SCIENCE CHINA
Information Sciences

• PERSPECTIVE •

December 2023, Vol. 66 226101:1–226101:3
https://doi.org/10.1007/s11432-023-3881-x

SAM struggles in concealed scenes — empirical study on "Segment Anything"

Ge-Peng JI[1], Deng-Ping FAN[2], Peng XU[3*], Bowen ZHOU[3],
Ming-Ming CHENG[4] & Luc VAN GOOL[2]

[1]*School of Computing, Australian National University, Canberra* 2601*, Australia;*
[2]*Computer Vision Lab (CVL), ETH Zurich, Zurich* 8092*, Switzerland;*
[3]*Department of Electronic Engineering, Tsinghua University, Beijing* 100084*, China;*
[4]*College of Computer Science, Nankai University, Tianjin* 300071*, China*

Received 15 April 2023/Revised 1 July 2023/Accepted 28 September 2023/Published online 8 November 2023

Citation Ji G-P, Fan D-P, Xu P, et al. SAM struggles in concealed scenes — empirical study on "Segment Anything". Sci China Inf Sci, 2023, 66(12): 226101, https://doi.org/10.1007/s11432-023-3881-x

Large models open up new opportunities for artificial intelligence. In the past few months, there has been a boom in training foundation models on the vast linguistic corpus to produce amazing applications, e.g., ChatGPT, GPT-4. Both natural language processing and multimodal learning communities have been revolutionized. Large models' capacity for generalization and emergent makes it easy for users to believe that large models can solve anything.

Very recently, the "Segment Anything" [1] project was released, and its Segment Anything Model (SAM) is a large ViT-based model trained on the large visual corpus (SA-1B). This is a ground-breaking step toward artificial general intelligence, as SAM demonstrates promising segmentation capabilities in various scenarios and the great potential of the foundation models for computer vision. Like all computer vision researchers, we cannot wait to probe the performance traits of SAM to help the community comprehend it further. Moreover, it is interesting to explore the situations in which SAM does not work well.

In this study, we compare SAM quantitatively with cutting-edge models on camouflaged object segmentation tasks and present diversified visualization results in three concealed scenes, i.e., camouflaged animals, industrial defects, and medical lesions. Our main observation is that SAM looks not skillful in concealed scenes.

Experiment. We use three frequently used camouflaged object segmentation (COS) benchmarks to evaluate SAM. If under the unprompted setting, SAM generates multiple binary masks and can pop out several potential objects within an input. For a fair evaluation of interesting regions in a specific segmentation task, we take a strategy to select the most appropriate mask based on its ground-truth mask. Formally, given N binary predictions $\{\boldsymbol{P}_n\}_{n=1}^{N}$ and the ground-truth $\boldsymbol{G}$ for an input image, we calculate intersection over union (IoU) scores for each pair to generate a set of evaluation scores $\{\mathrm{IoU}_n\}_{n=1}^{N}$. We finally select the mask with the highest IoU score from this set.

Our evaluation protocols are following the standard practice as in [2]. (1) Datasets: We use three commonly-used COS benchmarks, including CAMO [3], COD10K [4], and NC4K [5]. (2) Models: To ensure a fair comparison with SAM, we choose the current top-performing COS models using transformer architecture, i.e., CamoFormer-P/-S [6], HitNet [7]. (3) Metrics: We use five commonly-used evaluation metrics, S-measure (S_α), E-measure (E_ϕ), F-measure (F_β), weighted F-measure (F_β^w), and MAE (M).

We report the quantitative comparison in Figure 1(a) [8–10], SAM demonstrates significant improvements as model capabilities increase from ViT-B to ViT-L, with an increase in F_β^w score from 0.353 to 0.655 on CAMO. However, the improvement is limited when the model becomes larger, increasing only from 0.655 (ViT-L) to 0.700 (ViT-H). Moreover, we observe that there remains a large gap between SAM even with ViT-H and current top-performing COS models on three benchmarks. For example, as presented in Figure 1(a), the difference of E_ϕ^{mx} score between SAM (ViT-H) and CamoFormer-S [6] reaches 13.8% on COD10K, 25.6% on CAMO, and 16.9% on NC4K. This gap indicates that the perception ability of SAM needs further improvement for concealed scenes.

