OmniSeg3D: Omniversal 3D Segmentation via Hierarchical Contrastive Learning

Supplementary Material

Summary. In this supplementary material, we elaborate on the following subjects. We present an ablation study on 2D foundation models for segmentation in Sec. 7. We discuss the details of the mesh-based and point-based implementations in Sec. 8 and Sec 9. We provide more results and implementation details for hierarchical segmentation in Sec. 10, and for instance segmentation in Sec. 11. For more qualitative results, please refer to our video.

7. Ablation Study on 2D Foundation Models for Segmentation

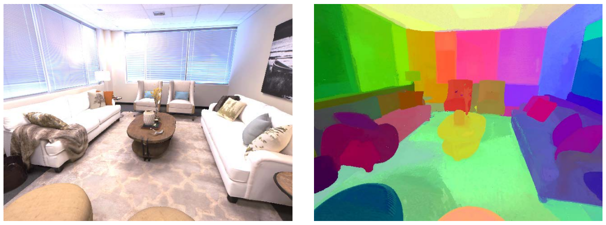
We propose the methodology of OmniSeg3D as a **general paradigm** for lifting inconsistent 2D segmentations to 3D, as opposed to a model catering to any specific 2D segmentation model. Though we use SAM [26] as the 2D foundation model in our implementation, any click-based segmentation methods can be used as an alternative. For instance, we adopt RITM [48] and SimpleClick [33] to demonstrate the generalizability of our method.

Implementation details. Similarly to SAM, we implement an automatic mask generator with the substitute backbone. We feed the image with a grid of 32×32 point prompts to the 2D segmentation model, retrieve all masks and corresponding logit maps, filter out unstable masks according to their sensitivity to the logit threshold, and filter duplicates with non-maximum suppression. With the resulting overlapping 2D masks, we follow Sec. 3.1 to build the hierarchical 2D representation and follow Sec. 3.2 to train the 3D feature field. We then follow Sec. 4.1 to benchmark these implementations.

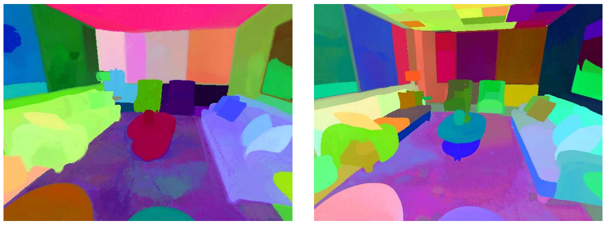
Results. Tab. 5 lists the quantitative comparisons on hierarchical and instance segmentation. Due to the fact that RITM and SimpleClick do not specialize in part-level or small object segmentation, switching to these backbones results in degraded performance on level-1 hierarchical segmentation and instance segmentation. However, the level-2 results remain comparable with the SAM-based OmniSeg3D, both outperforming vanilla SAM (see Tab. 1). Fig. 7 shows the UMAP visualizations of the learned semantic features on the `room-0` scene in the Replica dataset [49], volume rendered to a specific view. The SAM-based implementation captures the most fine-grained hierarchies within an object and demonstrates the sharpest segmentation boundaries, but all three variants achieve consistent high-level semantic clustering.

Method	Hierarchical mIoU			Instance
	Lv.1	Lv.2	Avg.	mIoU
Ours, w/ RITM [48]	77.9	90.6	84.3	74.9
Ours, w/ SimpleClick [33]	76.1	**93.6**	84.9	74.0
Ours, w/ SAM [26]	**93.6**	93.1	**93.3**	**83.0**

Table 5. Comparison of our method with different 2D foundation models, on the `room-0` scene in the Replica dataset.



(a) Image (b) Ours, w/ RITM [48]



(c) Ours, w/ SimpleClick [33] (d) Ours, w/ SAM [26]

Figure 7. Visualizations of 3D semantic features trained with alternative segmentation backbones on the Replica dataset.

8. Details on Mesh-based Implementation

Our method is not restricted by the underlying 3D representations and can be easily extended to mesh based rendering pipeline. For mesh-based representation, we implement a rasterization-based rendering pipeline based on NVDiffrast [28], in which only the points located on the mesh will be sampled for rendering and optimization. Meanwhile, the network architecture remains the same as the volume rendering pipeline in the main paper.

