Don't waste SAM

Nermeen Abou Baker[1] and Uwe Handmann[1] *

1- Ruhr West University of Applied Sciences - Dept of Computer Science
Lutzowstrassse 5, 46236 Bottrop - Germany

Abstract. Meta AI has recently released the Segment Anything Model (SAM), which demonstrates exceptional zero-shot image segmentation performance across various tasks with remarkable accuracy. Despite its inability to provide accurate segmentation across multiple research fields, SAM still serves as a valuable starting point for supporting the segmentation pipeline process, particularly for tasks that require extensive and senior skills annotations. This study aims to evaluate the generalization of SAM and fine-tuning SAM models using three **waste** segmentation datasets. Although they are captured from real scenes as SAM was pretrained on, these datasets present several challenges, including occlusions, deformable objects, transparency, and objects easily confused with backgrounds. In our findings, the fine-tuned SAM-ViT-H model outperforms the state-of-the-art Zerowaste, and TACO datasets with a significant increase of $+30$ in IoU, and it closely approaches performance levels of TrashCan 1.0, with only a -1.44 difference. After evaluating these popular **waste** datasets, it became evident that fine-tuning SAM as a foundational model is a crucial step for providing better generalization for downstream waste segmentation tasks. Therefore, SAM should not be disregarded or **wasted**.

1 Introduction

Deep learning models benefit from training on a lot of labeled data, like ImageNet, JFT, COCO, etc., to achieve high accuracy. So, using these benchmarking datasets to start building knowledge and then refining it for specific tasks, known as transfer learning, is essential. Large Language Models (LLM) are a recent advance, that undergo training on vast textual datasets. These models have garnered considerable attention from researchers due to their impressive performance in Natural Language Processing (NLP) tasks, like Generative Pre-trained Transformer (GPT) [1]. GPT excels across a range of language tasks, sparking interest in real-world applications. ChatGPT has received special attention for commercial success [2]. This trend is influencing computer vision models too. Diffusion models, for instance, are reshaping text-image generation [3]. CLIP [4] and ALIGN [5] boosted computer vision (CV), but challenges remain unresolved. These foundational CV models, while promising, are still in their early stages and currently show performance levels that are relatively less satisfactory than achievements in NLP.

Recently, Meta AI introduced SAM, a general image segmentation model. This

---

motivates data-centric AI applications to receive substantial support from foundation models. SAM can segment "almost" any object in any image without training (zero-shot). This capability comes from training the huge SA-1B dataset, which has so far the largest CV dataset, including 11 million images and 1.1 billion masks. The huge dataset contains images taken by photographers with their cameras, and it preserves object privacy through blurring faces and licensing plates in all images [6].

In addition, In addition, SAM represents a versatile image segmentation model that generates segmentation results for foreground, background, text, etc. It operates in three distinct modes: *click mode*, where users can select objects through one or more clicks; *box mode*, in which users delineate bounding boxes around objects; and *everything mode*, which automatically masks all objects within an image. SAM consists of three components: a flexible prompt encoder that can be sparse (points, boxes, text) or dense (masks), a fast mask decoder, and an image encoder based on the vision transformer model, and it is built on the vision transformer (ViT) model. SAM offers three versions with different backbone sizes: ViT-B, ViT-L, and ViT-H, with 91M, 308M, and 636M parameters, respectively [7]. This study explores how well SAM performs in real waste segmentation images. Although they are natural images, they have many challenges like being deformable, occluded, transparent, and easy to confuse with background objects.

2 Challenges and potentials of SAM

SAM applications are vast, and some of them even extend further for "anything tasks", to name but a few; Inpaint Anything [8], anything-3D [9], any-to-any style transfer [10], and track Anything [11]. More interesting applications, e.g. [12] SAM combined with Grounding DINO and CLIP for remote sensing semantic segmentation tasks via text prompts. Another application for generating pseudo labels as a prior step to pre-train thermal infrared image segmentation tasks using SAM as a teacher in knowledge distillation approach [13]

Although SAM proves its versatility in many applications, it has many shortcomings. For example, SAM shows successful segmentation performance in agriculture applications, such as pest and leaf disease identification. However, it requires additional knowledge of crop segmentation for example because it handles crops as a background. In addition, SAM has a relatively appealing segmentation performance on normal-sized objects, but it shows some conflicts in irregular size and shaped objects, even for blurred boundaries or small disconnected components. This applies to aerial images, remote sensing, etc. that contain small-sized objects like streets, buildings, agricultural objects, etc. [14]

Medical image scenarios took the lion's share of SAM applications since the segmentation task is fundamental in medical imaging. In addition, they have privacy concerns, and labeling them by clinical experts is expensive. Besides, SAM was not trained on similar images, therefore zero-shot is desired. For example, a skin lesion segmentation task faces multiple challenges since objects

have vision obstacles in colors, contrast, and blurred boundaries [15].