We further qualitatively evaluate SAM in three concealed scenarios, and several interesting findings are as follows. All the visualization results are generated by the online demo of SAM. (1) Camouflaged animal. As presented in Figure 1(b), it is difficult for SAM to detect concealed animals in their natural habitat. For instance, SAM fails to segment a mantis crouching on a leaf (in the second column) and a seahorse swimming in an orange coral reef (in the last column). In these two cases, SAM struggles to distinguish the target semantics from their surroundings because the foreground and background share similar appearances of shape and color. As a result, SAM becomes more dependent on unreliable

*Corresponding author (email: peng_xu@tsinghua.edu.cn)

Model	Pub/Year	Backbone	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$E_\phi^{ad} \uparrow$	$E_\phi^{mn} \uparrow$	$E_\phi^{mx} \uparrow$	$F_\beta^{ad} \uparrow$	$F_\beta^{mn} \uparrow$	$F_\beta^{mx} \uparrow$
COD10K [4]											
CamoFormer-P [6]	arXiv_{23}	PVTv2-B4 [8]	0.869	0.786	**0.023**	0.931	0.932	0.939	0.794	0.811	0.829
CamoFormer-S [6]	arXiv_{23}	Swin-B [9]	0.862	0.772	0.024	0.932	0.931	**0.941**	0.780	0.799	0.818
HitNet [7]	AAAI_{23}	PVTv2-B2 [8]	**0.871**	**0.806**	**0.023**	**0.936**	**0.935**	0.938	**0.818**	**0.823**	**0.838**
SAM [1]	arXiv_{23}	ViT-B [10]	0.585	0.353	0.108	0.535	0.533	0.535	0.423	0.422	0.423
		Difference (Δ)	−28.6%	−45.3%	+8.5%	−40.1%	−40.2%	−40.3%	−39.5%	−40.1%	−41.5%
		ViT-L [10]	0.751	0.655	0.065	0.766	0.764	0.766	0.718	0.716	0.718
		Difference (Δ)	−12%	−15.1%	+4.2%	−17%	−17.1%	−17.2%	−10%	−10.7%	−12%
		ViT-H [10]	0.781	0.700	0.054	0.800	0.798	0.800	0.756	0.754	0.756
		Difference (Δ)	−9%	−10.6%	+3.1%	−13.6%	−13.7%	−13.8%	−6.2%	−6.9%	−8.2%
CAMO [3]											
CamoFormer-P [6]	arXiv_{23}	PVTv2-B4 [8]	0.872	0.831	0.046	0.931	0.929	**0.938**	0.853	0.854	0.868
CamoFormer-S [6]	arXiv_{23}	Swin-B [9]	**0.876**	**0.832**	**0.043**	**0.935**	**0.930**	**0.938**	**0.856**	**0.856**	**0.871**
HitNet [7]	AAAI_{23}	PVTv2-B2 [8]	0.849	0.809	0.055	0.910	0.906	0.910	0.833	0.831	0.838
SAM [1]	arXiv_{23}	ViT-B [10]	0.462	0.238	0.219	0.402	0.401	0.402	0.312	0.312	0.312
		Difference (Δ)	−41.4%	−59.4%	+17.6%	−53.3%	−52.9%	−53.6%	−54.4%	−54.4%	−55.9%
		ViT-L [10]	0.630	0.534	0.162	0.628	0.626	0.628	0.617	0.615	0.617
		Difference (Δ)	−24.6%	−29.8%	+11.9%	−30.7%	−30.4%	−31%	−23.9%	−24.1%	−25.4%
		ViT-H [10]	0.677	0.594	0.136	0.682	0.680	0.682	0.670	0.668	0.670
		Difference (Δ)	−19.9%	−23.8%	+9.3%	−25.3%	−25%	−25.6%	−18.6%	−18.8%	−20.1%
NC4K [5]											
CamoFormer-P [6]	arXiv_{23}	PVTv2-B4 [8]	**0.892**	**0.847**	**0.030**	**0.941**	**0.939**	**0.946**	**0.863**	**0.868**	**0.880**
CamoFormer-S [6]	arXiv_{23}	Swin-B [9]	0.888	0.840	0.031	**0.941**	0.937	**0.946**	0.857	0.863	0.877
HitNet [7]	AAAI_{23}	PVTv2-B2 [8]	0.875	0.834	0.037	0.928	0.926	0.929	0.854	0.853	0.863
SAM [1]	arXiv_{23}	ViT-B [10]	0.544	0.334	0.166	0.494	0.493	0.494	0.403	0.403	0.403
		Difference (Δ)	−34.8%	−51.3%	+13.6%	−44.7%	−44.6%	−45.2%	−46%	−46.5%	−47.7%
		ViT-L [10]	0.728	0.643	0.101	0.735	0.733	0.735	0.706	0.704	0.706
		Difference (Δ)	−16.4%	−20.4%	+7.1%	−20.6%	−20.6%	−21.1%	−15.7%	−16.4%	−17.4%
		ViT-H [10]	0.763	0.696	0.087	0.777	0.775	0.777	0.752	0.750	0.752
		Difference (Δ)	−12.9%	−15.1%	+5.7%	−16.4%	−16.4%	−16.9%	−11.1%	−11.8%	−12.8%