Automatic discretization. We show **automatic** 3D discretization results in Fig. 8. Given an optimized 3D feature field, we can distill the feature onto the mesh vertices. Then a feature clustering algorithm is implemented, sim-

Scene	SAM [26]			Ours, w/o hierar.			Ours, w/o coord.			OmniSeg3D (Ours)		
	Lv.1	Lv.2	Avg.	Lv.1	Lv.2	Avg.	Lv.1	Lv.2	Avg.	Lv.1	Lv.2	Avg.
Office 0	91.4	87.0	89.2	93.4	74.5	83.9	90.9	89.5	90.2	89.7	88.8	89.3
Office 1	94.1	75.2	84.7	92.2	81.3	86.8	86.8	88.5	87.7	91.3	90.4	90.9
Office 2	92.6	79.1	85.8	93.1	76.4	84.7	92.8	82.9	87.9	93.0	87.0	90.0
Office 3	93.6	73.9	83.8	94.8	75.1	85.0	94.2	84.4	89.3	94.2	86.0	90.1
Office 4	90.7	82.5	86.6	91.7	81.4	86.5	88.2	86.4	87.3	87.2	90.1	88.7
Room 0	95.8	86.7	91.2	95.9	87.8	91.8	93.8	92.3	93.1	93.6	93.1	93.3
Room 1	93.3	75.9	84.6	93.1	85.8	89.4	92.0	89.4	90.7	91.8	90.7	91.3
Room 2	91.2	81.7	86.5	90.6	80.9	85.7	89.4	82.3	85.8	89.6	85.3	87.4
Mean	92.8	80.2	86.5	**93.1**	80.4	86.7	91.0	87.0	89.0	91.3	**88.9**	**90.1**

Table 6. Detailed quantitative comparison on point prompt based hierarchical segmentation on the Replica dataset [49].

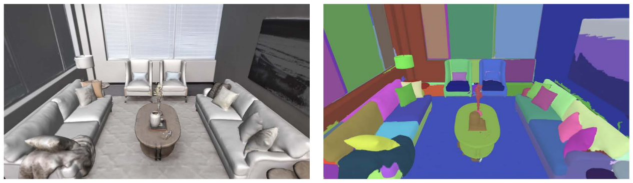


Figure 8. Scene discretization by feature clustering on mesh automatically without click.

ilar to the one proposed in ScanNet [13], where the similarity is modelled as feature distance instead of geometric smoothness. As shown in Fig. 8, OmniSeg3D provides high-quality mesh segmentation results. However, since no clear hierarchy level is specified, different objects may be segmented at different levels. To address this problem, introducing more textual or image guidance to determine the specific level in the hierarchy is worth exploring.

9. Details on Point-based Implementation

We provide the details of point-based implementation. Specifically, we integrate OmniSeg3D into Guassian Splatting [23]. This point-based representation supports easier 3D segmentation since we can simply cluster the interested objects by thresholding all the points in the scene. Firstly, we assign a view-independent feature vector ($D = 16$) to each gaussian sphere. Then we follow the hierarchical contrastive learning framework proposed in the main paper to optimize the per-point feature via differentiable rendering. The difference is that we found the sphere surface normalization term $\mathcal{L}_{norm}$ may cause instability to training. Therefore, we just keep the $\mathcal{L}_H$ for feature field optimization. During inference stage, we normalize each point feature before calculating the similarity score. For each scene, training usually costs about $30min$ on a single RTX 3090 GPU. Please check our code for more implementation details.

Besides, we believe our OmniSeg3D will also be a simple plug-in for SDF-based [31, 53] rendering pipelines.

10. Results on Hierarchical Segmentation

We present detailed results and comparisons on hierarchical 3D segmentation, where our OmniSeg3D is compared with SAM [26] and the basic implementation (without hierarchical modelling) from Sec. 3.2. Tab. 6 shows the quantitative results for hierarchical segmentation on the Replica dataset [49]. More implementation details about the evaluation are also provided.