SAM exhibits a strong generalization on natural scenes and has been tested in a broad range of experiments. While for task-specific image segmentation, SAM has limited generalization and requires additional knowledge to achieve high performance [16]. Even for natural images, such as scenes with shadows, SAM failed to extract these shadowed objects, because it was trained on foreground objects rather than background objects. Moreover, SAM struggles with segmenting objects with similar enclosed objects like camouflaged objects with similar foreground/background image contrast or cluttered objects. Furthermore, SAM struggles to capture objects tolerant of light conditions. This includes transparent objects that reflect light like glass, plastic, bottles, etc. [17]

Most of the previous tasks lack training instances in task-specific domains, which requires additional examples, and fine-tuning SAM could potentially deliver better performance than zero-shot transfer more effectively. On the other hand, SAM-Adaptor is proposed instead of fine-tuning SAM model [18]. It found that SAM-Adapter enhances SAM performance and even achieves state-of-the-art for the applications specific to segmenting foreground objects, like shadow and camouflaged object detection, but it only supports single-class segmentation. According to [17], to alleviate the open-set problem, prior knowledge is required to gain high accuracy. Therefore, there is a growing need to fine-tune the SAM model to downstream datasets to improve its robustness and generalizability.

3   Method and discussion

We chose the waste segmentation task as mentioned in section 1. We pretrained SAM models using three waste segmentation datasets with COCO format. This include varying datasets: For the Zerowaste dataset, we chose the full polygon annotated version that consists of the largest unbalanced deformable objects (cardboard, soft plastic, rigid plastic, and metal) with $4,661$ frames in an extremely cluttered environment that has been captured from conveyor belt [19]. TrashCan 1.0 dataset consists of $7,212$ images of underwater trash, plants, and animals. It was created for the marine robots community so they can develop efficient and autonomous trash segmentation for removal [20]. TACO dataset for waste in the wild consists of $1,500$ images and $4,784$ annotations in multiple environments such as woods, roads, and beaches [21].

We fine-tune the mask decoder only and keep the image and prompt encoders because the original SAM uses masks, points, and bounding boxes (bboxes). However, we validated only the bbox prompts derived from the ground truth masks for ease of use [22]. Practically, we used pretrained SAM weights (for ViT-L, ViT-B, and ViT-H) to initialize the weight of the mask decoder and fine-tune it in the training loop. The loss is calculated as follows: 20 * focal loss + dice loss + MSE loss.

For the tested datasets, we maintained the original dataset's training/testing split. If the split was not provided, we used an 80%/20% ratio. We imple-

mented the training process with the following hyperparameters: Batch-size=8, learning rate= 8.10^{-4}, with Adam optimizer on a single NVIDIA Quadro RTX 8000 GPU with 48 GB. We controlled overfitting using the early stop scenario when further training would not yield improvement. "The performance of mask quality was evaluated using mean Intersection over Union (IoU), which computes the overlap between ground truth and predicted masks and divides the intersection of the masks by their union. Additionally, mean precision and mean F1 pixel accuracy were calculated to determine the percentage of correctly classified pixels, as shown in Table 1.

As mentioned in section 2, the waste segmentation task has many challenges. The work of [19] shows that popular detection and segmentation models like Mask-RCNN, TridentNet, and DeepLabV3+(where the latter was the best) have unsatisfactory generalization to the Zerowaste-f dataset. In addition, for this dataset in the original SAM paper [7], SAM has +9.1 mean IoU compared to RITM [23], which is also a strong interactive segmenter. Our fine-tuned SAM-ViT-H model demonstrates a mean IoU increase of +30 compared to DeepLabv3+, and +21 compared to the original SAM.
For the TrashCan 1.0, a mean IoU decrease of -15.0 compared to RITM according to the SAM original paper [7], but the fine-tuned SAM-ViT-H has -1.44 mean IoU because SAM predictions have not matched with ground truth due to