(a)

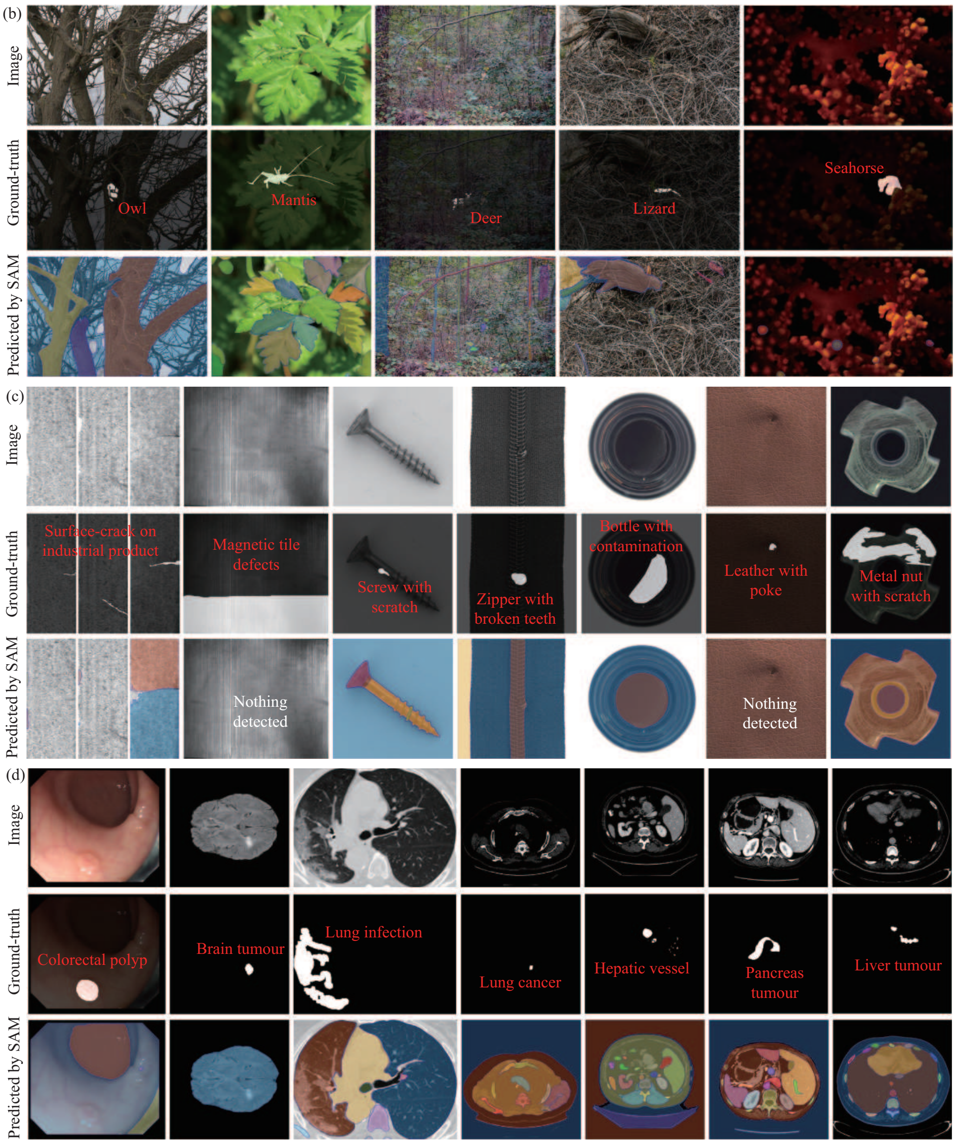

Figure 1 (Color online) (a) Quantitative comparison on three popular COS benchmarks. The symbols ↑/↓ indicate that a higher/lower score is better. The highest scores are marked in bold. Δ represents the difference between SAM and the highest score achieved by current cutting-edge COS models. (b) SAM [1] fails to perceive the animals that are visually "hidden" in their natural surroundings; (c) SAM [1] performs unskillfully in detecting concealed defects in industrial scenes; (d) SAM [1] fails to detect the lesion regions in various medical modalities.