Comparison with SAM. We compare OmniSeg3D with SAM, which predicts hierarchical masks with point prompts as input. As shown in Tab. 6, even though OmniSeg3D is not specifically designed for point-based segmentation, it still outperforms SAM on the overall mIoU metric, especially on level-2. As illustrated in Fig. 9, SAM occasionally delivers incomplete and inconsistent results for the same object in different views, which means the hierarchical relationship modelled by SAM is unstable across views. As a comparison, OmniSeg3D achieves much more stable performance through implicitly aligning multi-view inconsistent 2D segmentations and produces a stable cluster of semantic features, where the hierarchical structure is well preserved. Moreover, the underlying neural 3D reconstruction encourages the alignment of the 3D feature field with the scene geometry. This contributes to improved foreground-background separation, resulting in geometrically-guided robust segmentations which may otherwise be ambiguous from certain viewpoints.

Comparison with our basic implementation. Fig. 11 compares the intermediate score maps and final segmentations of our method with those of the basic implementation in Sec. 3.2 (with contrastive learning, without hierarchical modelling). Tab. 6 shows the quantitative results. As mentioned in Sec. 4.1, despite a minor drop in level-1 metrics, OmniSeg3D achieves large improvements in overall cross-level segmentation. The baseline method struggles to rec-

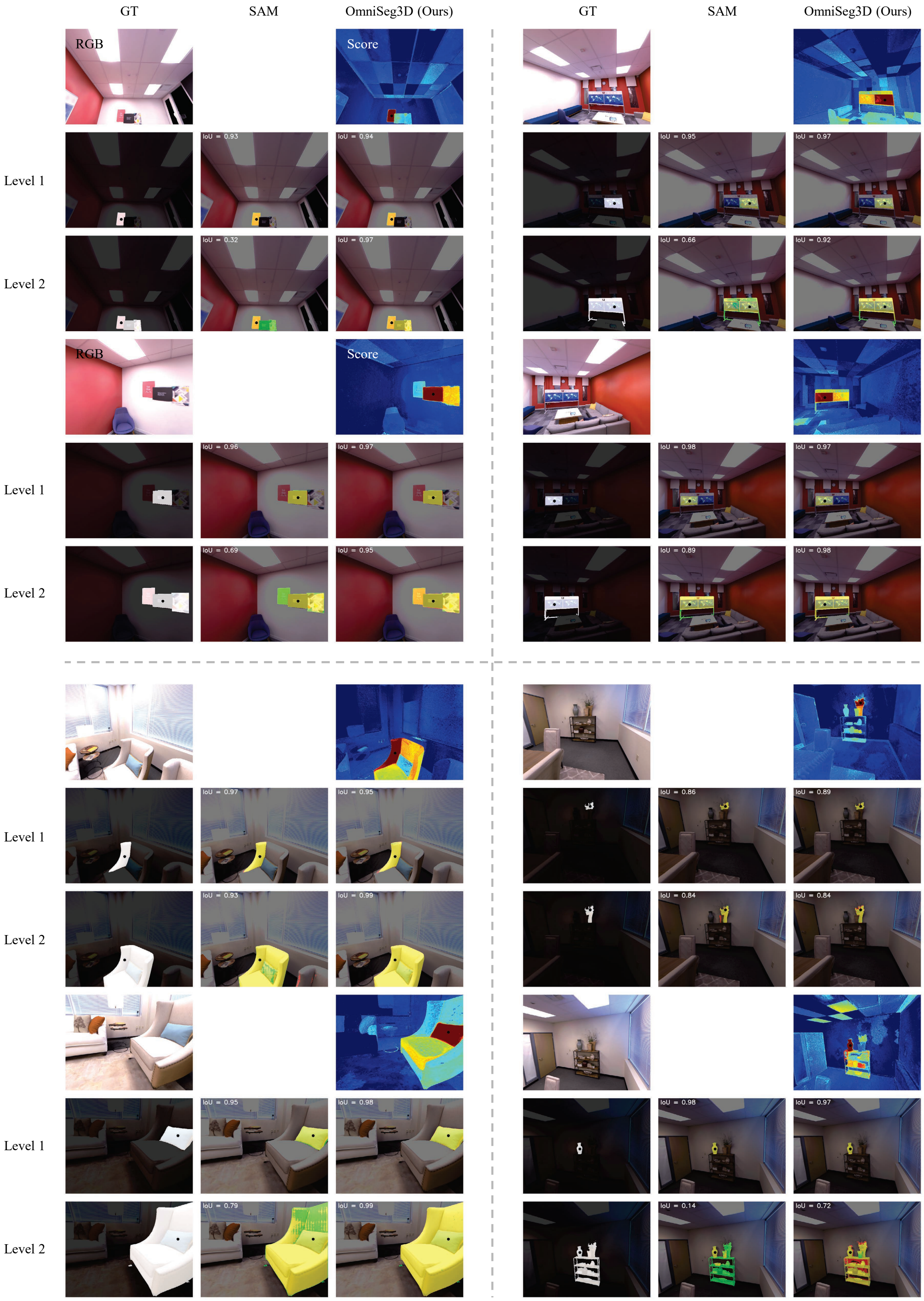
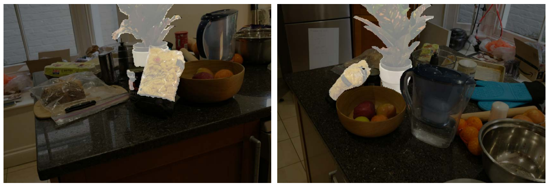
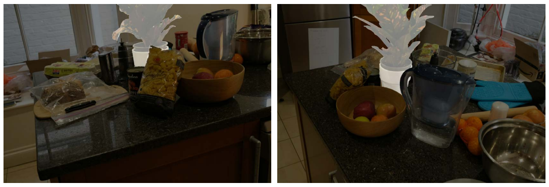


