CAN SAM COUNT ANYTHING? AN EMPIRICAL STUDY ON SAM COUNTING

Zhiheng Ma[1], *Xiaopeng Hong*[2,3], *Qinnan Shangguan*[2]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Science
[2]Harbin Institute of Technology, Harbin, P. R. China
[3]Peng Cheng Laboratory, Shenzhen, P. R. China

ABSTRACT

Meta AI recently released the Segment Anything model (SAM), which has garnered attention due to its impressive performance in class-agnostic segmenting. In this study, we explore the use of SAM for the challenging task of few-shot object counting, which involves counting objects of an unseen category by providing a few bounding boxes of examples. We compare SAM's performance with other few-shot counting methods and find that it is currently unsatisfactory without further fine-tuning, particularly for small and crowded objects. Code can be found at `https://github.com/Vision-Intelligence-and-Robots-Group/count-anything`.

Index Terms— Object counting, few-shot object counting, segment anything

1. INTRODUCTION

Large models have brought about a revolution in the field of AI, transforming it in numerous ways. In the past few months, Natural Language Processing (NLP) has undergone a significant shift towards the development of large-scale language models, resulting in some remarkable applications such as ChatGPT [1] and GPT-4 [2]. The success of large models in NLP has also inspired researchers to explore their application in the field of Computer Vision (CV).

Recently, Meta AI has released the Segment Anything model (SAM) [3], which was trained on over 1 billion masks using 11 million licensed and privacy-respecting images. Its exceptional performance in segmenting unknown images has rapidly gained attention. SAM has demonstrated its impressive capabilities in segmenting various types of images and scenes. We are therefore excited to investigate SAM's performance in few-shot counting to further explore its potential.

In this report, we compare SAM with other existing few-shot counting methods. Whether SAM can successfully count target objects depends on two aspects: first, whether SAM can segment each object. Second, whether SAM can distinguish the target objects from others using the reference examples. Instead of introducing an additional zero-shot object detector such as Grounding DINO [**?**], or an additional zero-shot classifier such as CLIP [4], we use the original image features of SAM to distinguish different objects, which greatly save the computational cost.

We provide a detailed account of our implementation as follows. Firstly, we extract the dense image feature through the use of the image encoder (ViT-H) of SAM for a given image. Secondly, we utilize the given bounding boxes as prompts to generate segment masks of the reference examples. These masks are then multiplied with the dense image feature and subsequently averaged to produce the feature vectors of the reference objects. Thirdly, we use the point grids (32 points on each side) as prompts to segment everything, and the output masks are multiplied with the dense image feature before being averaged to generate feature vectors of all masks. Finally, we compute the cosine similarity between the feature vectors of the predicted masks and the reference examples. If the cosine similarity exceeds the predetermined threshold, we consider it as the targeted object. The total count can then be obtained by counting all the target objects.

2. EXPERIMENTS

2.1. Datesets

FSC-147 [5] comprises a total of 6,135 images that have been collected for the purpose of few-shot counting. For each of these images, three object instances are randomly selected and annotated using bounding boxes, while the remaining instances are annotated using points. The training set consists of 89 object categories with 3,659 images. Meanwhile, the validation and testing sets each consist of 29 categories, with 1,286 and 1,190 images respectively.

MS-COCO [6] is a large dataset that is commonly used in object detection and instance segmentation. The val2017 set contains 5,000 images of 80 object categories that are present in complex everyday scenes. We adopt the same approach as in [7] and generate four train/test splits, each of which contains 60 training categories and 20 testing categories. It is noteworthy that while other methods utilize the training set for fine-tuning, SAM refrains from doing so.

