R3DS: Reality-linked 3D Scenes for Panoramic Scene Understanding

Qirui Wu[1] Sonia Raychaudhuri[1] Daniel Ritchie[2] Manolis Savva[1] Angel X. Chang[1,3]
[1]Simon Fraser University [2]Brown University [3]Alberta Machine Intelligence Institute (Amii)
https://3dlg-hcvc.github.io/r3ds/

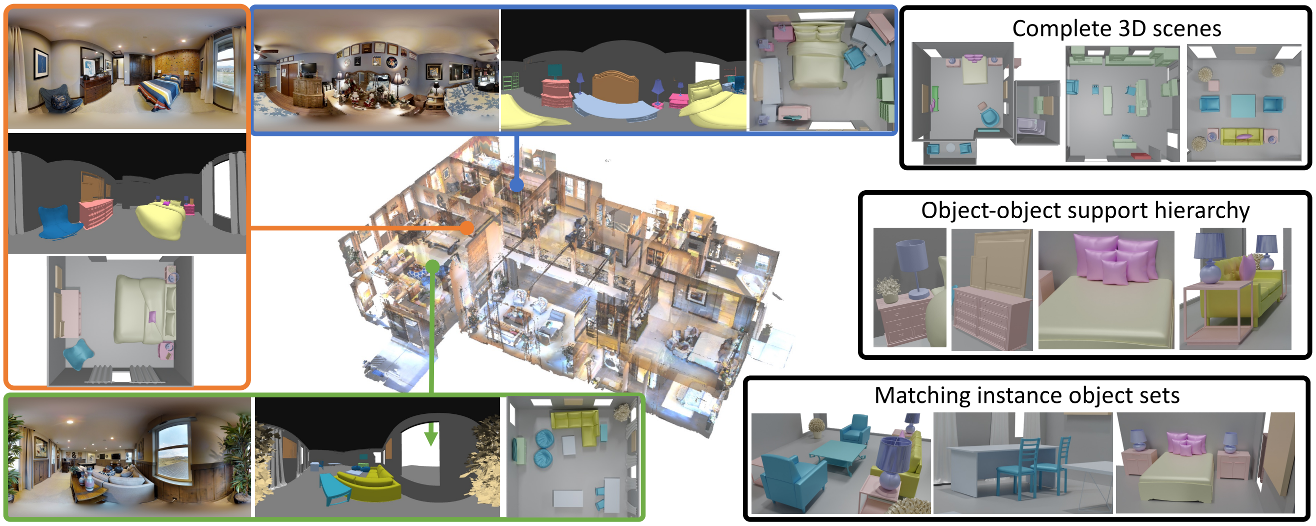
Complete 3D scenes
Object-object support hierarchy
Matching instance object sets

Figure 1. **Left**: the Reality-linked 3D Scenes dataset (R3DS) fills a gap between synthetic 3D scenes and reconstructions of real-world environments by providing 3D scene proxies linked to real-world panoramas from Matterport3D (three example panoramas and 3D scenes shown). **Right**: our dataset contains scenes with higher density and completeness compared to prior datasets, and provides additional annotations such as object support (what objects or architectural elements support other objects), and matching object sets (e.g., pairs of the same nightstand).

Abstract

We introduce the Reality-linked 3D Scenes (R3DS) dataset of synthetic 3D scenes mirroring the real-world scene arrangements from Matterport3D panoramas. Compared to prior work, R3DS has more complete and densely populated scenes with objects linked to real-world observations in panoramas. R3DS also provides an object support hierarchy, and matching object sets (e.g., same chairs around a dining table) for each scene. Overall, R3DS contains 19K objects represented by 3,784 distinct CAD models from over 100 object categories. We demonstrate the effectiveness of R3DS on the Panoramic Scene Understanding task. We find that: 1) training on R3DS enables better generalization; 2) support relation prediction trained with R3DS improves performance compared to heuristically calculated support; and 3) R3DS offers a challenging benchmark for future work on panoramic scene understanding.

