EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything

Supplementary Material

Method	Params (M)	Throughput (Images/Second)
SAM[1]	636	2
EfficientSAM-Ti (ours)	10	54
EfficientSAM-S (ours)	25	47

Table 1. Inference efficiency results. All models are prompted with one ViTDet[2] box for benchmarking the speed (throughput) of instance segmentation on a single NVIDIA A100.

In this supplementary material, we provide more results to demonstrate instance segmentation capabilities of our EfficientSAM.

1. Efficiency Evaluation

Throughput and number of parameters of our models are recorded in Tab. 1. We measure throughput (images per second) on a single NVIDIA A100 with one box prompt. The input image resolution is 1024×1024.

2. Qualitative Evaluation

To study how well our model is able to produce segmentation masks based on the prompt, we use our model to perform prompt-based instance segmentation including point-based and box-based prompt prompt segmentation. We also take our model to perform segment everything and salient instance segmentation without manually creating point and box prompt.

For each task, we share 4 examples for showing the instance segmentation capabilities of our model. These results provide direct evidence for the competing instance segmentation capabilities of our EfficientSAM with different prompts. For example, in the case of point-prompt instance segmentation, our model is able to give reasonable instance segmentation results (see Fig. 1). In the case of box-prompt instance segmentation, our model also generates expected object segmentation (see Fig. 2). In the case of segment everything, our model provides decent segmentation performance (see Fig. 3). In the case of salient instance segmentation, our model has the ability of generating mask and gives automatic instance segmentation without manually creating points or boxes (see Fig. 4). But we still need to note that our model may sometimes produce noisy segmentation, shown in Fig. 5.

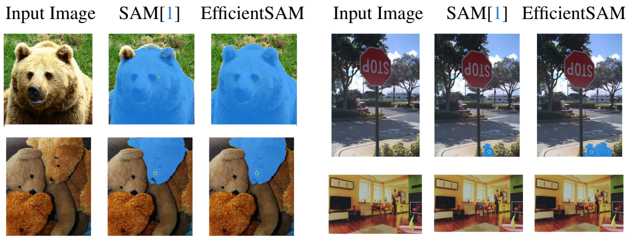

Figure 1. Visualization results on point-prompt input.

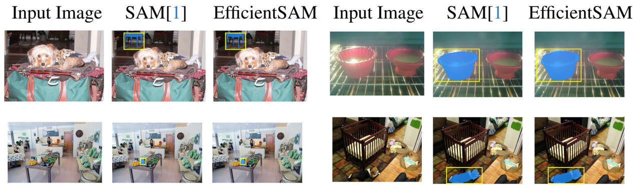

Figure 2. Visualization results on box-prompt input.

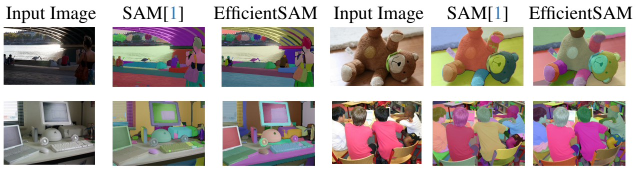

Figure 3. Visualization results on segment everything.

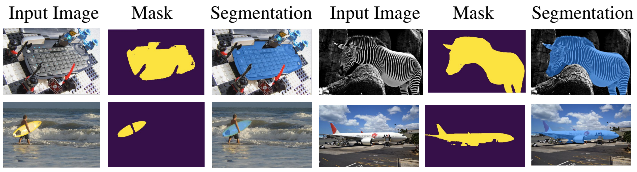

Figure 4. Saliency-based automatic instance segmentation results.

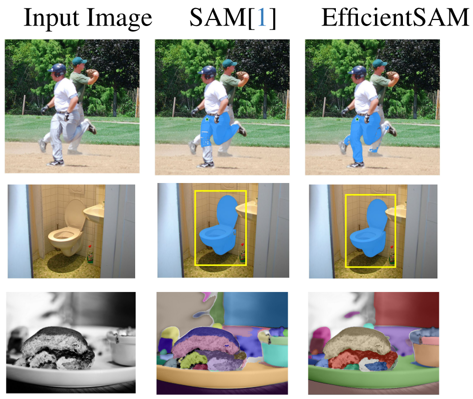

Figure 5. Segmentation results with noise. Our model may sometimes provide masks with noise, which can also be observed in the results of SAM[1].

To show the ability of processing text prompts like SAM, we also consider the task of instance segmentation with text prompts. We provide some qualitative results for EfficientSAM on the text-to-mask task by identifying the mask with the highest similarity between segmented object embedding and text embedding from CLIP encoders. In Fig. 6, EfficientSAM can segment objects well based on three different text prompts.

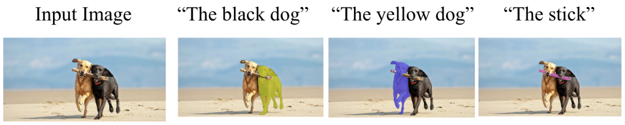

Figure 6. Segmentation results of EfficientSAM with text prompts.

References

[1] 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

[2] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022. 1