Figure 9. Comparison of our method with SAM [26] on point-based hierarchical segmentation, on the Replica dataest [49]. Point prompts are shown as black dots. The top-right image in each set is the score map obtained by our method. Colored pixels denote TP, FP and FN respectively. IoUs of each predicted mask are shown in the top-left corner.

(a) Reference view



(b) Target views, SA3D



(c) Target views, ours

Figure 10. Qualitative comparison of our method with SA3D [26].

ognize entire objects in the scene due to the lack of proper hierarchical modeling – only weak part-whole hierarchical clues can be retained from the multi-view inconsistency of SAM predictions. (For instance, an object might be segmented as a whole in one view and broken into patches in another view.) In contrast, our full method retains richer hierarchical structures in the 2D image by exhaustively correlating the segmented patches using a voting-based correlation matrix, as formulated in Sec. 3.1. The correlations are then implicitly aggregated and averaged in 3D with hierarchical contrastive learning. As depicted in the score maps in Fig. 11, our hierarchical modelling pulls together parts of the same object in the feature space while lowering the similarities with the surroundings.

Implementation details. We elaborate on how we derive the score maps in Eq. 7 in Sec. 4.1. We retrieve the volume rendered feature maps $\mathbf{f}$ (normalized so that $\|\mathbf{f}\| = 1$) from the reference and target views, and compute the spatial coordinate $\mathbf{x}$ corresponding to each pixel $\mathbf{p}$ from the rendered depth. The similarity sim between pixels $\mathbf{p}_1$ and $\mathbf{p}_2$ is defined as a distance-weighted feature similarity:

$$\mathrm{sim}\left(\mathbf{p}_1, \mathbf{p}_2\right) = \left(1 + \mathbf{f}_1 \cdot \mathbf{f}_2\right) e^{-\alpha \|\mathbf{x}_1 - \mathbf{x}_2\|} \tag{9}$$

where α is a non-negative constant depending on the spatial extent of the dataset, and $\|\mathbf{x}_1 - \mathbf{x}_2\|$ is the Euclidean distance between the rendered 3D coordinates $\mathbf{x}_1$ and $\mathbf{x}_2$.

Given the point prompt $\mathbf{p}_0$, the score for any other pixel $\mathbf{p}_i$ in the image is defined as

$$\mathrm{score}\left(\mathbf{p}_i\right) = \mathrm{sim}\left(\mathbf{p}_0, \mathbf{p}_i\right) = \left(1 + \mathbf{f}_0 \cdot \mathbf{f}_i\right) e^{-\alpha \|\mathbf{x}_0 - \mathbf{x}_i\|} \tag{10}$$

The predicted mask in image $\mathbf{I}$ is then produced by thresholding the score map as in Eq. 8. Since the feature field is queried and optimized in the 3D space, the spatial coordinates $\mathbf{x}$ and the Euclidean distances serves as a free enhancement to the semantic feature field (with a similarity metric defined on the concatenated feature $(\mathbf{f}; \mathbf{x})$), down-weighting the similarity of an object with the background. As shown in the quantitative comparisons regarding the use of spatial coordinates in Tab. 6, the involvement of $\mathbf{x}$ slightly boosts the overall performance but is not crucial for our algorithm.