1. Introduction

Datasets of 3D indoor environments are increasingly used for research on scene understanding [1, 3, 14], embodied AI [2, 10, 12], and scene generation [8, 13]. However, constructing 3D scene datasets is time-consuming and require expertise. Compared to reconstruction based on 3D scans, synthetic 3D scenes are complete and easy to manipulate but often do not match the statistics of real-world spaces and are artificially "clean". There have been some attempts to create "synthetic" replicas of real world by matching CAD models to objects in scans [10, 12]. These efforts have been limited in scale and often result in partial and sparsely populated synthetic counterparts of the real environments.

We design a framework that allows users to author 3D scenes from RGB panoramas and create **R3DS**: a dataset of '**R**eality-linked' **3D S**cenes. Each 3D scene in our dataset is a complete proxy of an environment from the Matterport3D [3] dataset, representing both the 3D architecture and the objects. Thus, each scene is linked to a real space, with correspondences established between each object observation and the synthetic object. These reality-linked scenes reflect denser real-world arrangements of objects.

Compared to prior efforts such as Scan2CAD [1] and CAD-Estate [7], our dataset provides more complete scenes, with salient observed objects being captured in the

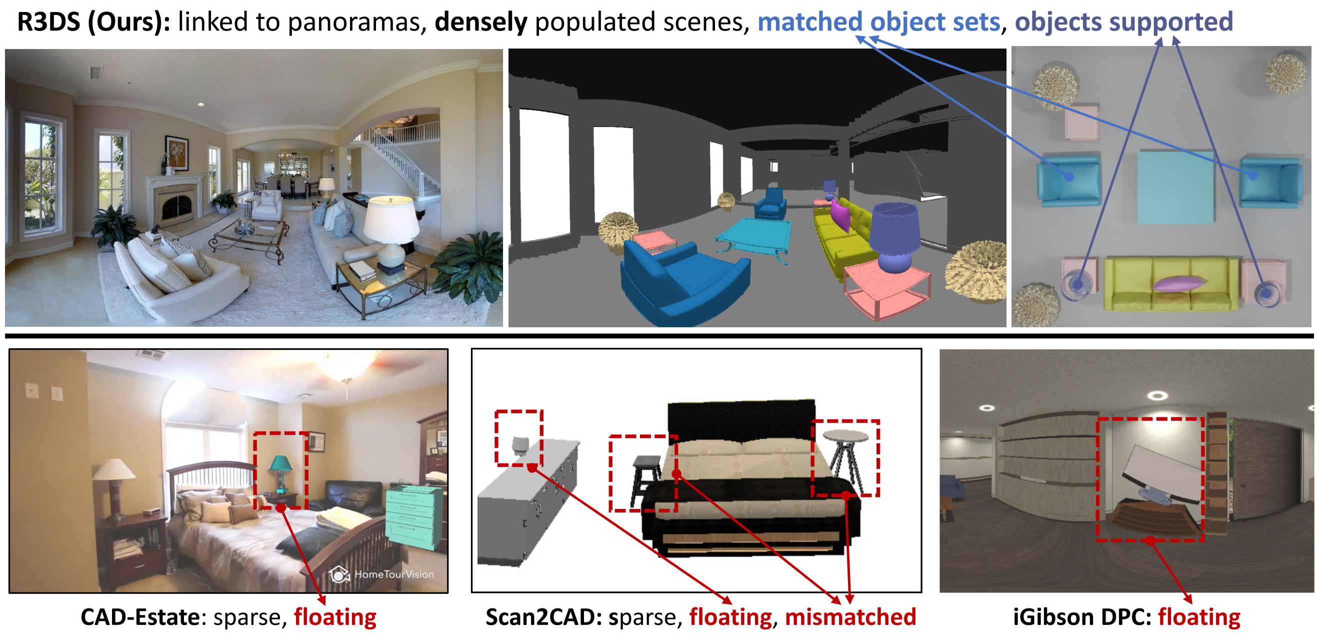
R3DS (Ours): linked to panoramas, densely populated scenes, matched object sets, objects supported
HomeTourVision
CAD-Estate: sparse, floating
Scan2CAD: sparse, floating, mismatched
iGibson DPC: floating

