SAD: Segment Any RGBD

Jun Cen[1,3] **Yizheng Wu**[2,3] **Kewei Wang**[2,3] **Xingyi Li**[2,3] **Jingkang Yang**[3]
Yixuan Pei[4] **Lingdong Kong**[5] **Ziwei Liu**[3✉] **Qifeng Chen**[1✉]

[1]The Hong Kong University of Science and Technology
[2]Huazhong University of Science and Technology [3]Nanyang Technological University
[4]Xi'an Jiaotong University [5]National University Singapore
jcenaa@connect.ust.hk, {yzwu21, wangkewei, xingyi_li}@hust.edu.cn,
jingkang001@e.ntu.edu.sg, peiyixuan@stu.xjtu.edu.cn, lingdong@comp.nus.edu.sg,
ziwei.liu@ntu.edu.sg, cqf@ust.hk

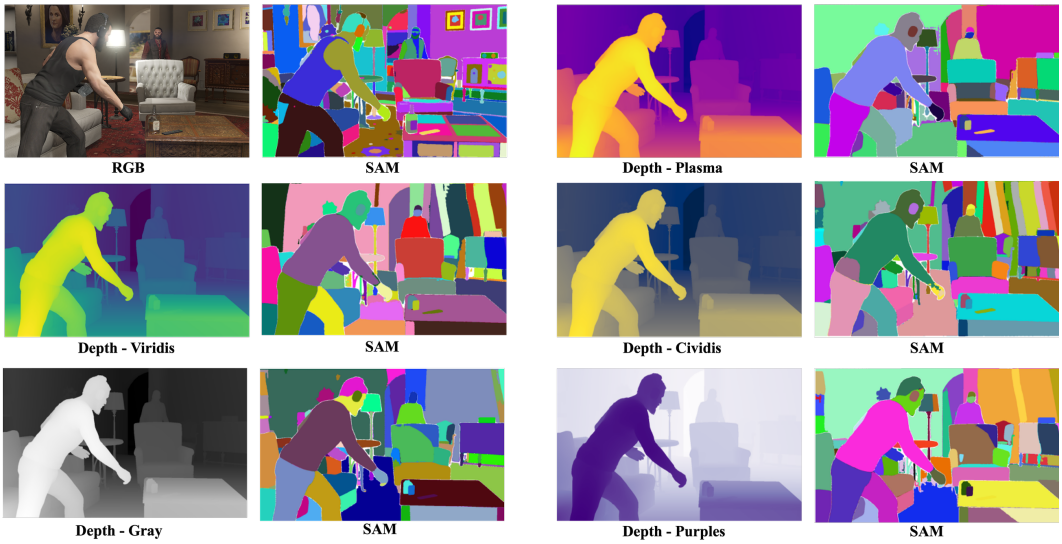

Figure 1: **Segmentation results of depth map by SAM.** In contrast to RGB images, segmentation results derived from depth maps inherently encompass a richer set of geometric information.

Abstract

The Segment Anything Model (SAM) has demonstrated its effectiveness in segmenting any part of 2D RGB images. A lot of SAM-based applications have shown amazing performance. However, SAM exhibits a stronger emphasis on texture information while paying less attention to geometry information when segmenting RGB images. To address this limitation, we propose the Segment Any RGBD (SAD) model, which is specifically designed to extract geometry information directly from images. Inspired by the natural ability of humans to identify objects through the visualization of depth maps, SAD utilizes SAM to segment the rendered depth map, thus providing cues with enhanced geometry information and mitigating the issue of over-segmentation. Compared to other SAM-based projects, we are the first to use SAM to segment non-RGB images. We further include the open-vocabulary semantic segmentation in our framework to provide the semantic labels of each segment.

37th R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023(NeurIPS 2023).