11. Results on Instance Segmentation

Comparison with SA3D [6]. In addition to the quantitative results in Sec. 4.2, we compare our method qualitatively with SA3D on the `counter` scene in the Mip-NeRF 360 [2] dataset. The images in the dataset are extracted from a video sequence. We follow the inference procedure of SA3D by selecting the best out of the three masks of the foreground object (partial observation of the plant and the vase) in the first frame provided by SAM [26], and predict object segmentations in the other frames. As illustrated in Fig. 10, SA3D propagates the foreground segmentation to irrelavant contents that are occluded in the first frame. In comparison, our method correctly handles occlusions and yields view-consistent segmentations of the object, thanks to the learned feature field for all the objects in 3D. Please refer to the supplementary video for the complete result.

Details on quantitative results. We elaborate on the benchmarks in Sec. 4.2, including scribble-based segmentation on NVOS dataset [45] (built upon LLFF real dataset [37]) and multi-view mask propagation on MVSeg [39] (6 forward-facing scenes from LLFF and 4 360° scenes) and Replica [49] (instance labels provided by Semantic-NeRF [67], object list provided by SA3D [6]) datasets. For scribble-based segmentation, a pair of foreground and background scribbles is specified in the reference view, which serves as input to the 3D segmentation algorithm. The model then generates a mask in an unseen target view. The predicted mask is compared with the ground truth 2D instance segmentation. For multi-view mask propagation, given the ground truth 2D mask of an object in the reference view, the algorithm is supposed to lift the mask to 3D and propagate it to all other views. The predicted masks are compared with the ground truth mask for each view. Results for each scene are shown in Tab. 8, Tab. 9 and Tab. 7.

