

# SceneSAM: Integrating 2D Labels for Weakly Supervised 3D Scene Understanding - Supplementary Material

BMVC 2024 Submission # 933

## 1 Training Parameters

In this supplementary material, we will go over the hyperparameters and their relevance to our architecture. Each category will be introduced and explained in their respective sections. We report the most important parameters to our training in in Table 1. We also separate our synthetic experiments on the Replica [1] dataset, ran on clean inputs, and our real-world experiments on the ScanNet [2] dataset.

## 2 Tracking Implementation

Starting with the tracking parameters, it is essential to recognize that this combination was used without ground truth poses provided by the corresponding datasets. While tracking, a random subset of pixels is sampled from the image. The number of pixels sampled and the amount of iterations that are done on the respective frame are correlated with the accuracy of the tracking. This setting can be tuned in a per-scene basis and per-domain basis as well, where for noisier measurements we have to associate more computation on the pose tracking module.

The *Ignore Edge* parameters, also lead to more precise tracking, since sampling near the edges of the images is unreliable and contour bleeding is common, as a reason of the motion blur, lens distortion and depth sensor resolution. As seen from the Table 1, we needed to invest more resources during tracking while training on real-world scenes. It needs to be mentioned that the ignore edge variables were set to lower numbers since we empirically found out that it yielded better results if a higher number of pixels were sampled.

## 3 Segmentation Implementation

One of our main contributions, the segmentation algorithm is based on the Segment Anything Model [3], and this part of the architecture included many design choices 1. Starting with the flexibility our architecture provides, the choice of running in parallel, meaning as a SLAM system, or running the segmentation part subsequently. This choice has no effect on the generated segmentations and the resulting outputs. Additionally, we segment only every

$N^{\text{th}}$  frame to increase runtime speed and limit redundancy in heavily overlapping camera frustrums. Additionally we introduce a simple, yet effective filtering for noise segmentations. The "Smallest Mask Size", the size of the segmented objects in pixels, was a crucial choice. The choice for allowing smaller masks results in finer segmentations but leads to relatively higher inconsistencies between the frames affecting our algorithm. As the scenes from ScanNet include motion blur and depth inconsistencies [1], stabilizing segmentations with smaller mask sizes is harder. Therefore, we opted for having higher confidence in segmentations than having finer segmentations in real-world scenes.

The "Border", how far from the edges of the image we are sampling, "Object Contour Farther", the margin between the neighbouring masks, and "Depth Condition", the depth error that we expect from each sampled point, is again set for preventing undesired behaviour during the creation of the self-consistent frames.

## 4 Other Variables

Two vital hyperparameters, concerning the mapping section of the architecture, are the frequency of the mapping, and the selection of the frequency of the keyframes. The mapping part has also an effect on how well the tracking performs, therefore, for real-world scenes we opted again for higher frequency.

In our multi-stage training strategy we first optimize for a fine geometry and color representation, and only in the final stage start to optimize for scene instances. Here during the instance stage we deactivate the learning rates for the other stages and only train for the object masks. This separation of learning tasks was shown to be successful to reduce the effect of noisy gradients stemming from inconsistencies in the video masks.

For the camera parameters, we downsized the images in synthetic scenes to be able to process the images faster, and the Segment Anything Model was trained also on similar sized images. The meshing confidence threshold filters out the instance predictions where occlusions or incomplete coverage limits the instance supervision amount.

## References

- [1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.
- [2] 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.
- [3] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The replica dataset: A digital replica of indoor spaces, 2019.

Category	Key	Synthetic Scene Values	Real-World Scene Values
Tracking	Ground Truth Camera	False	False
	Ignore Edge Width (in Pixels)	50	<b>20</b>
	Ignore Edge Height (in Pixels)	50	<b>20</b>
	Pixels	400	<b>2000</b>
	Iterations	25	<b>300</b>
	Learning Rate	0.001	0.001
Segmenter	Full Slam Configuration	True	True
	Every Frame Frequency	10	<b>2</b>
	Border (in Pixels)	10	10
	Normalize Point Number (in Px)	7	7
	Object Contour Farther (in Px)	2	2   <b>5</b>
	Depth Condition (in metres)	0.05	0.05   <b>0.1</b>
Smallest Mask Size (in Px)	600	<b>1000</b>   <b>2000</b>	
Mapping	Every Frame Frequency	5	<b>1</b>
	Keyframe Frequency	50	<b>6</b>
	Weight Instance Loss	10	10
	Instance Iteration Ratio	0.4	0.4
Instance	Decoders Learning Rate	0.005	0.005
	Coarse Learning Rate	0.0	0.0
	Middle Learning Rate	0.0	0.0
	Fine Learning Rate	0.0	0.0
	Color Learning Rate	0.0	0.0
	Instance Learning Rate	0.4	0.4
Camera	Crop size (in Pixels)	[340, 600]	-
Meshing	Confidence Threshold	0.4	0.4

Table 1: We provide full parameter list for our SceneSam paper to aid reproducibility for the community.