Methods	Fold 0		Fold 1		Fold 2		Fold 3		Average	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Segment [7]†	2.91	4.20	2.47	3.67	2.64	3.79	2.82	4.09	2.71	3.94
GMN [8]†	2.97	4.02	3.39	4.56	3.00	3.94	3.30	4.40	3.17	4.23
CFOCNet [9]†	2.24	**3.50**	1.78	2.90	2.66	3.82	2.16	3.27	2.21	3.37
FamNet [5]	2.34	3.78	1.41	2.85	2.40	2.75	2.27	3.66	2.11	3.26
CFOCNet [9]	2.23	4.04	1.62	2.72	1.83	3.02	2.13	3.03	1.95	3.20
LaoNet [10]	**2.20**	3.78	**1.32**	**2.66**	**1.58**	**2.19**	**1.84**	**2.90**	**1.73**	**2.93**
SAM	5.27	7.67	3.85	8.51	3.13	8.09	3.23	7.87	3.87	8.03

Table 1. The results for each of the four folds of COCO val2017 are presented above.Methods marked with † follow the experimental settings in [9].

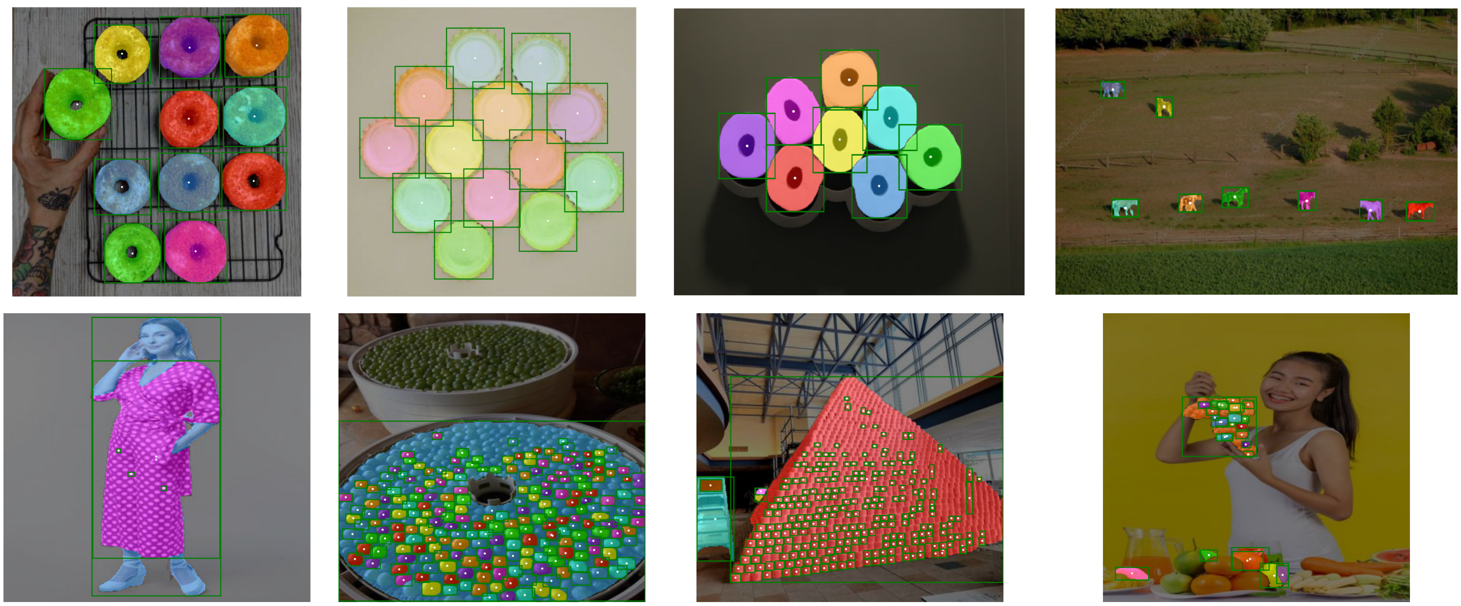

Fig. 1. Visualizations of the resulting mask obtained through SAM segmentation and counting are presented. The counts of the four images in the first row are exactly accurate, while the results of the four images in the second row are suboptimal. All the examples are from the FSC-147 dataset.

2.2. Comparison with Few-Shot Approaches

SAM's performance was evaluated through experiments conducted on the two few-shot counting datasets mentioned earlier. Table 1 presents the results of a four-fold cross-validation performed on the COCO val2017 dataset. Methods marked with † in the upper part of the table adhere to the experimental setup outlined in [9]. While SAM's performance is comparatively inferior to that of other few-shot methods, we observe that the performance gap is *not significant*, with an about 2-unit gap in average MAE, owing to the fact that COCO contains fewer small and congested objects.