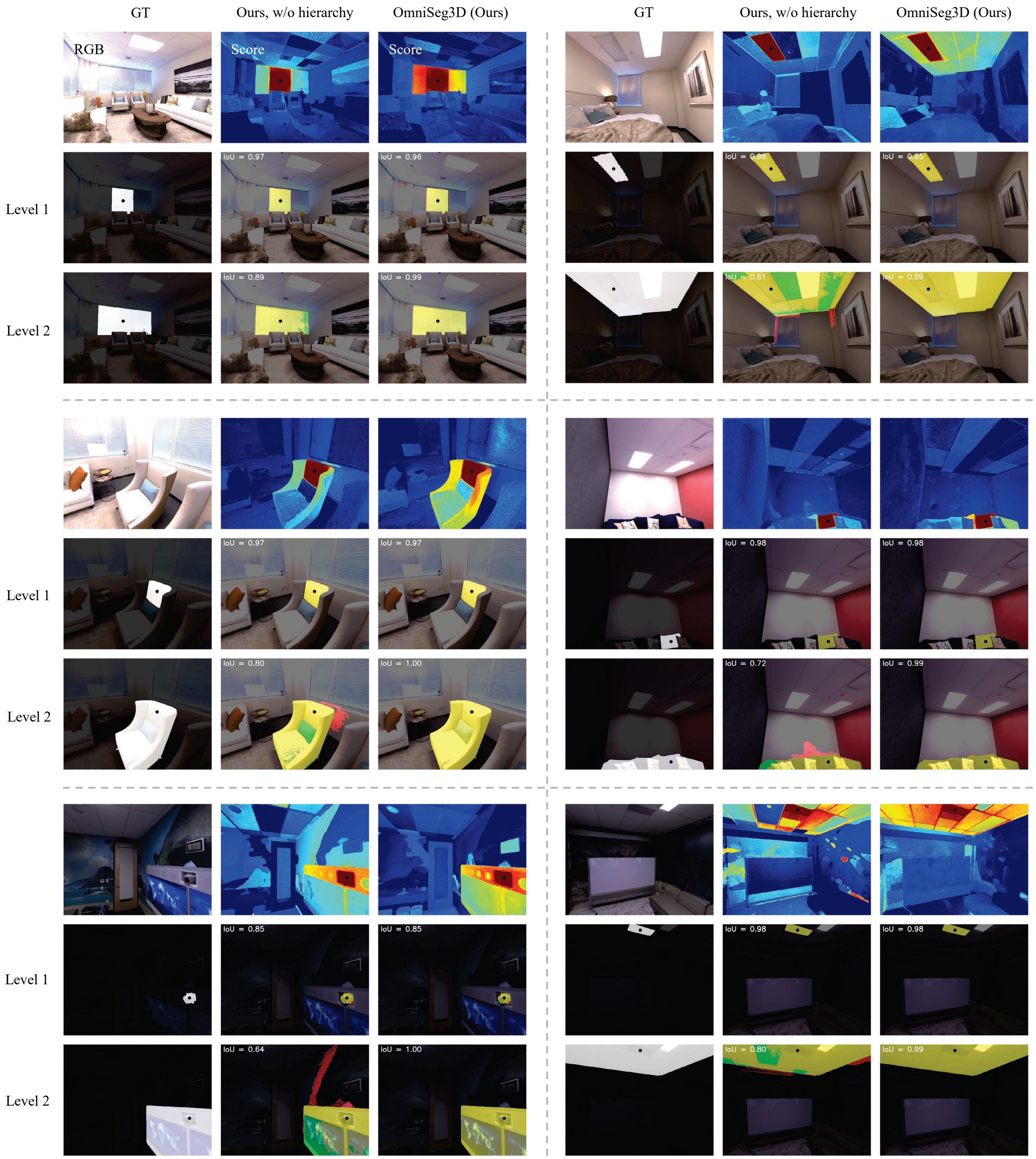
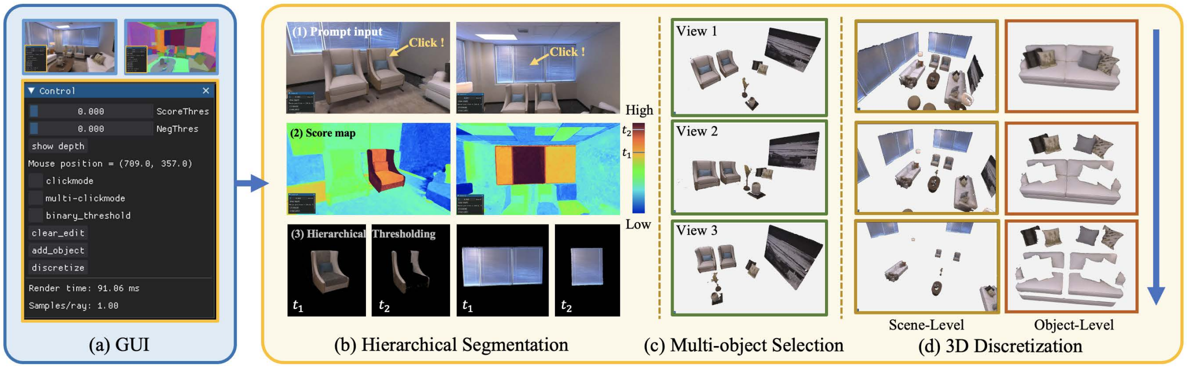


Figure 11. Comparison of the intermediate score maps and final segmentations of our method, with and without hierarchical modelling, on point-based hierarchical segmentation on the Replica dataest [49]. Point prompts are shown as black dots. Colored pixels denote TP, FP and FN respectively. IoUs of each predicted mask are shown in the top-left corner.

Implementation details. We sample positive pixels $\{\mathbf{p}_i \,|\, i \in S_{pos}\}$ uniformly from the foreground scribble or instance mask and negative pixels $\{\mathbf{p}_j \,|\, j \in S_{neg}\}$ from the background in the reference view. The score for each pixel



0.000 ScoreThres

0.000 NegThres

Mouse position = (709.0, 357.0)

Render time: 91.06 ms

Samples/ray: 1.00

(a) GUI (b) Hierarchical Segmentation (c) Multi-object Selection (d) 3D Discretization

Figure 12. Interactive 3D segmentation with (a) a graphical user interface. For `room-0` of Replica, we show the segmentation performance on (b) hierarchical inference, (c) multi-object selection, and (d) 3D discretization with our GUI.

