parameter sizes (91 M). Now, we aim to generate adversarial samples based on larger and more powerful models, such

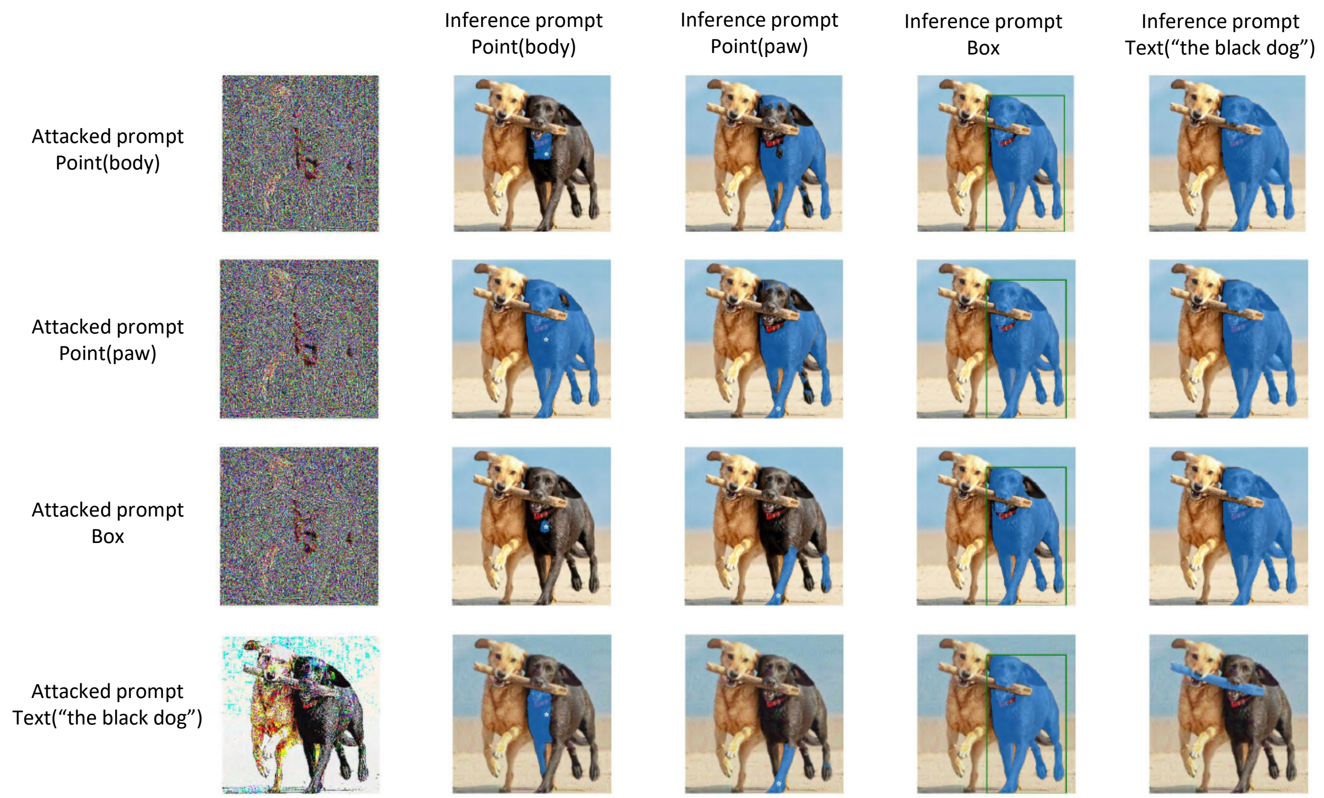

Figure 7. We claim that prompt-specific attacks exhibit fundamental differences in the adversarial noise they generate, and their transferability is limited to a narrow range of prompts. Adversarial examples tend to overfit to the prompts used during the attack phase and have limited impact on unseen prompts.