Table 2 presents the performance of SAM on the FSC-147 datasets. While SAM's performance surpasses that of some earlier methods, it is *significantly inferior* to that of more recent methods, with a gap of over 10 in MAE. We conducted a visualization study on the FSC dataset, and the results are presented in Figure 1. The first row of images in the figure depicts count predictions that are exactly accurate, and we observe that the items to be counted in these images are relatively sparse. Conversely, the second row of images represents some of the bad examples, where the predicted count and ground truth count differ significantly. The first three images in this row have one thing in common - small and crowded objects are predicted as a single object during prediction. In the fourth image, SAM erroneously predicts other categories of fruit as grapes.

Methods	Val		Test	
	MAE	RMSE	MAE	RMSE
3-shot				
Mean	53.38	124.53	47.55	147.67
Median	48.68	129.70	47.73	152.46
FR detector [11]	45.45	112.53	41.64	141.04
FSOD detector [12]	36.36	115.00	32.53	140.65
GMN [8]	29.66	89.81	26.52	124.57
MAML [13]	25.54	79.44	24.90	112.68
FamNet [5]	23.75	69.07	22.08	99.54
CFOCNet [9]	21.19	61.41	22.10	112.71
BMNet+ [14]	15.74	58.53	14.62	91.83
SAFECount [15]	15.28	47.20	14.32	85.54
HMFENet [16]	**13.10**	**44.90**	**12.74**	**84.63**
SAM	31.20	100.83	27.97	131.24
1-shot				
CFOCNet [9]	27.82	71.99	28.60	123.96
FamNet [5]	26.55	77.01	26.76	110.95
LaoNet [10]	17.11	56.81	15.78	97.15
SAM	36.68	116.75	33.53	142.28

Table 2. The table displays a comparison between our few-shot method and previous state-of-the-art methods on FSC-147 dataset. The top part of the table shows the results for the 3-shot setting, while the bottom part shows the results for the 1-shot setting.

3. CONCLUSION

Although the Segment Anything model (SAM) has shown impressive performance in many scenarios, it currently lags behind state-of-the-art few-shot counting methods, especially for small and congested objects. We believe that this is due to two main reasons. Firstly, SAM tends to segment congested objects of the same category with a single mask. Secondly, SAM is trained with masks that lack semantic class annotations, which could hinder its ability to differentiate between different objects. Nevertheless, further exploration of adapting SAM to the object counting task is still worth studying.

Acknowledgements

This work is funded by the National Natural Science Foundation of China (62206271, 62076195), the Guangdong Basic and Applied Basic Research Foundation (2020B1515130004), and the Fundamental Research Funds for the Central Universities (AUGA5710011522).

4. 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, "Gpt-4 technical report," 2023.

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

[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] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai, "Learning to count everything," in *CVPR*, 2021.

[6] 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 *ECCV*. Springer, 2014.

[7] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S Ecker, "One-shot instance segmentation," *arXiv preprint*, 2018.

[8] Erika Lu, Weidi Xie, and Andrew Zisserman, "Class-agnostic counting," in *ACCV*, 2018.

[9] Shuo-Diao Yang, Hung-Ting Su, Winston H Hsu, and Wen-Chin Chen, "Class-agnostic few-shot object counting," in *WACV*, 2021.

[10] Hui Lin, Xiaopeng Hong, and Yabin Wang, "Object counting: You only need to look at one," 2021.

[11] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell, "Few-shot object detection via feature reweighting," in *ICCV*, 2019.

[12] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *CVPR*, 2020.

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.

[14] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao, "Represent, compare, and learn: A similarity-aware framework for class-agnostic counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 9529–9538.

[15] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le, "Few-shot object counting with similarity-aware feature enhancement," in *Proceedings of the IEEE/CVF Winter Conference on Applications*

of Computer Vision (WACV), January 2023, pp. 6315–6324.

[16] Zhiquan He, Donghong Zheng, and Hengyou Wang, "Accurate few-shot object counting with hough matching feature enhancement," *Frontiers in Computational Neuroscience*, vol. 17, 2023.