Figure 2. **Dataset comparison.** (Top) shows different views of a scene annotated in R3DS. Comparison with previous datasets (bottom) shows (1) R3DS has more complete scenes than the previous datasets; (2) Objects in R3DS are properly supported by either architecture or other objects unlike the others (e.g. floating objects with no proper support); (3) R3DS is annotated using the same 3D model for objects arranged together (chairs by the dining table, couches arranged together).

layout. Moreover, we provide a support hierarchy defining what objects are placed on other objects and specify sets of identical objects such as dining chairs around a table, allowing for creating realistic variations of the scene by swapping the entire set to a different chair design.

We demonstrate the value of our dataset by using it for the Panoramic Scene Understanding task. We show that leveraging the denser layouts and support hierarchy information in our scenes leads to improved object detection performance and better generalization compared to training using other datasets previously used for this task.

In summary, we make the following contributions:

- We design a framework for efficient construction of synthetic scenes from real panoramas and use it to create R3DS: a dataset of reality-linked 3D scenes.
- R3DS provides more complete and realistic scenes with correspondences between real and synthetic objects, and object-object support relations.
- We show that the more complete layouts and support relations in our dataset enable better performance and generalization in the Panoramic Scene Understanding task, and that our dataset offers a challenging benchmark for future work in scene understanding.

2. The R3DS Dataset

We describe the construction of the R3DS dataset and present a statistical analysis of the scenes it contains. Figure 1 shows example annotations from our dataset.

We collect annotations for 20 Matterport3D houses with 808 panoramas in total. We discard panoramas taken on stairs or outside a house, since they have a limited number of objects that can be placed. After filtering we have 769 panoramas for our analysis and experiments. For 73 panoramas, we collect two sets of annotations for each, to obtain a total of 842 annotated object arrangements across 22 different Matterport3D region types. The panoramas with two annotations serve as a test of annotator consistency and add diversity. In total, R3DS contains 19,050 object instances from 3,784 unique 3D CAD models spanning over 110 fine-grained object categories. Table 1 shows a comparison of overall statistics with previous 3D indoor scene datasets. See the supplement for statistics about annotated objects.

Compared to prior datasets that align CAD models to real-world scenes, R3DS is more complete, providing annotated object support hierarchies and matching object instances. CAD-Estate [7] annotates RGB videos with incomplete 3D objects and partial architectural room layouts under limited view. This results in annotations of objects floating mid-air and not properly supported. Scan2CAD [1] also lacks support structure without providing clean 3D architecture. See examples in Fig. 2. In contrast, our R3DS scenes have an accurate support hierarchy by construction. OpenRooms [6] augments Scan2CAD with room layouts representing the architecture. However, the architecture in R3DS is more complex and realistic, especially due to inclusion of more doors (1.92 doors per room in R3DS vs 0.67 in OpenRooms).

Dataset	Source	CAD Alignment	Type	Houses/Rooms	Panos	#CAD	#Objects	#Cat	Ave Obj	Ave Cat	Sup	Match
Scan2CAD [1]	ScanNet [4]	Annotator	scan	- / 1506	✗	3,049	14,225	35	9.4	4.1	✗	✗
OpenRooms [6]	ScanNet [4]	Scan2CAD [1]	scan	- / 1288	✗	2,651	16,014	38	12.4	6.3	✗	✗
ReplicaCAD [12]	Replica [11]	Artist recreation	scan	- / 105*	✗	92	2,293	44	21.8	14.4	✗	✗
CAD-Estate [7]	RealEstate10K [17]	Annotator	video	19,512	✗	12,024	100,882	49	6.3	3.4	✗	✗
Replica-Pano [5]	Replica [11]	Heuristic	pano	- / 27	2700	-	-	25	-	-	✗	✗
iGibson-DPC [14]	iGibson [10]	Heuristic	pano	15 / 100	1500	500	26,998	57	17.9	10.2	✗	✗
R3DS (Ours)	Matterport3D [3]	Annotator	pano	20 / 370	842	3,784	19,050	110	22.9	10.4	✓	✓

