SCIENCE CHINA
Information Sciences

• LETTER •

April 2024, Vol. 67, Iss. 4, 149101:1–149101:2
https://doi.org/10.1007/s11432-023-3943-6

SAM3D: zero-shot 3D object detection via the segment anything model

Dingyuan ZHANG[1], Dingkang LIANG[1], Hongcheng YANG[1], Zhikang ZOU[3], Xiaoqing YE[3], Zhe LIU[2] & Xiang BAI[1*]

[1]*School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan* 430074*, China;*
[2]*School of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan* 430074*, China;*
[3]*Baidu Inc., Beijing* 100085*, China*

Received 16 August 2023/Accepted 18 November 2023/Published online 25 March 2024

In the past few years, foundation models have thrived and succeeded in linguistic and visual tasks, showing astonishing zero-shot and few-shot capabilities. Their advances encourage researchers and industries to extend the boundaries of what artificial intelligence can do and have shown some fantastic products (e.g., ChatGPT [1]) with the potential to change the world.

Recently, Kirillov et al. [2] proposed a new vision foundation model for image segmentation, the segment anything model (SAM), trained on a huge dataset called SA-1B. The flexible prompting support, ambiguity awareness, and vast training data endow the SAM with powerful generalization, enabling the ability to solve downstream segmentation problems using prompt engineering. Some following studies leverage the excellent zero-shot capability of SAM to solve other 2D vision tasks (e.g., medical image processing [3] and camouflaged object segmentation [4]). Although SAM presents great power on some 2D vision tasks, whether it can be adapted to 3D vision tasks still needs to be discovered. With this inspiration, a few studies attempt to combine SAM with pre-trained 3D models to learn 3D scene representation (e.g., SA3D [5]) and single-view reconstruction (e.g., anything-3D [6]), showing promising results.

3D object detection, one of the fundamental tasks in 3D vision, has a wide range of real-world applications (e.g., autonomous driving). Although plenty of studies aim to solve this task, the zero-shot setting on 3D object detection still needs to be explored. Thus, considering the advance of SAM, it is natural to question: Can we adapt the zero-shot capability of SAM to 3D object detection?

In this study, we aim to explore the zero-shot 3D object detection with SAM [2] alone. Considering SAM is initially built for 2D images, many challenges exist when using SAM for 3D detection (please refer to the appendix for more discussion). The key insight is that we can leverage the powerful capability of SAM for 3D object detection by using the bird's eye view (BEV), which carries crucial 3D information (e.g., depths) with a 2D image-like data format. Thus, the challenges to using SAM for 3D detection can be significantly solved. With this observation, we present SAM3D, which uses SAM to segment on BEV maps and predicts objects based on the masks from its outputs.

We evaluate our method on the large-scale Waymo open dataset [7], and the results show the great potential of SAM on 3D object detection. Although this study is only an early attempt, it gives a positive signal for applying vision foundation models like SAM for 3D vision tasks, especially for 3D object detection.

Proposed method. We consider point cloud as the input of our method, which is a 3D representation and naturally sparse, while SAM is trained for 2D images with dense semantics. Our basic idea is to translate LiDAR points into a 2D image-like representation with 3D information that narrows the domain gap; thus BEV is a straightforward choice. We build the whole pipeline with SAM based on BEV, shown as Figure 1(a). Our method mainly contains five steps.

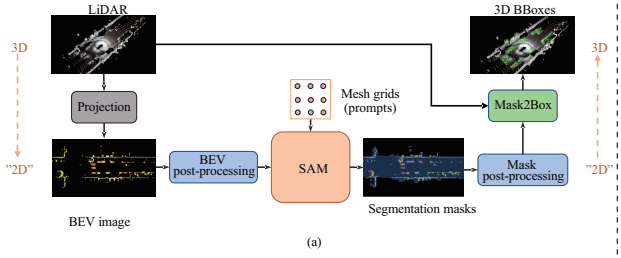
Firstly, our method conducts the LiDAR-to-BEV projection, which translates sparse LiDAR signals to discriminative BEV images. At this step, we use the projection equations to determine each point's coordinate on the image plane and a predefined intensity-to-RGB mapping to get RGB vectors for pixels in a BEV image, making it more discriminative during processing.

Then, the BEV post-processing modifies the original BEV images with the morphology dilation (interpreted as a max pooling) since SAM is trained on natural images with "dense" signals, which differs from the "sparse" BEV images. This step helps form more suitable inputs for SAM, leading to easier segmentation and better performance.

After obtaining the desired BEV images, we segment the BEV images using SAM, which supports various prompts like point, box, and mask prompts. Our goal in this step is to segment foreground objects as many as possible, so we choose to cover the whole image with mesh grid prompts. Additionally, we prune the prompts in this step without performance sacrifice to accelerate the segmentation.

