

Supplementary Material for SAMPro3D

Outline

This supplementary document is arranged as follows:

- (1) Sec. **A** illustrates more qualitative results across ScanNet200 [13], our ScanNet200-Fine50, ScanNet++ [16] and KITTI-360 [8];
- (2) Sec. **B** reports the results on Matterport3D [1];
- (3) Sec. **C** discusses an augmented version of using prompt propagation across 2D frames;
- (4) Sec. **D** provides more ablation studies of our method;
- (5) Sec. **E** integrates several variants of SAM [4, 6, 18] into our framework;
- (6) Sec. **F** elaborates our method implementations and dataset constructions;
- (7) Sec. **G** introduces a supplementary evaluation scheme when lacking fine-grained GT annotations.

A. More Qualitative Results

A.1. On ScanNet200

Following the main paper, Fig. **I** presents more qualitative comparisons on the ScanNet200 validation set, where our method is compared with SAM3D [15], Mask3D [14], and the original annotations of ScanNet200 [13]. Note that Mask3D does not treat floor and wall as instances, resulting in the absence of these two labels in its results.

Consequently, our method consistently achieves remarkable 3D scene segmentation results across diverse scenes and objects, from holistic views to focused perspectives. Notably, our approach significantly outperforms SAM3D in terms of segmentation quality and diversity. When compared to Mask3D (*trained and evaluated both on ScanNet200*), our method demonstrates competitive or superior segmentation quality and diversity. Importantly, our results not only match the quality of human annotations but also exhibit greater diversity in many cases.

We provide an animated visualization of the qualitative comparison in a visually appealing video format, where we simulate the moving camera in the segmented 3D scene. You may refer to the supplementary file folder and access the video file named *qualitative_comparison.mp4*, to have a clear view of our impressive qualitative results. In this video, the first 2 minutes and 12 seconds showcase a qualitative comparison across different indoor scene areas. Specifically, the segment from the beginning to 1 minute

and 16 seconds compares our method with SAM3D, while the segment from 1 minute and 17 seconds to 2 minutes and 12 seconds compares our method with Mask3D (*trained and evaluated both on ScanNet200*). Starting from the 2 minutes and 12 seconds mark until the end, the video simulates a complete camera moving throughout an entire indoor room. From 2 minutes and 12 seconds to 2 minutes and 43 seconds, the comparison is made with SAM3D, and from 2 minutes and 44 seconds to the end, the comparison is made with Mask3D.

A.2. On ScanNet200-Fine50

We have introduced a fine-grained test set called ScanNet200-Fine50. In Fig. **II**, we provide more qualitative comparisons among the predictions from our method, SAM3D [15], and Mask3D [14], as well as the original annotations from ScanNet200 [13] and the fine-grained annotations from our ScanNet200-Fine50. Here, we mainly show *focused* views, aiming to highlight the segmentation performance specifically for fine-grained instances.

In comparison to the initial annotations in ScanNet200, our ScanNet200-Fine50 test set offers significantly more finely detailed annotations of high quality. Furthermore, our method’s predictions exhibit better alignment with ScanNet200-Fine50 when compared to SAM3D and Mask3D, demonstrating the fine-grained segmentation capability of our approach.

B. Results on Matterport3D

As mentioned in the main paper, we also applied our method to the Matterport3D dataset [1]. In this dataset, the RGB frames exhibit *larger view changes* compared to ScanNet [2] and ScanNet++ [16], which presents additional challenges when performing segmentation solely on 2D frames. We follow [11] to use undistorted images in the official Matterport3D repo * and test on 160 classes of the validation set.

As shown in Tab. **I**, our method outperforms both SAM3D [15] and SAI3D [17] on Matterport3D dataset, which is further supported by the qualitative results in Fig. **III**. This further proves the robustness of our method on novel 3D scenes.