ZeroWaste Dataset (5 Epochs)	Mean IoU	Mean Precision	Mean F1 Pixel Accuracy
DeepLabv3+ [19]	52.13	75.63	59.46
SAM-ViT-B	55.39	79.63	64.65
Fine-tuned SAM-ViT-B	79.22	87.88	86.11
SAM-ViT-L	57.22	80.13	65.46
Fine-tuned SAM-ViT-L	80.64	89.93	88.82
SAM-ViT-H	61.12	83.39	69.10
Fine-tuned SAM-ViT-H	**82.79**	**85.14**	**90.53**
TrashCan1.0 (5 Epochs)	Mean IoU	Mean Precision	Mean F1 Pixel Accuracy
DeepLabv3+	**80.32**	**86.7**	**91.25**
SAM-ViT-B	62.87	71.2	74.3
Fine-tuned SAM-ViT-B	76.91	80.81	85.04
SAM-ViT-L	63.38	71.7	75.98
Fine-tuned SAM-ViT-L	77.26	81.21	86.56
SAM-ViT-H	66.43	73.30	78.08
Fine-tuned SAM-ViT-H	78.88	82.76	87.38
TACO Dataset (10 Epochs)	Mean IoU	Mean Precision	Mean F1 Pixel Accuracy
DeepLabv3+	29.19	44.86	33.21
SAM-ViT-B	40.25	51.90	53.21
Fine-tuned SAM-ViT-B	50.40	60.41	63.20
SAM-ViT-L	41.33	53.06	55.25
Fine-tuned SAM-ViT-L	52.67	61.85	64.76
SAM-ViT-H	46.29	65.85	59.40
Fine-tuned SAM-ViT-H	**61.06**	**71.59**	**73.54**

Table 1: SAM and Fine-tuned SAM segmentation performance for the tested datasets
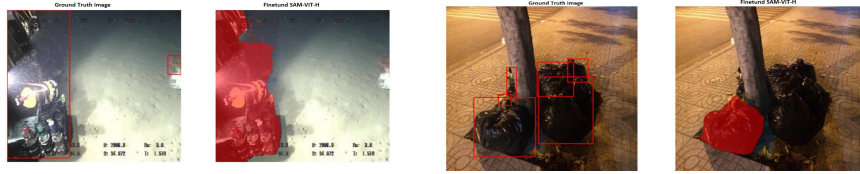
Fig. 1: Segmentation results for the pair of ground truth with original box and fine-tuned SAM-ViT-H on datasets: (left) TrashCan1.0 & (right) TACO

the ambiguity of the mask and small connections between objects, as shown in Figure 1.

The TACO dataset is very challenging because it was annotated roughly, and it has an extremely variable background, contrasts, outdoor variant lighting conditions, and object sizes, as shown in Figure 1. Therefore, SAM struggles to distinguish shadows, tiny objects, and occluded objects. Fine-tuned SAM-ViT-H provides the most superior performance compared to the other two versions because it has the highest number of parameters making it more general at capturing complex patterns. However, it requires more computational resources.

We found that SAM can be fine-tuned to create better segmentation performance for waste segmentation tasks. Therefore, fine-tuning SAM shows the ability to generalize tasks related to waste segmentation.

4   Conclusion

The study demonstrates the effectiveness and powerful impact of SAM as well as its limitations. We investigate the potential of leveraging SAM as a pre-trained model for segment-specific waste. Although the studied datasets were natural images, they are very challenging because of the variety of object sizes, orientation, lighting, background, and interconnection. We found that fine-tuned SAM-ViT-H is great potential since it approaches, or outperforms the state-of-the-art. The limitation of this study is that it evaluates the datasets that have been labeled based on bounding boxes. However, it can be alleviated using an additional prior step to draw them using detection models to locate them, then pass the dataset into the fine-tuned SAM. Eventually, this opens new possibilities for CV applications, like vision tasks analysis and understanding. On the other hand, it encourages the CV community to develop more robust foundation models. Future work aims to study more case studies and multiple datasets to support our findings.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner,

Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[2] OpenAI. Chatcpt, November 2022.

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[4] 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, 2021.

[5] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[6] MetaSA1B. Metasa1b, April 2023.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollã¡r, and Ross Girshick. Segment anything, 2023.

[8] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting, 2023.

[9] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild, 2023.

[10] Songhua Liu, Jingwen Ye, and Xinchao Wang. Any-to-any style transfer: Making picasso and da vinci collaborate, 2023.

[11] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023.

[12] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Lan Mu, Mengxuan Hu, and Sheng Li. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models, 2023.

[13] Junzhang Chen and Xiangzhi Bai. Learning to "segment anything" in thermal infrared images through knowledge distillation with a large scale dataset satir, 2023.

[14] Lv Tang, Haoke Xiao, and Bo Li. Can sam segment anything? when sam meets camouflaged object detection, 2023.

[15] Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps?, 2023.

[16] Sheng He, Rina Bao, Jingpeng Li, P. Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks, 2023.

[17] 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", 2023.

[18] 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, medical image segmentation, and more, 2023.

[19] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes, 2022.

[20] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris, 2020.

[21] Pedro F Proenca and Pedro Simoes. Taco: Trash annotations in context for litter detection, 2020.

[22] Luca Medeiros. lightning-sam, April 2023.

[23] Konstantin Sofiiuk, Ilia A. Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation, 2021.