Towards Scalable Spatial Intelligence via 2D-to-3D Data Lifting

Supplementary Material

1. Appendix

Statistics of COCO-3D Fig. 2 shows the number of instances for each category. The x-axis lists the categories, while the y-axis represents the instance count. Fig. 3 illustrates the percentage distribution of points across different categories. The x-axis represents the various categories, and the y-axis indicates the percentage of points assigned to each category. From the figures, it is evident that most points are concentrated in the "person" category, which accounts for 30% of the total points—far exceeding the other categories. Compared to other domain-specific 3D datasets, our dataset exhibits notable differences. COCO-3D is derived from the transformation of COCO data, which enables us to retain the rich semantic information and diverse annotations found in COCO. Our experiments have demonstrated that our synthetic data performs well in zero-shot transfer, giving us confidence in leveraging this dataset to enhance 3D object detection and recognition. It is particularly worth mentioning that our dataset includes a large number of scenes involving people, with especially abundant data in the "person" category. This makes our dataset more realistic when addressing human-related tasks. Pretraining on synthetic data followed by fine-tuning on real data can, to some extent, alleviate the challenges posed by the scarcity of real data.

Compare with Other 3D Datasets Compared to traditional databases Sec. 1 (such as ShapeNet [6], ModelNet [20], 3D-Future [11] that mainly focus on single objects, ScanNet [9], Matterport3D [5] that are limited to small-scale scenes), or SUN-RGBD [17] and Omni3D [4] only include monocular 3D representation datasets of indoor scenes, our COCO-3D and object365-v2-3D datasets are significantly ahead in terms of the number of scenes and categories. Specifically, COCO-3D contains 122K scene instances and 81 categories, while object365-v2-3D has 2M scene instances and 365 categories. Our dataset includes indoor and outdoor scenes. Although the data is synthetic, rich experimental results prove that it has zero shot capabilities and can be generalized to other datasets, providing sufficient data support for tasks such as 3D perception.

Discussion with SpatialVLM SpatialVLM [7] improves the spatial QA performance of VLM by converting 2D images into 3D point clouds and generating many spatial QA pairs. However, it does not calibrate the point cloud's geometric accuracy or camera parameters, nor does it carry out systematic validation on low-level 3D vision tasks such as segmentation, etc. It only addresses QA tasks about relative positions and sizes of objects. In contrast, our work builds a full 3D representation, of which the point cloud is only one part. For each scene, we calibrate gravity direction, camera parameters, and metric scale. Moreover, our experiments cover a range of spatial reasoning tasks, from low-level (semantic segmentation, instance segmentation, few-shot learning, zero-shot learning) to high-level (QA, captioning, and referring segmentation).

Dataset	Number	Categories	Class	Scenes/Objects
ShapeNet [6]	51k	55	-	Objects
ModelNet [20]	12k	40	-	Objects
3D-Future [11]	16k	34	-	Objects
ABO [8]	8k	63	-	Objects
Toys4K [18]	4k	105	-	Objects
CO3D V1 / V2 [15]	19 / 40k	50	-	Objects
ScanObjectNN [19]	15k	15	-	Objects
GSO [10]	1k	17	-	Objects
AKB-48 [13]	2k	48	-	Objects
OmniObject3D [21]	6k	190	-	Objects
LLFF [14]	35	-	-	Scenes
DTU [1]	124	-	-	Scenes
BlendedMVS [22]	133	-	-	Scenes
ScanNet [9]	1509	-	20	Scenes
Matterport3D [5]	90	-	21	Scenes
Tanks and Temples [12]	21	-	-	Scenes
ETH3D [16]	25	-	-	Scenes
ARKitScenes [3]	1004	-	-	Scenes
ScanNet++ [23]	460	-	100	Scenes
S3DIS [2]	271	-	13	Scenes
Structured3D [24]	3500	-	25	Scenes
COCO-3D	122K	-	81	Scenes
object365-v2-3D	2M	-	365	Scenes

Table 1. **A comparison between COCO-3D, Object365-v2-3D, and other commonly-used 3D scenes/object datasets.**

More Visualization In Fig. 4 and Fig. 5, we provide more visualization results of the zero-shot experiments on ScanNet for Uni3D.

Data Quality Assurance In the process of constructing the dataset from 2D images to 3D representations, we implemented a series of data quality assurance mechanisms to ensure that the generated data meets high standards in terms of authenticity, accuracy, and consistency. First, through