Table 1. **Comparison with 3D indoor scene datasets aligned with real-world images, videos, or scans.** Our R3DS dataset contains more densely populated annotations compared to other datasets, with objects from 110 different categories. We report the unique models (#CAD), object categories (#Cat), object instances (#Objects) as well as average number of objects and object categories per annotation.

Evaluation of CAD object annotation quality is non-trivial as the 'ground truth' from the semantically annotated 3D reconstructions is itself imperfect. We measured how closely our annotated CAD objects conform to the real objects using the average 2D IoU between CAD object mask and ground truth 2D mask. R3DS is at 42.6% vs 38.5% for Scan2CAD, across 8 common object categories (bed, sofa, chair, cabinet, tv/monitor, table, shelving, bathtub).

Of the datasets previously used for Panoramic Scene Understanding, Replica-Pano [5] has not been released, and iGibson-DPC [14] is the only dataset with synthetic panoramic images annotated with 3D objects and room layout. iGibson-DPC is built on scenes from iGibson [9] by randomly replacing objects with different models from the same category and rendering using the iGibson simulator to render panoramas. The selection and placement of objects in iGibson-DPC is based on heuristic algorithms, while our R3DS is manually annotated and placed 3D models are verified in terms of match and alignment to the object masks. Moreover, iGibson-DPC contains unrealistic object arrangements (e.g., floating TV in Fig. 2).

3. Experiments

We showcase the value of R3DS on the Panoramic Scene Understanding (PanoSun) task [5, 14, 15]. Given an input RGB panorama, the goal is to estimate the room layout, detect objects in 2D, estimate their 3D oriented bounding boxes and also reconstruct 3D object meshes. We train and evaluate DPC [14] on the iGibson-DPC (IG) [5, 9, 14], Structured3D (S3D) [16], and R3DS datasets. We consider three variants of R3DS based on the input panorama: *R3DS-real* where we use the Matterport3D panoramas, *R3DS-syn* where we use rendered panoramas (at the same camera poses) from the annotated synthetic scenes, and *R3DS-mix* where we combine the two types of panoramas and double the available data. Our experiments show that methods trained on R3DS data benefit from its realism and generalize better when evaluated on photorealistic images. We also investigate the role of object support hierarchy information in improving performance. See the supplement for details about metrics.

Test	Train	3D detection ↑		Collision ↓		Attachment F1 ↑			
		IoU	mAP	mesh	arch	obj	wall	floor	ceil
IG	IG	**27.5**	**30.3**	1.662	2.594	53.1	**76.8**	**95.0**	**86.2**
	IG+R3DS	24.0	30.2	1.404	2.254	**59.7**	64.1	94.6	2.7
	R3DS-real	17.3	13.4	**0.242**	1.456	38.8	64.0	92.8	0.0
	R3DS-syn	23.2	14.2	0.480	1.938	48.5	46.7	93.8	28.6
	R3DS-mix	21.6	15.6	0.434	**1.248**	43.1	67.2	90.1	9.8
S3D	IG	19.5	3.5	1.016	2.651	**50.9**	**68.7**	90.8	**11.6**
	IG+R3DS	**19.7**	7.0	0.868	2.089	52.0	67.4	**91.2**	1.8
	R3DS-real	18.4	7.1	0.600	2.598	45.0	61.6	89.7	0.7
	R3DS-syn	19.0	4.8	0.644	2.561	49.3	49.7	**91.2**	2.4
	R3DS-mix	19.6	**7.5**	**0.463**	**1.673**	47.8	64.1	87.2	0.9
R3DS	IG	15.6	5.9	0.575	1.959	**53.8**	50.7	51.2	0.0
	IG+R3DS	17.5	14.1	0.281	1.267	49.5	**61.6**	58.6	0.0
	R3DS-real	16.4	15.0	0.226	1.562	44.0	57.3	58.9	0.0
	R3DS-syn	14.0	8.4	0.390	1.664	54.1	40.6	49.1	0.0
	R3DS-mix	**17.6**	**15.8**	**0.171**	**1.007**	48.5	58.3	**60.1**	0.0