* <https://github.com/niessner/Matterport>

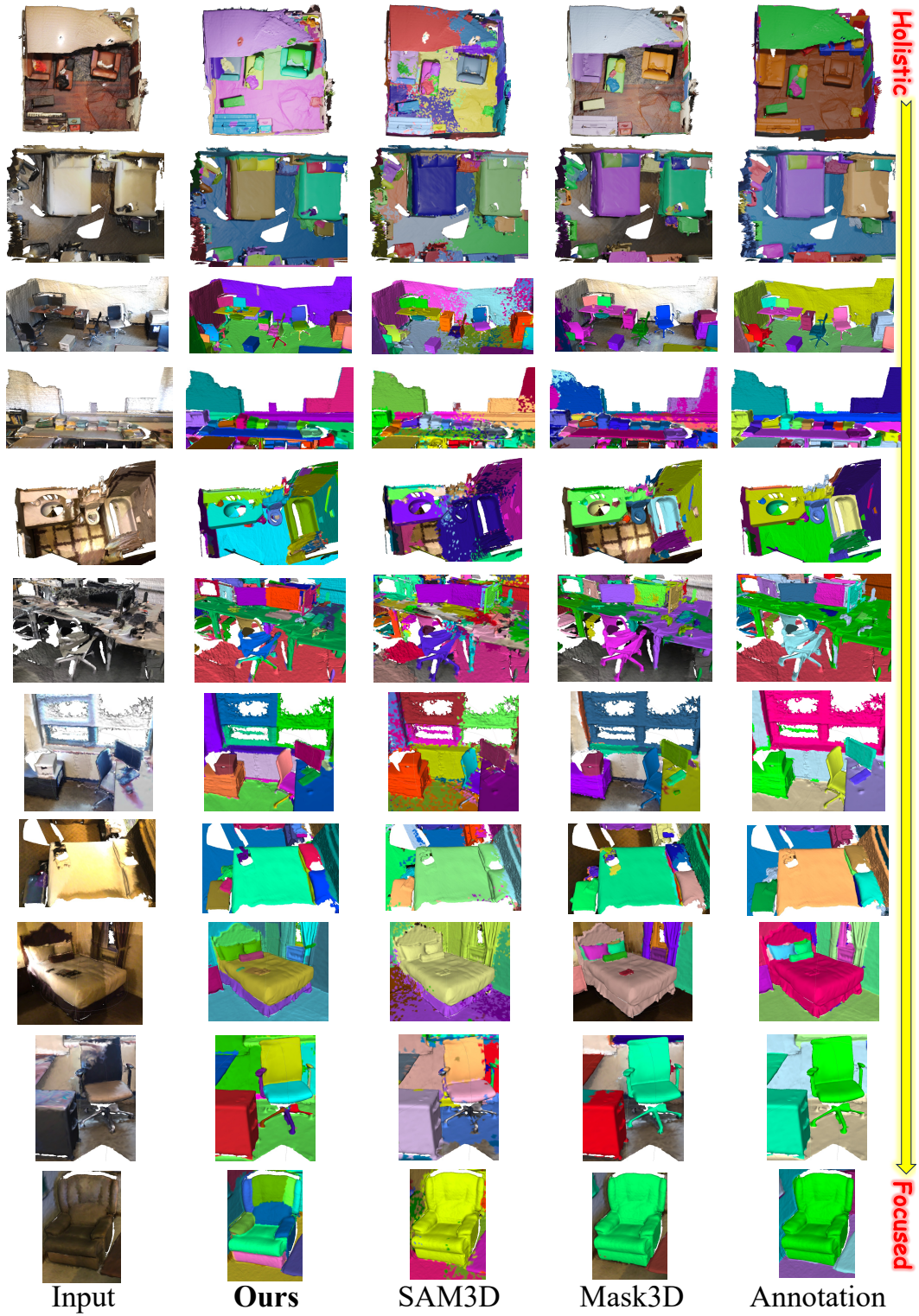


Figure I. The qualitative comparison of **our method**, SAM3D [15], Mask3D [14] and ScanNet200’s annotations [13], across various scenes in the **ScanNet200 validation set**, from holistic to focused view. Note that Mask3D does not treat floor and wall as instances, resulting in the absence of these two labels in its results.