Method	Office 0	Office 1	Office 2	Office 3	Office 4	Room 0	Room 1	Room 2	Mean
MVSeg [39]	31.4	40.4	30.4	30.5	25.4	31.1	40.7	29.2	32.4
SA3D [6]	84.4	77.0	88.9	84.4	82.6	77.6	79.8	89.2	83.0
OmniSeg3D (Ours)	83.9	85.3	89.0	87.2	78.3	83.0	79.4	88.9	**84.4**

Table 7. Breakdown results for instance segmentation on the Replica dataset [49].

$\mathbf{p}_t$ in the target view is defined through the difference of maximal similarities with positive and negative samples:

$$\text{score}\left(\mathbf{p}_t\right) = \max_{i \in S_{pos}} \text{sim}\left(\mathbf{p}_t, \mathbf{p}_i\right) - \beta \max_{j \in S_{neg}} \text{sim}\left(\mathbf{p}_t, \mathbf{p}_j\right) \tag{11}$$

where $\beta = 0.15$ and sim is defined in Eq. 9. In practice, we substitute the $\max_{i \in S_{pos}}$ operator for positive samples with a 95th percentile to suppress noise. The binarization threshold in Eq. 8 is determined by maximizing the IoU between the predicted and ground truth masks in the reference view $\mathbf{I}_{ref}$: $\max_{th} \text{IoU}\left(\{\mathbf{p}_t \in \mathbf{I}_{ref} \,|\, \text{score}\left(\mathbf{p}_t\right) > th\}, M_{GT}\right)$, then the same threshold is applied to all other views for evaluation.

12. Interactive 3D segmentation

We further show the details of our Graphic User Interface (GUI) for convenient 3D segmentation based on OmniSeg3D. In Fig. 12(a), we show the options and operation buttons. By click on the screen, user can choose the object of interest and achieve hierarchical segmentation by tuning the score threshold as shown in Fig. 12(b). Besides, the multi-click mode enable user to select multiple objects (c). By combining (b) and (c), user can discretize the whole scene in a hierarchical manner as shown in Fig. 12(d). When using InstantNGP [42] based implementation of OmniSeg3D, the rendering speed consistently reaches 20-30fps, and each interactive segmentation operation can be completed within 50ms.

Scene	IoU (%)	Acc (%)
Fern	82.7	94.3
Flower	95.3	98.9
Fortress	98.5	99.7
Horns (center)	97.7	99.6
Horns (left)	95.6	99.7
Leaves	92.7	99.5
Orchids	84.0	97.1
Trex	87.4	98.3
Mean	91.7	98.4

Table 8. Breakdown results for NVOS dataset [45].

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 3

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 6, 4

[3] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. *arXiv preprint arXiv:2306.04633*, 2023. 2, 3, 5

[4] WANG Bing, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3

Scene	MVSeg [39]		SA3D [6]		Ours	
	mIoU	Acc	mIoU	Acc	mIoU	Acc
Fern	94.3	99.2	97.1	99.6	97.5	99.7
Fortress	97.7	99.7	98.3	99.8	97.9	99.7
Horns	92.8	98.7	94.5	99.0	91.5	98.5
Leaves	94.9	99.7	97.2	99.9	96.0	99.8
Orchids	92.7	98.8	83.6	96.9	92.3	98.7
Room	95.6	99.4	88.2	98.3	97.9	99.7
Fork	87.9	99.5	89.4	99.6	90.4	99.6
Lego	74.9	99.2	92.2	99.8	90.8	99.7
Pinecone	93.4	99.2	92.9	99.1	92.1	99.0
Truck	85.2	95.1	90.8	96.7	96.1	98.7
Mean	90.9	98.9	92.4	98.9	**94.3**	**99.3**

Table 9. Breakdown results for MVSeg dataset [39].