Approach \ Models	SAM-B (source)			SAM-L			SAM-H			FastSAM		
	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑
UAD	**51.53**	43.89	20.79	**66.07**	**26.44**	**12.27**	**68.96**	**23.42**	**10.23**	28.83	69.95	59.63
UAD + DI [47]	56.74	41.01	18.51	70.19	24.54	11.07	71.25	22.15	8.83	28.88	65.38	56.69
UAD + MI [6]	55.99	**44.29**	**21.54**	67.79	25.29	11.80	69.62	23.22	9.97	**27.44**	**70.97**	**59.81**
UAD + MI [6] + DI [47]	56.49	42.06	19.24	67.94	24.76	11.32	69.26	23.03	9.51	27.60	70.36	59.58

Table 2. Results of combining our method with gradient momentum and input augmentation.

as SAM-L (308 M parameters) and SAM-H (636 M parameters). Additionally, we conducted experiments by ensembling SAM-B and SAM-L. We couldn't ensemble SAM-H due to its high GPU memory consumption, as we have limited computational resources.

As indicated in Table 3, our attack, much like many other adversarial attacks, demonstrates a tendency to overfit to the source model. Specifically, when the source and target models are identical, the attack performs significantly better when the source model does not encompass the target model. Interestingly, model ensembling further enhances attack results; for instance, ensembling SAM-B and SAM-L surpasses the performance of using SAM-B alone by a considerable margin. Ensembling exhibits a stronger impact on the global results, leading to a lower mean IoU (mIoU).

3. Algorithm Pseudo-code

We present the pseudo-code of our attack in Alg.1.

4. Notations

We put all symbols appeared in this paper in Tab. 4 for reference.

5. More visualization results

We visualize the attack effect under segment everything mode (which provides a panoramic view without prompts) on SAM-B and SAM-H models in Fig.8, Fig. 9 and 10. We

Source Models \ Target Models	SAM-B			SAM-L			SAM-H			FastSAM		
	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑	mIoU↓	ASR@50↑	ASR@10↑
SAM-B (91 M)	51.53	43.89	20.79	66.07	26.44	12.27	68.96	23.42	10.23	28.83	69.95	59.63
SAM-L (308 M)	61.67	35.65	13.45	**55.41**	**28.50**	**15.53**	68.98	23.37	10.41	31.27	63.23	52.27
SAM-H (636 M)	61.06	35.69	13.87	63.92	24.32	13.04	**63.31**	**25.85**	**12.61**	30.60	64.59	53.75
SAM-B + SAM-L (Ensemble)	**50.54**	**44.18**	**22.61**	59.67	26.88	14.40	68.35	23.81	11.35	**27.36**	**71.38**	**60.90**

Table 3. Ablation study on source models used to craft adversarial examples.

Algorithm 1 Unsegment Anything by Simulating Deformation

Input: Input image: I;
Deformation parameters: w;
Maximal deformation iterations: T_D;
Maximal proxy perturbation iteration: T_f;
Maximal perturbation iterations: T;
Perturbation step size: α;
Perturbation range: ϵ;
Output: Adversarial perturbation: r

procedure UAD($I, w, T_D, T_f, T, \alpha, \epsilon$)
  $I' = I, r = 0, t_D = 0, t = 0$;
  Initialize w so that D_w produces identity mapping;
  While $t_D < T_D$ **do** ▷ Stage 1: Deformation
    $\hat{I} = \mathcal{D}_w(I)$; ▷ Get deformed image
    $I'' = I$;
    $t_f = 0$;
    While $t_f < T_f$ **do** ▷ proxy adversarial sample
      $I'' = I'' - \alpha \cdot sign(\nabla_{I''}\mathcal{L}_F(\hat{I}, I''))$;
      $I'' = clip_\epsilon(I'' - I) + I$
      $I'' = clip_{0,1}(I'')$
    end While
    $\mathcal{L}(w) = \mathcal{L}_D(\hat{I}, I) + \mathcal{L}_C(w) + \mathcal{L}_F(\hat{I}, I'')$
    $w = w - \nabla_w \mathcal{L}(w)$ ▷ Update deformation parameters
  end While
  $\hat{I} = \mathcal{D}_w(I)$; ▷ End of stage 1, deformed target fixed
  While $t < T$ **do** ▷ Stage 2: Simulation
    $I' = I' - \alpha \cdot sign(\nabla_{I'}\mathcal{L}_F(\hat{I}, I'))$
    $I' = clip_\epsilon(I' - I) + I$
    $I' = clip_{0,1}(I')$
  end While
  $r = I' - I$
  return r
end procedure

compare the attack results with Attack-SAM and PATA++ to highlight the difference in failure patterns and our effectiveness.