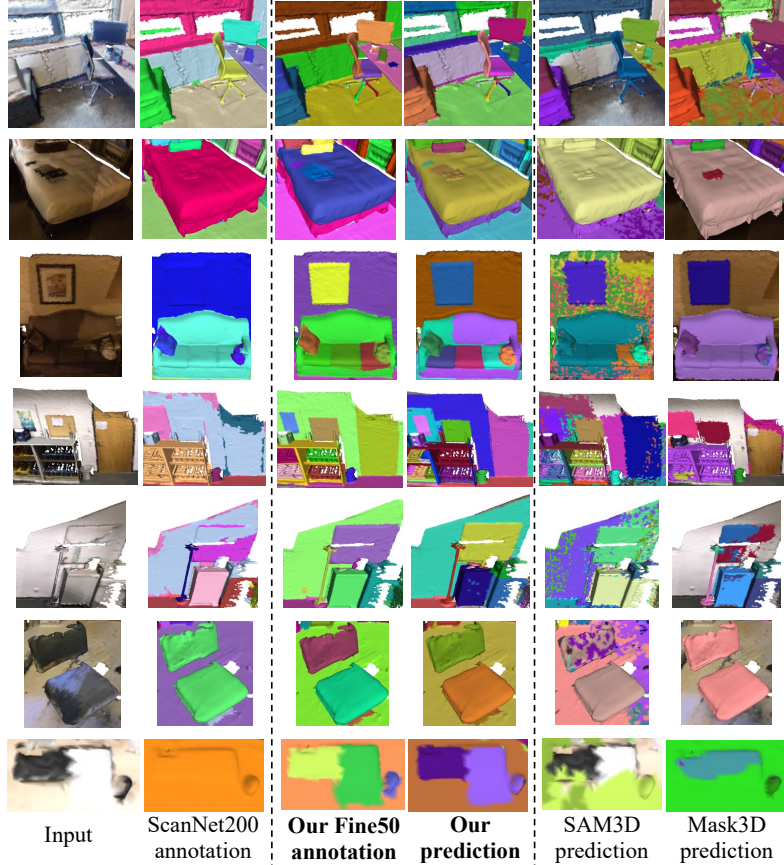


Figure II. The qualitative comparison among the predictions from our method, SAM3D [15], and Mask3D [14], as well as the original annotations from ScanNet200 [13] and the fine-grained annotations from our ScanNet200-Fine50, across diverse scenes from focused views. Note that Mask3D does not treat floor and wall as instances, resulting in the absence of these two labels in its results.

Model	AP	AP ₅₀	AP ₂₅
SAM3D [15]	10.1	19.4	36.1
SAI3D [17]	21.5	38.3	59.1
Ours	24.3	42.1	65.4

Table I. The quantitative comparison on Matterport3D [13].

Model	AP	AP ₅₀	AP ₂₅
Augmented 2D propagation	20.3	38.6	59.7
Ours	26.3	47.2	68.6
Ours+HQ-SAM [4]	28.5	47.9	69.8
Ours+Mobile-SAM [18]	20.9	40.8	61.3

Table II. The quantitative results on ScanNet200 [13]. “Augmented 2D propagation” is detailed in Sec. C. “+HQ.” and “+Mob.” respectively indicate incorporating HQ-SAM [4] and Mobile-SAM [18] in our framework.

C. An Augmented 2D Propagation

In Sec. 1 and Fig. 2 (c) of the main paper, we discuss achieving prompt consistency by using automatic-SAM [5] on an initial frame to generate pixel prompts which can be propagated to subsequent frames, similar to SAM-PT [12] for video tracking. However, prompts generated on initial frames of 3D scenes cannot cover newly emerged instances in other frames, leading to the absence of segmentation for many instances.

In this section, we evaluate an alternative scheme. Instead of performing automatic-SAM only once on an initial frame, we check if any areas lack segmentation masks in a frame, indicating the presence of newly emerged objects.

In such cases, we reapply automatic-SAM to make prompts cover these objects.

As depicted in Fig. IV, although the augmented version of 2D propagation improves the completeness of 3D segmentation results, it still falls short in terms of both segmentation quality and diversity. Its deficiency is further highlighted by the comparison of mAP scores in Tab. II. The primary reason behind this inferior performance is that the augmented scheme only aligns pixel prompts within a



Figure III. The qualitative results of our method on **Matterport3D** [1], from holistic to focused view. The results are arranged in pairs where the left is the input and the right is our output.