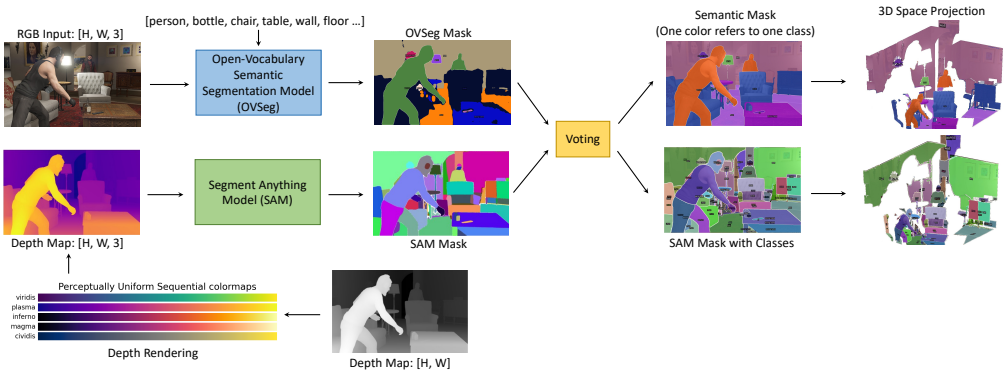

Figure 2: **The overview of the proposed SAD.** The original depth map is projected to the RGB space by a colormap function, so that it could be segmented by the Segment Anything Model. We also use open-vocabulary semantic segmentation to provide the semantic labels for each pixel. Then we conduct the voting process to determine the final semantic label of each SAM mask.

1 Introduction

Recently, a prominent model for image segmentation called the Segment Anything model (SAM) has been introduced [1]. SAM serves as a robust foundation model for effectively segmenting 2D images in various scenarios. However, during the segmentation of 2D RGB images, it has been observed that SAM primarily relies on texture information, such as color, leading to over-segmentation results. Consequently, the challenge lies in finding a way to obtain segmentation results that incorporate more geometric information through the utilization of SAM.

To address this issue, we draw inspiration from the remarkable ability of humans to identify objects by visualizing depth maps. We first map a depth map ($\mathbb{R}^{H\times W}$) to the RGB space ($\mathbb{R}^{H\times W\times 3}$) by a colormap function and then feed the rendered depth image into SAM. Compared to RGB images, the rendered depth image ignores the texture information and focuses on the geometry information, as shown in Fig. 1. Notably, it is worth mentioning that while previous SAM-based projects [2] like SSA [3], Anything-3D [4], and SAM 3D [5] primarily employ RGB images as inputs, we are the first to utilize SAM for directly segmenting rendered depth images.

2 Related Work

Recently, designing a general class of foundation models, which are trained on broad data that can be adapted to various downstream tasks, is causing a revolution within the community. The Segment Anything Model (SAM) [1] has achieved remarkable strides in pushing the boundaries of segmentation, profoundly advancing the development of foundational models for computer vision. SAM is a developed large Vision Transformer (ViT [6])-based model trained over 1 billion masks on 11 million images that enables powerful zero-shot generalization on diverse styles of 2D images. A large number of extended works have been proposed to explore the limits of SAM's capabilities and its application to diverse tasks, including image editing [7], style transfer [8], real-world detection and segmentation [9, 10, 11, 12], segmentation in complex scenes [13, 14, 15, 16], medical image analysis [17, 18, 19, 20, 21], video object tracking [22, 23], data annotations [24, 16, 25], video text spotting [24], 3D segmentation and reconstruction [26, 27, 28], image captioning [29], text-based segmentation [30], audio-visual localization and segmentation [31]. Several concurrent works, such as SegGPT [32], SEEM [33], X-Decoder [34], and OpenSeeD [35], have also delved into the potential of vision foundation models for tackling various visual tasks.