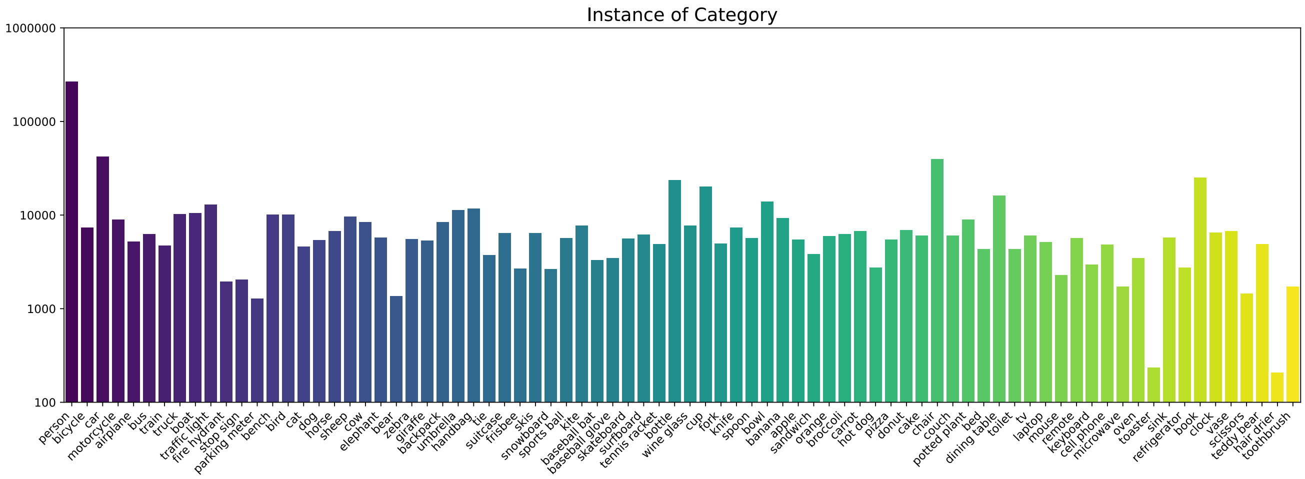

Figure 2. **Statistic of COCO-3D.** The number of instances for each category.

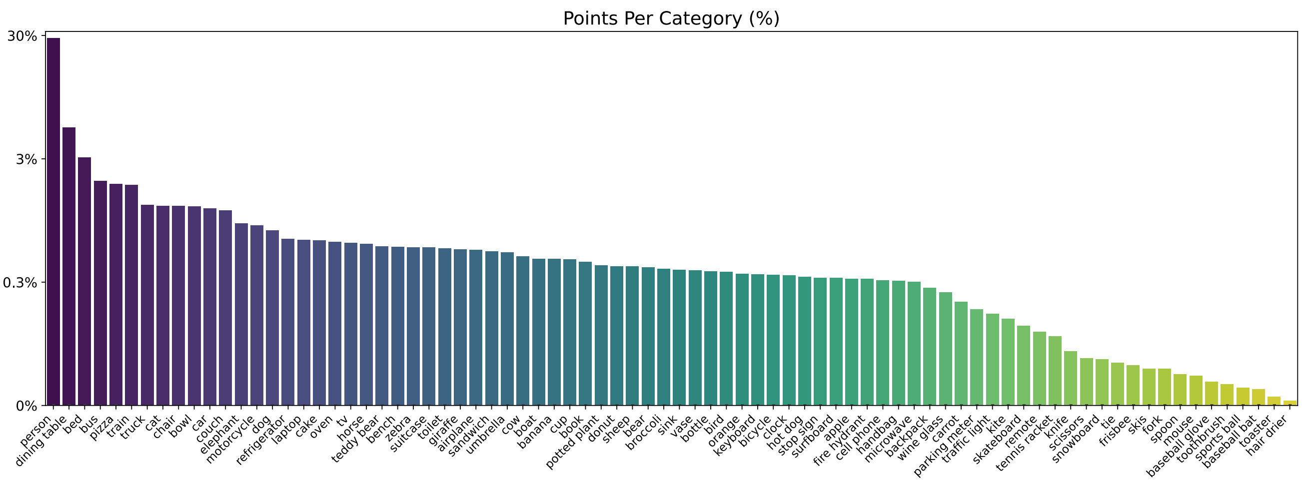

Figure 3. **Statistic of COCO-3D.** The percentage distribution of number points across various categories.

depth estimation and camera parameter prediction, we use an automatic filtering algorithm after generating a preliminary 3D representation to remove edge areas, undefined areas, and predicted abnormal points, and calculate the scale factor based on the relative depth and quantized depth distribution in the valid point set to achieve an effective fusion of depth information and absolute scale. Next, we select some samples and use Open3D visualization for manual verification to verify the consistency between the original 2D annotations and the generated 3D annotations, and check the correspondence between the 3D representation and the original 2D image, so as to promptly discover and correct possible errors in the automatic process. Finally, we further ensure the rationality of the data in scale and structure by statistically analyzing the size distribution of each category and comparing it with the actual physical size.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. 1

[2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 1

[3] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Bran-

Figure 4. **Visualization of zero-shot point cloud instance segmentation results.** Despite significant differences between synthetic and real data, models trained on COCO-3D can directly generalize to ScanNet.

don Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 1

[4] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023. 1

[5] 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. *International Conference on 3D Vision (3DV)*, 2017. 1

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1

[8] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision*

Figure 5. **Visualization of zero-shot point cloud semantic segmentation results.** Despite significant differences between synthetic and real data, models trained on COCO-3D can directly generalize to ScanNet.

and pattern recognition, pages 21126–21136, 2022. 1

[9] 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. 1

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 1

[11] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 1

[12] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 1

[13] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022. 1

[14] 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. 1

[15] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction, 2021. 1

[16] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[17] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1

[18] Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1798–1808, 2021. 1

[19] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019. 1

[20] Kashi Venkatesh Vishwanath, Diwaker Gupta, Amin Vahdat, and Ken Yocum. Modelnet: Towards a datacenter emulation environment. In *2009 IEEE Ninth International Conference on Peer-to-Peer Computing*, pages 81–82. IEEE, 2009. 1

[21] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 1

[22] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1790–1799, 2020. 1

[23] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 1

[24] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. 1