Table 2. **Cross-dataset evaluation for the Panoramic Scene Understanding task.** We evaluate 3D detections with class-agnostic IoU and mAP at IoU of 0.15, and report object collisions. The highlighted rows indicate the most challenging scenario.

3.1. Results

1) Does R3DS help DPC generalize to real images? Since the original DPC work only trained and evaluated on synthetic data, it is unclear how well it performs on realistic panoramic imagery. We hypothesize that training on R3DS will lead to better performance. We separately show results on 3D object detection and scene relatioin classsification.

Object detection. For 3D object detection, we train DPC on different data settings and conduct a cross-dataset evaluation (see Tab. 2). To investigate how models perform on out-of-distribution scenes, we evaluate models on Structured3D, as its images are near-realistic. To explore whether DPC training benefits from R3DS given the same amount of data, we create a special data input *IG+R3DS* that combines iGibson and R3DS panoramas by randomly replacing half (500) of iGibson data with *R3DS-real* data. The results show that *IG+R3DS* performs almost the same as *IG* with fewer collisions on iGibson, but it remarkably outperforms *IG* on the test set of *R3DS* and *S3D* by 8.2 and 3.5 improvements on 3D mAP, respectively. It also averages 0.221 fewer mesh collisions. There are noticeable performance gaps on iGibson for models trained on R3DS data likely due to the data domain shift. Among the three variants of R3DS

	3D detection		Collision ↓		Support F1 ↑		
Train	IoU↑	mAP↑	mesh	arch	obj	floor	ceil
IG	14.2	5.2	0.703	1.639	**4.1**	85.3	0.0
S3D	16.4	**10.0**	**0.112**	1.226	3.1	**86.8**	0.0
IG+S3D	**17.1**	9.7	0.133	**1.162**	3.3	84.8	0.0

Table 3. Performance of models trained on three synthetic datasets (IG, S3D, and IG+S3D) evaluated on the *R3DS-full* dataset, where "full" indicates all 840 panoramas are used for testing.

data, *R3DS-mix* outperforms the others on all three test sets regarding 3D IoU and 3D mAP with the fewest mesh and architecture collisions. Although *R3DS-syn* underperforms on *R3DS* and *S3D* test sets, it achieves better performance than *IG* with even less data.

Scene relation classification. We report F1 scores for identifying attachment relationships of objects to other objects and architecture elements (see Tab. 2). We note that models trained with synthetic renderings perform better than those trained on real images. That is because synthetic renderings present cleaner and simpler scenes with fewer objects than real world and simpler illumination such that DPC finds it easier to learn object-object and object-architecture relations. Also, note that the predictions of object-ceiling attachments can be extremely low because few objects are attached to the ceiling in the ground truth data. We show qualitative examples in the supplement.

2) Is R3DS a challenging, high-quality test set? How would a model trained on pure synthetic data perform on complex real data (R3DS)? Due to its modest scale, we propose using R3DS as a challenging, high-quality test set. Specifically, we evaluate the synthetic-to-real performance of DPC by training on iGibson and/or Structured3D and testing on all panoramas in R3DS-real. Table 3 shows that a model trained with Structured3D performs the best (10.0 3D mAP and 0.112 mesh collision) as it observes the most photo-realistic images. DPC benefits from the synthetic data for higher bounding box IoUs, since it possesses accurately aligned 3D bounding box and more unoccluded objects. However, mAP performance is lower due to worse object recognition ability. All models struggle to predict correct object-wise support relations but do a better job of predicting object-floor support relations.