Figure 3: **Segmentation results with the RGB image input and the rendered depth image inputs.**

Method	FS	ZS	Input	mIoU	person	car	motorcycle	truck	bird	bottle	cup	bowl	chair	potted plant	dining table	TV	laptop	cell phone	bin	box	door	road barrier	other object
DKNet [36]	✓		PC	13.68	93.41	51.06	0	0	0	5.28	1.95	0	22.29	80.67	0.02	0	0	0	0	0.02	2.31	0	2.97
SAD		✓	RGB	37.90	91.96	56.67	42.93	46.63	39.52	36.60	19.26	19.85	51.01	15.53	10.41	59.73	69.35	37.84	26.93	41.10	54.40	0.44	0
SAD		✓	D	34.35	91.02	43.61	36.20	41.55	40.90	42.84	39.88	0	47.73	14.86	9.04	56.45	63.89	17.86	20.60	34.67	51.50	0.13	0

Table 1: **Quantitative Results on Sailvos3D.** FS and ZS denote Fully-supervised and Zero-shot, respectively. PC means point cloud.

3 Method

3.1 Preliminaries

Segment Anything Model (SAM). The Segment Anything model (SAM) [1] is a recently developed large Vision Transformer (ViT)-based model. SAM has been trained on an extensive visual corpus known as SA-1B. The training process on this large-scale dataset has endowed SAM with the remarkable ability to perform zero-shot segmentation on diverse styles of 2D images. All SAM-related methods consider the RGB images as the inputs, while we are the first to use SAM to segment depth maps.

Open-Vocabulary Semantic Segmentation (OVSeg). Given the class candidates in the text format, Open-Vocabulary Semantic Segmentation (OVSeg) [37] can segment an image into semantic regions even if the categories are not seen during training. We use OVSeg as the off-the-shelf model to provide the semantic labels for each pixel.

3.2 Segment Any RGBD

In this paper, we propose Segment Any RGBD (SAD) that leverages both SAM and OVSeg to achieve semantic segmentation results utilizing the geometry information derived from depth maps. The overview of SAD is shown in Fig. 2. The process can be divided into the following parts:

Rendering depth maps. We notice that depth maps tend to emphasize geometry information over texture information when compared to RGB images, as visually depicted in Fig. 1. Capitalizing on this characteristic, our approach involves initially utilizing a colormap function [38] to render the depth maps to the RGB space. We try different colormaps such as Viridis, Gray, Plasma, Cividis, and Purples, as shown in Fig. 1. Consequently, the rendered depth maps are employed as inputs for SAM.

Segmentation with SAM. Following the rendering process, we apply SAM to the rendered depth images to generate initial SAM masks. It is worth noting that these initial SAM masks are class-agnostic and still over-segmented, as illustrated in the Fig. 1. We find that the depth images rendered by different colormaps have different SAM outputs, indicating that SAM has different color preferences.