	Module Impact		Initial Prompts				θ_{retain}								Selection	
	w/o Sel.	w/o Con.	1%	1.5%	2%	5%	0.3	0.4	0.5	0.6	0.7	0.8			soft	top-k
Normal	43.2	42.6	46.7	47.8	48.1	44.2	40.7	47.9	48.1	47.9	47.3	42.5	47.5	47.7		
Small	25.6	25.1	29.1	30.1	30.3	26.1	25.6	29.2	30.3	30.1	29.9	25.7	30.0	29.8		
Tiny	22.9	22.5	24.3	25.3	25.6	23.7	23.2	24.1	25.6	25.4	24.9	23.0	25.2	25.0		

Table III. The quantitative ablation studies **on our ScanNet200-Fine50** test set. We report AP_{50} across different mask sizes of our GT annotations (Normal, Small, Tiny). “w/o Sel.” and “w/o Con.” respectively denote discarding prompt selection and consolidation. We also evaluate our method using different ratios (1%, 1.5%, 2%, 5%) of input points as our initial prompts. θ_{retain} is the threshold in prompt selection. “soft” and “top-k” are two voting schemes used during prompt selection.

limited range, from the frame where automatic-SAM is applied to the next time reapplying it. Consequently, the mask consistency is restricted to a few frames. In contrast, our 3D prompts globally align pixel prompts across *all* frames, resulting in comprehensive frame-consistent pixel prompts and 2D masks, as well as superior 3D segmentation results.

D. More Ablation Studies

D.1. The Ablation Results on ScanNet200-Fine50

Following the main paper, we conduct similar ablation studies on our ScanNet200-Fine50 test set and report the result

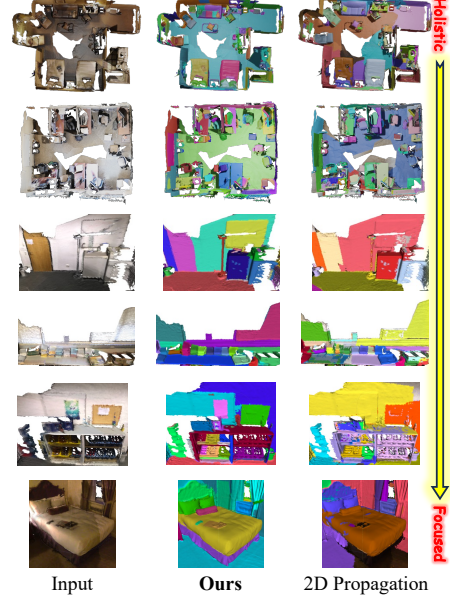


Figure IV. The qualitative comparison between **our method** and **the augmented 2D propagation**, across various scenes in the ScanNet200 [13] validation set, from holistic to focused view.

in Tab. III. Similar to the conclusions in the main paper, removing prompt selection or consolidation leads to a performance drop across different mask sizes. Besides, using 5% initial prompts or setting $\theta_{retain} = 0.3, 0.8$ results in worse performance.

However, the fine-grained segmentation results (Small, Tiny) also *slightly* degrade when using 1% initial prompts or setting $\theta_{retain} = 0.4$. One possible reason is that when selecting fewer prompts, prompts may have a lower probability of being *accurately scattered* onto fine-grained instances, as this kind of instance only occupies a small area. In this scenario, prompts may tend to be located on large-sized instances, resulting in the absence of segmentation for fine-grained instances.

D.2. Frame Gaps

In the context of performing SAM [5] on 2D image frames, an alternative approach is to skip frames with a certain gap. Fig. V illustrates the qualitative results obtained by skipping frames with different gap numbers. Fig. VI depicts the quantitative results on the ScanNet200 validation set [13], considering both segmentation AP_{50} and time cost.

The results indicate that the segmentation accuracy remains satisfactory with a gap of 5, while there is a degradation when using a gap of 10 or 20. Besides, according to Fig. V, our framework stably maintains good segmentation diversity across different gap settings. It’s also worth mentioning that our method consistently outperforms SAM3D [15] in view of both segmentation accuracy and efficiency

same soft mask. We kept the prediction (*i.e.*, binary mask resulting from thresholding logits at 0) if the IoU between its pair of -1 and +1 thresholded masks was 60.0 or higher.

Details of building ScanNet200-Fine50. To build our ScanNet200-Fine50 test set, we handpicked 50 scenes from the ScanNet200 [13] validation set. These selected scenes predominantly feature a higher number of fine-grained instances that lack mask annotations, such as multiple small instances on tables. Subsequently, we engaged the expertise of five experienced 3D data annotators, assigning each of them 10 scenes for annotation. Throughout the annotation process, they were instructed to meticulously examine each instance and provide annotations with the utmost level of detail possible. For instance, their annotations encompass not only each small instance on a table but also different removable parts of a chair. At present, our fine-grained annotations are agnostic to specific categories and do not include explicit category labels. Furthermore, the annotators will cross-check the initial annotations provided by their peers and offer feedback through an online communication system. This process ensures the meticulous and high-quality annotation of the data.