pixel intensity changes along the boundaries. (2) Industrial defect. In this scenario, given a product, it is usually collected by the close-range shots, occupying a large area in the image, and SAM's behavior appears to segment the main part of the object, such as the screw in the 3rd column and bottle in the 5th column in Figure 1(c). Furthermore, we notice that it is difficult for SAM to distinguish defective areas from the textured background. For instance, products with surface cracks (in the 1st column) and leather with a poke (in the 6th column) are challenging to identify accurately. This phenomenon is not surprising because SAM is trained on natural objects with standard sizes and high-contrast attributes. (3) Medical lesion. As illustrated in the 1st column in Figure 1(d), we observe that SAM does not handle medical data with concealed patterns well, such as benign colorectal polyps that share similar colors with the surrounding tissues. The remaining samples in Figure 1(d) are grayscale slices from three-dimensional MRI and CT scans. SAM can roughly segment the organ regions since they have distinct boundaries, but it does not perform well in recognizing amorphous lesion regions, e.g., cancer, vessels, and tumors. This suggests that SAM lacks the medical domain knowledge of these anatomical and pathological cases. To alleviate this limitation, the intrinsic relationships and semantics of anatomical structures can be injected into SAM, such as the assumption that liver tumors should be inside the liver, rather than the brain.

Discussion. From the above empirical analyses, our conclusion is: (1) We observe that SAM often segments an occluded object into multiple separated masks, indicating that its semantic capabilities in concealed scenes can be improved. (2) Unlike self-supervised large language models, SAM employs supervised training; in our experiments its emergent and reasoning abilities have not been observed. Thus, it would be interesting to try if more challenging training tasks improve its performance. (3) Considering the practical open-set problem, now the granularity and uncertainty are the bottlenecks of SAM, limiting its applications to the scenes that require high accuracy, e.g., autonomous driving and clinical diagnosis. To alleviate this issue, one potential solution is to support the model with prior knowledge. (4) SAM's great success is demonstrating the power of data-centric AI in the large model era. We see a significant trend that human feedback-based learning and large foundation models bring new opportunities for the vision community.

In summary, this work presents an empirical study for SAM. Firstly, we quantitatively evaluate SAM using cutting-edge models on the camouflaged object segmentation task. Secondly, we present several failure cases in three concealed scenarios: camouflaged animals, industrial defects, and medical lesions. We expect that this study helps the readers to comprehend SAM's performance traits in concealed scenes and brings new ideas to computer vision researchers.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2022ZD0119101).

References

1 Kirillov A, Mintun E, Ravi N, et al. Segment anything. 2023. ArXiv:2304.02643

2 Fan D-P, Ji G-P, Cheng M-M, et al. Concealed object detection. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 6024–6042

3 Le T-N, Nguyen T V, Nie Z, et al. Anabranch network for camouflaged object segmentation. Comput Vision Image Understanding, 2019, 184: 45–56

4 Fan D-P, Ji G-P, Sun G, et al. Camouflaged object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2777–2787

5 Lv Y, Zhang J, Dai Y, et al. Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 11591–11601

6 Yin B, Zhang X, Hou Q, et al. CamoFormer: masked separable attention for camouflaged object detection. 2023. ArXiv:2212.06570

7 Hu X, Fan D-P, Qin X, et al. High-resolution iterative feedback network for camouflaged object detection. In: Proceedings of AAAI Conference on Artificial Intelligence, 2023

8 Wang W, Xie E, Li X, et al. PVT v2: improved baselines with pyramid vision transformer. Comp Visual Media, 2022, 8: 415–424

9 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 10012–10022

10 Kolesnikov A, Weissenborn D, Zhai X, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of International Conference on Learning Representations, 2021