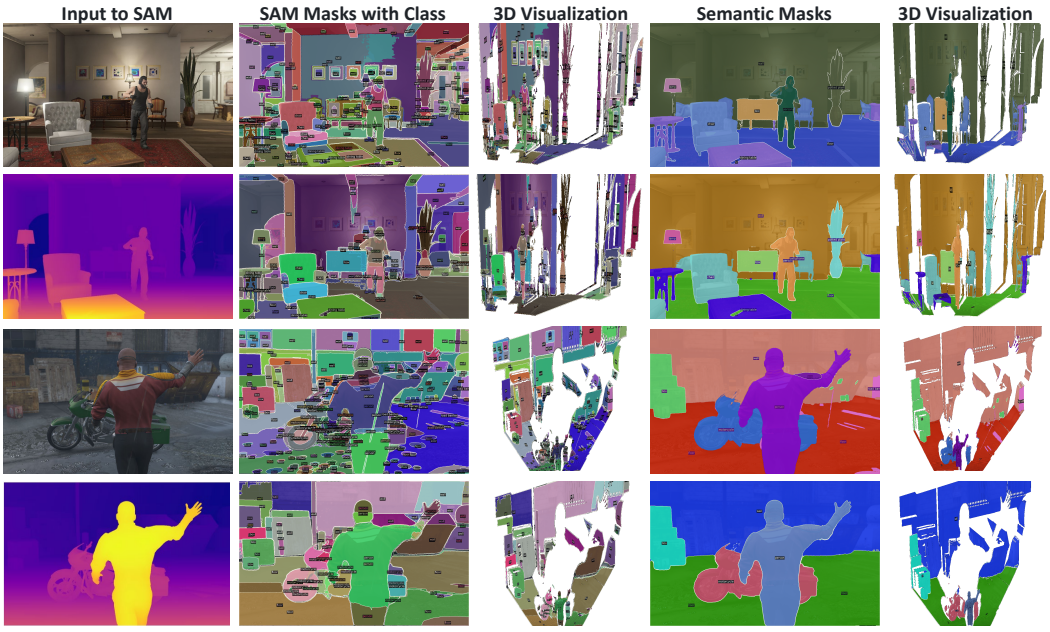

Figure 4: **Qualitative results on the Sailvos3D dataset.**

Method	FS	ZS	Input	mIoU	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	frige	s. curtain	toilet	sink	bathtub	other
U-Net [39]	✓		RGB	48.9	73.6	84.5	50.3	62.6	23.5	55.3	38.5	41.6	56.6	26.6	32.1	37.9	49.8	52.4	54.0	12.8	81.6	60.8	55.3	28.6
BPNet [40]	✓		RGB+PC	67.0	83.1	93.1	65.9	79.5	48.1	77.0	57.4	56.7	67.5	83.6	39.0	45.1	65.6	76.9	70.0	53.4	57.4	68.9	82.2	39.5
DKNet [36]	✓		PC	61.3	76.0	95.1	49.9	62.3	86.4	75.5	65.6	48.4	56.0	66.8	25.0	52.6	51.1	63.1	28.5	61.3	87.5	62.6	84.8	26.5
SAD		✓	RGB	33.3	53.1	56.6	26.5	58.9	47.9	45.5	17.8	36.2	32.6	40.2	17.3	10.5	19.6	33.8	57.2	12.7	34.6	26.5	32.3	6.8
SAD		✓	D	28.8	44.8	42.1	24.9	56.9	47.6	44.4	17.4	28.0	25.5	34.1	10.6	7.5	18.6	24.1	54.7	11.5	31.9	22.3	21.9	6.6

Table 2: **Quantitative Results on ScanNetV2.** FS and ZS denote Fully-supervised and Zero-shot, respectively. PC means point cloud.

Semantic segmentation with OVSeg. By employing RGB images as input and leveraging text prompts, OVSeg exhibits the ability to generate coarse masks that encompass significant semantic information. These coarse semantic segmentation masks serve a dual role: firstly, they assist in guiding the clustering process of the over-segmented parts within the SAM masks, and secondly, they provide crucial category insights that contribute to refining the fine-grained SAM results.

Semantic voting. For each pixel in the SAM mask, we first find its corresponding predicted class from the OVSeg mask. Subsequently, we assign the class of each segment based on the majority class of pixels contained within it. Following this, we can proceed to cluster adjacent segments that belong to the same class.

3D Projection. Finally, the semantic segmentation results can be projected to 3D-world based on the depth map for stereoscopic visualizations. This projection enables a comprehensive understanding and visual representation of the segmented results in their spatial context.

4 Comparison with RGB Image Input

We compare the proposed rendered depth image input with the RGB image input. RGB images predominantly capture texture information, while depth images primarily contain geometry information. As a result, RGB images tend to be more vibrant and colorful compared to the rendered depth images. Consequently, SAM produces a larger number of masks for RGB inputs compared to depth inputs, as illustrated in Fig. 3. The utilization of rendered depth images mitigates the issue of over-segmentation in SAM. For instance, when examining the table, the RGB images segment it into four distinct parts, with one part being misclassified as a chair in the semantic results (indicated by yellow circles in Fig.3), while it is accurately classified in the depth image. It is also important to note that when