G. A Supplementary Evaluation Scheme

As mentioned in the main paper, different from previous zero-shot or fully-supervised methods, our approach preserves the zero-shot power of SAM, *often segmenting fine-grained instances that lack* corresponding accurate Ground Truth (GT) annotations (as in Fig. 1). Consequently, as illustrated in Fig. IX (“Problem”), if we directly compare our predictions with GT annotations during mAP calculation, our successfully-segmented fine-grained instances will be counted as False Positive, which hurts the accurate evaluation. A similar problem occurs in the SAM paper [5] (Tab. 5), where evaluating SAM on classical coarsely-annotated datasets [3, 9] leads to inferior mAP results compared to the fully-supervised ViTDet [7]. However, SAM outperforms ViTDet according to the user study.

To handle this issue, we add a grouping process. We begin by selecting an annotated instance g from the validation data. We then traverse our segmentation outputs to identify all predictions $\{O_m | i = m, \dots, M\}$ that *belonging* to g by checking if the most area ($> 80\%$) of O_m is included by g . We group all such predictions O_m into a single prediction which is then compared with g to decide whether it is a True Positive, as shown in Fig. IX (“Our process”). This process is repeated for all annotated instances.

We apply the proposed grouping process when calculating the AP scores of Mask3D [14] and our method. The results on the ScanNet200 validation set are listed in Tab. IV. First, it is evident that our grouping method has a *minimal* impact on Mask3D. This is primarily because Mask3D is

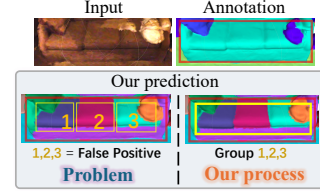


Figure IX. Evaluation of the fine-grained predictions that lack GT annotations. Predictions 1,2,3 (sofa cushions) are **visually good**. However, if we compare them with the whole annotated sofa, they will all be **False Positive** due to small IoU. Thus, we group them for better evaluation.

Method	AP		AP ₅₀		AP ₂₅	
	w/o group	w/ group	w/o group	w/ group	w/o group	w/ group
Mask3D [14]	53.3	54.0 (0.7↑)	71.9	72.7 (0.8↑)	81.6	82.2 (0.6↑)
Ours	26.3	33.7 (7.4↑)	47.2	54.3 (7.1↑)	68.6	77.2 (8.6↑)

Table IV. Quantitative comparison with Mask3D (pretrained on ScanNet) on ScanNet200. “w/o group” and “w/ group” denote scores **without and with our grouping process**. While grouping operation **improves** our scores, it has **minimal impact** on Mask3D which can always find GT annotations.

pretrained on ScanNet200 training data, which possesses similar or even finer granularity compared to the test data. Consequently, most of Mask3D’s predictions successfully match the corresponding GT annotations in the validation set (as shown in Fig. 1). In contrast, for our SAM-powered model, it is common for predictions to lack annotations, so the grouping operation leads to improved AP scores. Notably, our **AP₂₅** with grouping is even comparable with fully-supervised Mask3D. This experiment indicates that our grouping scheme not only mitigates the issue of evaluating predictions without annotations but also has minimal impact on evaluating the predictions with matched annotations, showcasing the fairness and rationality of our grouping process.

References

- [1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017. 1, 4
- [2] 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. 1
- [3] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 6
- [4] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 1, 3, 5
- [5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

- head, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023. 3, 4, 5, 6
- [6] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 1, 5
- [7] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 6
- [8] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *PAMI*, 2022. 1
- [9] 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*, 2014. 6
- [10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5
- [11] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D Scene Understanding with Open Vocabularies. In *CVPR*, 2023. 1
- [12] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023. 3
- [13] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 1, 2, 3, 4, 6
- [14] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *ICRA*, 2023. 1, 2, 3, 6
- [15] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. In *ICCVW*, 2023. 1, 2, 3, 4, 5
- [16] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 1
- [17] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *CVPR*, 2024. 1, 3
- [18] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 1, 3, 5