Table 4. Notation Table

Variable	**Description**
I	Original clean image
P	Prompt
M	Mask
f_{θ^I}	Image encoder of the promptable segmentation model
h_{θ^P}	Prompt encoder of the promptable segmentation model
g_{θ^M}	Prompt encoder of the promptable segmentation model
w	Deformation control parameters
$\mathcal{D}_w$	Deformation function
$\hat{I}$	Deformed (target) image
w_{ff}	Parameters of flow field which controls deformation
$w_{ff}^{(i,j)}$	Flow vector of position (i, j) in w_{ff}
∇u	Movement in width indicated by flow field
∇v	Movement in height indicated by flow field
r	Adversarial perturbation (adversarial noise)
$I' = I + r$	Adversarial image
T_D	Deformation iterations
T_f	Proxy adversarial update iterations
T	Adversarial update iterations
α	Adversarial perturbation step size
ϵ	Adversarial perturbation range
$\mathcal{L}_D$	Deformation loss
$\mathcal{L}_C$	Control loss
$\mathcal{L}_F$	Fidelity loss

Figure 8. Visualizations of attack results in panoramic view(I)

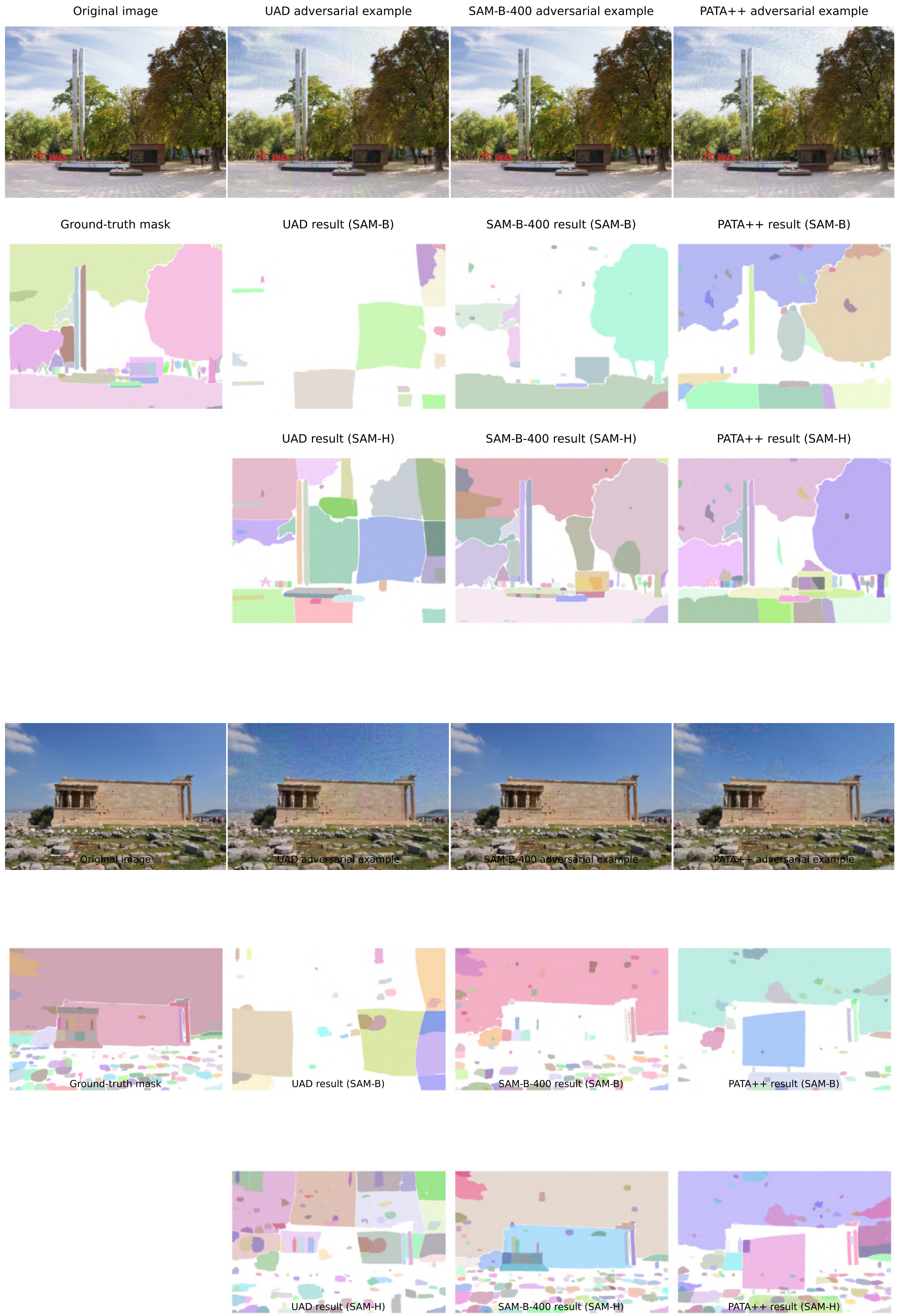

Figure 9. Visualizations of attack results in panoramic view(II)