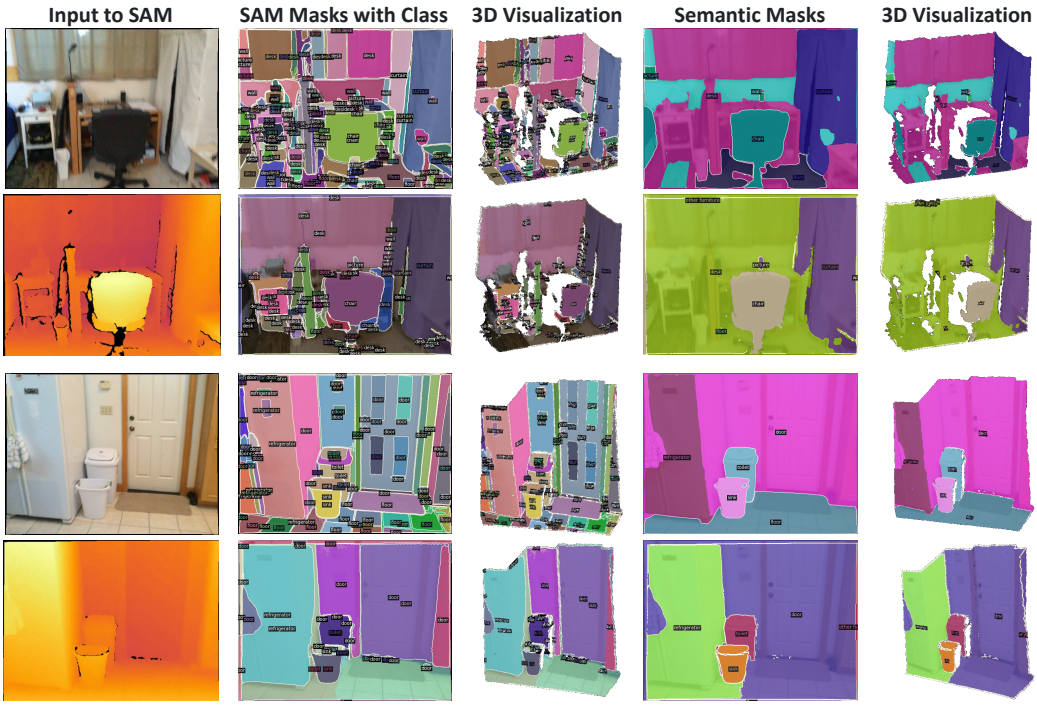

Figure 5: **Qualitative results on the ScannetV2 dataset.**

two objects are in close proximity, they may be segmented as a single object in the depth image, as depicted by the chair in the red circle of Fig. 3. In such cases, the texture information present in RGB images becomes crucial for accurately identifying and distinguishing the objects.

5 Quantitative Results

We conduct the experiments on Sailvos3D [41] and ScanNet [42] datasets. The corresponding quantitative results are shown in Table 1 and Table 2, respectively. It is reasonable to see that the overall performance of RGB inputs is slightly better than the depth inputs since the SAM is mainly trained on RGB images. However, the mIoU of SAD using the depth input is still acceptable, and some classes in Sailvos3D even have higher IoU than the RGB inputs, such as bird, bottle, and cup. Most importantly, we show that SAM has the potential to segment other modality data like depth maps rather than just segment RGB images. The overall IoU of SAD is not as good as other fully-supervised methods, which is reasonable since SAD is a zero-shot method and does not train the models at all. So there is still much room to improve based on our work.

6 Qualitative Results

We present qualitative results on Sailvos3D [41] and ScanNet [42], which are depicted in Fig. 4 and Fig. 5. The figures clearly demonstrate the enhanced performance of our method in generating geometric semantic segmentation results when utilizing depth map inputs. It is clear that the SAM results of depth maps have fewer segments than the RGB inputs, indicating that the over-segment problem of SAM is alleviated to some extent. Besides, the segmentation results of depth maps are highly influenced by the quality of the depth maps. For instance, the depth maps of Scannet might miss some depth values, as shown in the black areas in Fig. 5, which may cause the segment results not accurate.