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7

[6] 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. 2, 3, 8, 4, 6, 7

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[8] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 2, 3, 6

[9] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 45–54, 2020. 3

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3

[12] Guy Barrett Coleman and Harry C Andrews. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5): 773–785, 1979. 3

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3, 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[15] Peter Dorninger and Clemens Nothegger. 3d segmentation of unstructured point clouds for building modelling. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 35(3/W49A):191–196, 2007. 2

[16] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 3

[17] Rahul Goel, Dhawal Sirikonda, Saurabh Saini, and PJ Narayanan. Interactive segmentation of radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4201–4211, 2023. 2, 3, 8

[18] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2940–2949, 2020. 2, 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[20] Karl Heinz Höhne and William A Hanson. Interactive 3d segmentation of mri and ct volumes using morphological operations. *Journal of computer assisted tomography*, 16(2): 285–294, 1992. 2

[21] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11108–11117, 2020. 3

[22] Jing Huang and Suya You. Point cloud labeling using 3d convolutional neural network. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2670–2675. IEEE, 2016. 3

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 6

[24] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 2, 3, 5, 6

[25] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 2, 3

[26] 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, 3, 4, 6, 7, 1

[27] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022. 2, 3

[28] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 1

[29] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2, 7

[30] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2020. 5

[31] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 2

[32] Leyao Liu, Tian Zheng, Yun-Jou Lin, Kai Ni, and Lu Fang. Ins-conv: Incremental sparse convolution for online 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18975–18984, 2022. 2, 3

[33] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023. 2, 3, 6, 1

[34] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[36] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 6

[37] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 6, 4

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3, 4, 5, 6

[39] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. 2, 3, 8, 4, 6, 7

[40] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019. 3

[41] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019. 3

[42] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2, 4, 5, 6

[43] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 2

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 7

[45] Zhongzheng Ren, Aseem Agarwala, Bryan Russell, Alexander G Schwing, and Oliver Wang. Neural volumetric object selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2022. 8, 4, 6

[46] Ruwen Schnabel, Roland Wahl, and Reinhard Klein. Efficient ransac for point-cloud shape detection. In *Computer graphics forum*, pages 214–226. Wiley Online Library, 2007. 2

[47] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 2, 3

[48] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 2, 3, 6, 1

[49] 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. *arXiv preprint arXiv:1906.05797*, 2019. 6, 7, 1, 2, 3, 4, 5

[50] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Open-

mask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 2, 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[52] Guangyu Wang, Jinzhi Zhang, Kai Zhang, Ruqi Huang, and Lu Fang. Giganticnvs: Gigapixel large-scale neural rendering with implicit meta-deformed manifold. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[53] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[54] Weiyue Wang and Ulrich Neumann. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 135–150, 2018. 3

[55] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3268–3278, 2020. 3

[56] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022. 2, 3

[57] Yajie Xing, Jingbo Wang, and Gang Zeng. Malleable 2.5 d convolution: Learning receptive fields along the depth-axis for rgb-d scene parsing. In *European Conference on Computer Vision*, pages 555–571. Springer, 2020. 3

[58] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *Advances in neural information processing systems*, 32, 2019. 3

[59] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 3

[60] Haiyang Ying, Baowei Jiang, Jinzhi Zhang, Di Xu, Tao Yu, Qionghai Dai, and Lu Fang. Parf: Primitive-aware radiance fusion for indoor scene novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17706–17716, 2023. 2

[61] Fenggen Yu, Zhiqin Chen, Manyi Li, Aditya Sanghi, Hooman Shayani, Ali Mahdavi-Amiri, and Hao Zhang. Capri-net: Learning compact cad shapes with adaptive primitive assembly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11768–11778, 2022. 3

[62] Jinzhi Zhang, Mengqi Ji, Guangyu Wang, Zhiwei Xue, Shengjin Wang, and Lu Fang. Surrf: Unsupervised multi-view stereopsis by learning surface radiance field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7912–7927, 2021. 2

[63] Jianing Zhang, Jinzhi Zhang, Shi Mao, Mengqi Ji, Guangyu Wang, Zequn Chen, Tian Zhang, Xiaoyun Yuan, Qionghai Dai, and Lu Fang. Gigamvs: a benchmark for ultra-large-scale gigapixel-level 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7534–7550, 2021. 3

[64] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669, 2022. 5

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3

[66] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 3

[67] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 2, 3, 6, 4