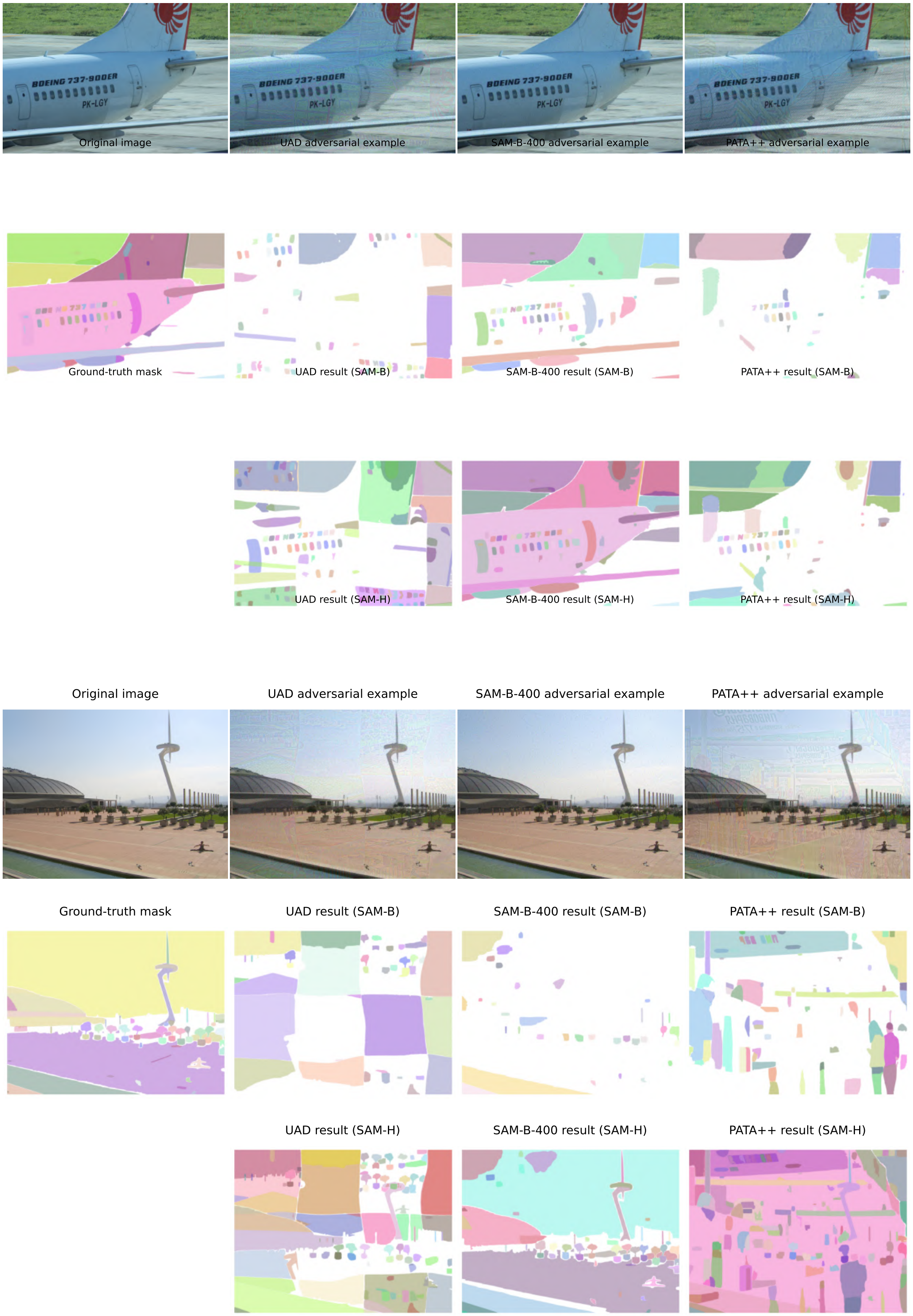

Figure 10. Visualizations of attack results in panoramic view(III)

6. Limitation and social impact

Limitation

While we have made a progressive step in the task of "Anything Unsegmentable", successfully creating an attack that is effective and transferable across several models trained under the Segment Anything task, we found it challenging when evaluating our attack on Segment Everything Everywhere Models (SEEM) [61]. The reason behind this lack of transferability may stem from fundamental differences in training data and tasks: SEEM is trained on COCO2017 [30] with panoptic segmentation annotations. Consequently, the feature space of the SEEM model inherently contains rich information about semantic labels, which is significantly different with SAM family. We believe that this divergence in feature space is the primary reason our attack did not transfer successfully.

However, we are optimistic about the potential for improvement. By introducing additional loss term that targets the category feature space, we anticipate the development of new and more powerful adversarial attacks capable of simultaneously compromising SAM, SEEM, and even more promptable segmentation models.

Social Impact

Our primary goal is to protect the personal digital content from potential copyright infringement and privacy breaches. We envision users employing our approach to preprocess their digital assets before uploading them to public websites, thereby reducing the risk of misuse or theft of their photos and digital creations.

An alternative approach, instead of incorporating adversarial attacks, could involve implementing protective measures directly within the segmentation models themselves. For instance, model publishers might consider adopting a consensus not to perform valid segmentations on protected data. However, establishing and enforcing such a consensus is a complex challenge. Moreover, addressing the issue of models that have already been downloaded and deployed by potential adversaries presents its own set of difficulties.

References

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

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 8

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[17] Yihao Huang, Yue Cao, Tianlin Li, Felix Juefei-Xu, Di Lin, Ivor W Tsang, Yang Liu, and Qing Guo. On the robustness of segment anything. *arXiv preprint arXiv:2305.16220*, 2023. 1

[18] Yihao Huang, Yue Cao, Tianlin Li, Felix Juefei-Xu, Di Lin, Ivor W Tsang, Yang Liu, and Qing Guo. On the robustness of segment anything. *arXiv preprint arXiv:2305.16220*, 2023. 4, 8

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[42] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting ev-

erything in context. *arXiv preprint arXiv:2304.03284*, 2023. 8

[43] Yuqing Wang, Yun Zhao, and Linda Petzold. An empirical study on the robustness of the segment anything model (sam). *arXiv preprint arXiv:2305.06422*, 2023. 1

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

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

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

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

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

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

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

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

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

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

[54] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 1

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

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

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

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

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

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

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