7 Conclusion

In summary, we introduce the Segment Any RGBD (SAD) model, which combines SAM and OVSeg for semantic segmentation using depth maps. SAD leverages the geometry information of depth maps by rendering them with colormap projection and feeding them into SAM. Initial SAM masks are generated, which are refined using OVSeg's coarse semantic segmentation masks. The clustering process is then applied to group adjacent segments of the same class, improving the segmentation results' coherence. Finally, the semantic segmentation results are projected into the 3D world based on the depth map, enabling comprehensive stereoscopic visualization. Overall, SAD enhances semantic segmentation by incorporating depth maps and leveraging both SAM and OVSeg, resulting in more accurate and context-aware results. Compared to other SAM-based projects, we are the first to try SAM on other modality images that cannot be fed into the SAM directly because of the shape inconsistency. This work opens up new possibilities for SAM on other modality images and provides valuable insights into real-world applications.

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.

[2] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. A comprehensive survey on segment anything model for vision and beyond. *arXiv preprint arXiv:2305.08196*, 2023.

[3] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. `https://github.com/fudan-zvg/Semantic-Segment-Anything`, 2023.

[4] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023.

[5] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. `https://github.com/Pointcept/Pointcept`, 2023.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

[8] Songhua Liu, Jingwen Ye, and Xinchao Wang. Any-to-any style transfer: Making picasso and da vinci collaborate. *arXiv e-prints*, pages arXiv–2304, 2023.

[9] Wei Ji, Jingjing Li, Qi Bi, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*, 2023.

[10] Mohsen Ahmadi, Ahmad Gholizadeh Lonbar, Abbas Sharifi, Ali Tarlani Beris, Mohammadsadegh Nouri, and Amir Sharifzadeh Javidi. Application of segment anything model for civil infrastructure defect assessment. *arXiv preprint arXiv:2304.12600*, 2023.

[11] Iraklis Giannakis, Anshuman Bhardwaj, Lydia Sam, and Georgios Leontidis. Deep learning universal crater detection using segment anything model (sam). *arXiv preprint arXiv:2304.07764*, 2023.

[12] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.

[13] 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". *arXiv preprint arXiv:2304.06022*, 2023.

[14] Dominic Williams, Fraser MacFarlane, and Avril Britten. Leaf only sam: A segment anything pipeline for zero-shot automated leaf segmentation. *arXiv preprint arXiv:2305.09418*, 2023.

[15] Junzhang Chen and Xiangzhi Bai. Learning to" segment anything" in thermal infrared images through knowledge distillation with a large scale dataset satir. *arXiv preprint arXiv:2304.07969*, 2023.

[16] Di Wang, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Scaling-up remote sensing segmentation dataset with segment anything model. *arXiv preprint arXiv:2305.02034*, 2023.

[17] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

[18] Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps? *arXiv preprint arXiv:2304.07583*, 2023.

[19] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics*, 13(11):1947, 2023.

[20] Yichi Zhang and Rushi Jiao. How segment anything model (sam) boost medical image segmentation? *arXiv preprint arXiv:2305.03678*, 2023.

[21] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[22] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.

[23] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023.

[24] Haibin He, Jing Zhang, Mengyang Xu, Juhua Liu, Bo Du, and Dacheng Tao. Scalable mask annotation for video text spotting. *arXiv preprint arXiv:2305.01443*, 2023.

[25] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *arXiv preprint arXiv:2305.11003*, 2023.

[26] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *arXiv preprint arXiv:2306.09347*, 2023.

[27] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. *arXiv preprint arXiv:2304.12308*, 2023.

[28] Youtan Yin, Zhoujie Fu, Fan Yang, and Guosheng Lin. Or-nerf: Object removing from 3d scenes guided by multiview segmentation with neural radiance fields. *arXiv preprint arXiv:2305.10503*, 2023.

[29] Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023.

[30] 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. *arXiv preprint arXiv:2304.10597*, 2023.

[31] Shentong Mo and Yapeng Tian. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023.

[32] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[33] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023.

[34] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.

[35] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023.

[36] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 235–252. Springer, 2022.

[37] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 2023.

[38] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[40] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14373–14382, 2021.

[41] Y.-T. Hu, J. Wang, R. A. Yeh, and A. G. Schwing. SAIL-VOS 3D: A Synthetic Dataset and Baselines for Object Detection and 3D Mesh Reconstruction from Video Data. In *CVPR*, 2021.

[42] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.