Despite SAM's powerful zero-shot capability, a non-negligible domain gap still exists. Hence, we propose mask post-processing for filtering noisy masks according to some rules drawn from priors, which reduces the number of false

* Corresponding author (email: xbai@hust.edu.cn)



Version	Level 1		Level 2	
	AP	APH	AP	APH
ViT-B	17.30	11.58	16.90	11.31
ViT-L	**19.94**	13.25	**19.48**	12.93
ViT-H	19.51	**13.30**	19.05	**12.98**

(b)

Pillar size (m)	Level 1		Level 2	
	AP	APH	AP	APH
0.05	16.45	11.20	16.30	11.11
0.075	18.70	12.59	18.25	12.29
0.1	**19.51**	**13.30**	**19.05**	**12.98**
0.125	19.25	13.12	18.80	12.81
0.15	18.16	12.42	17.73	12.11
0.175	16.16	10.88	15.77	10.60
0.2	14.06	9.41	13.73	9.18
0.4	~0	~0	~0	~0

(c)

Figure 1 (Color online) (a) Overall framework of our method. We first project LiDAR points to colorful BEV images via a predefined palette, then post-process BEV images to better fit the requirements of SAM. After the segmentation, we post-process the noisy masks and finally predict 3D bounding boxes with the aid of LiDAR points. (b) Results of SAM3D using different versions of SAM. (c) Results of SAM3D using different pillar sizes. We report metrics of VEHICLE in the range [0, 30) on Waymo validation set. Best performance is in bold.

positives and helps improve the final performance.

Finally, after the segmentation and post-processing, we predict 3D bounding boxes from the foreground masks. Since BEV images already carry depth information, we can directly estimate the horizontal attributes (i.e., horizontal object center, length, width, and heading) of 3D bounding boxes from the 2D masks. Meanwhile, for the vertical attributes (i.e., vertical object center and height), LiDAR points will be utilized as extra information compensation.

Please refer to the appendix for more detailed methods.

Experiments. We evaluate our method on the Waymo Open Dataset [7], one of the large-scale datasets for autonomous driving. The dataset is split into 798 training sequences, 202 validation sequences, and 150 testing sequences. Since our method performs zero-shot object detection, we only focus on the validation sequences. For the metrics, because of the natural sparsity of point clouds and the lack of semantic label outputs, we only care about the mAP and mAPH of VEHICLE with a distance of at most 30 m in this study.

Since SAM uses different backbones with different complexities, we conduct experiments to evaluate the effectiveness of our method, shown in Figure 1(b). It reveals that using SAM with less capacity performs worse. However, there is only a marginal difference between SAM with ViT-L and ViT-H. We argue that the model capacity is not the performance bottleneck when using large models, and the power of SAM still needs to be fully unleashed. For insurance purposes, we use SAM with ViT-H. We also conduct experiments to determine how the pillar size influences the performance in Figure 1(c). When using larger pillar sizes such as 0.2 and 0.4 m, the discretization errors are relatively large, and it is hard to distinguish different objects when they are close to each other. However, pillar sizes that are too small also harm performance. One possible reason is that due to the high resolution of the small pillar size and the sparsity of LiDAR signals, it is difficult for individual instances to form a completely connected region. SAM tends to separate one object into many parts. We set the pillar size as 0.1 m, which is a good balance. Please refer to the appendix for all detailed results.

Conclusion. This study explores the zero-shot 3D object detection with the visual foundation model SAM and proposes the SAM3D. To narrow the gap between the training data of SAM and 3D LiDAR signals, we use the BEV images to represent 3D outdoor scenes. We propose an SAM-powered BEV processing pipeline to utilize the great zero-shot capability of SAM for zero-shot 3D object detection. Qualitative and ablation experiments on the Waymo Open Dataset show promising results for adapting the zero-shot ability of SAM to 3D object detection. Although this study is only an early attempt, we believe it presents a possibility and opportunity to unleash the power of foundation models like SAM on 3D tasks with technologies like few-shot learning, model distillation, and prompt engineering in the future. The code has been released in https://github.com/DYZhang09/SAM3D.

Acknowledgements This work was supported in part by National Science Fund for Distinguished Young Scholars of China (Grant No. 62225603), Hubei Key R&D Program (Grant No. 2022BAA078), and "Qisun Ye" Science Fund (Grant No. U2341227).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 1877–1901
2 Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: Proceedings of IEEE International Conference on Computer Vision, 2023
3 Zhou T, Zhang Y, Zhou Y, et al. Can SAM segment polyps? 2023. ArXiv:2304.07583
4 Tang L, Xiao H, Li B. Can SAM segment anything? When SAM meets camouflaged object detection. 2023. ArXiv:2304.04709
5 Cen J, Zhou Z, Fang J, et al. Segment anything in 3D with NeRFs. 2023. ArXiv:2304.12308
6 Shen Q, Yang X, Wang X. Anything-3D: towards single-view anything reconstruction in the wild. 2023. ArXiv:2304.10261
7 Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2020. 2446–2454