3) Are R3DS support relations helpful for PanoSun? We investigate whether the support relationships between objects provided in our R3DS scene hierarchy help boost performance of holistic scene understanding. We augment DPC's Relation Scene-GCN module with additional support relation prediction branches. Besides obtaining explicitly annotated scene support relations from R3DS, it is also possible to compute heuristic support relations from object bounding boxes. Specifically, an object is supported by another if their bounding boxes intersect within tolerance distance of 0.1m and the centroid of the former object is higher than that of the latter. Support by wall/floor/ceiling is calculated in the same way without the height judgment. This definition is similar to how DPC defines object attachment. Table 4 shows that incorporating support relation prediction indeed influences the performance of DPC. Heuristic support information may worsen 3D object detection (mAP in *R3DS-real* and *R3DS-syn*), but it eliminates mesh collisions the most. Learning support relations from R3DS annotations leads to a 2.4 improvement on mAP in *R3DS-mix*, although the classification F1 score is low.

		3D detection		Collision ↓		Support F1 ↑		
Train	Supp.	IoU↑	mAP↑	mesh	arch	obj	floor	ceil
R3DS-real	none	16.4	15.0	0.226	1.562	-	-	-
	heur	16.2	14.1	**0.219**	**1.329**	3.2	69.9	0.0
	anno	**16.6**	**15.9**	0.349	1.404	**38.5**	**94.4**	0.0
R3DS-syn	none	14.0	**8.4**	0.390	1.664	-	-	-
	heur	**14.6**	7.8	**0.281**	1.301	4.6	82.9	**52.6**
	anno	14.3	8.2	0.349	**1.219**	**32.0**	**95.0**	0.0
R3DS-mix	none	17.6	15.8	0.171	**1.007**	-	-	-
	heur	**19.2**	17.7	**0.151**	1.267	3.6	83.7	0.0
	anno	18.6	**18.2**	0.158	1.308	**12.0**	**96.2**	**85.8**

Table 4. Performance on *R3DS-real* of DPC models trained on variants of R3DS with different support relation settings. We compare the original model without support (*none*) against models supervised with support that is heuristically computed (*heur*) or annotated from R3DS (*anno*).

4. Conclusion

We introduced the R3DS dataset. R3DS provides more complete, densely populated, and richly annotated synthetic 3D scene proxies of real-world environments with linked panoramic images. We showed the usefulness of R3DS on the Panoramic Scene Understanding task. Our experiments demonstrate the value of realistic synthetic recreations in this task, in particular through the use of object support information. While we focused on the PanoSun task, R3DS can also be useful for other tasks such as single-view shape retrieval, single-view object pose estimation, and panoramic scene graph prediction.

Acknowledgements. This work was funded in part by a CIFAR AI Chair, a Canada Research Chair, NSERC Discovery Grant, NSF award #2016532, and enabled by support from WestGrid and Compute Canada. Daniel Ritchie is an advisor to Geopipe and owns equity in the company. Geopipe is a start-up that is developing 3D technology to build immersive virtual copies of the real world with applications in various fields, including games and architecture. We thank Madhawa Vidanapathirana, Weijie Lin, and David Han for help with development of the annotation tool, and Denys Iliash, Mrinal Goshalia, Brandon Robles, Paul Brown, Chloe Ye, Coco Kaleel, Elizabeth Wu and Hannah Julius for data annotation, and Ivan Tam, Austin Wang, and Ning Wang for feedback on the paper draft.

References

[1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Nießner. Scan2CAD: Learning CAD model alignment in RGB-D scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[2] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, Manolis Savva, and Hao Su. Rearrangement: A challenge for embodied AI. *arXiv preprint arXiv:2011.01975*, 2020. 1

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 667–676. IEEE, 2017. 1, 3

[4] 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 (CVPR)*, pages 5828–5839, 2017. 3

[5] Yuan Dong, Chuan Fang, Zilong Dong, Liefeng Bo, and Ping Tan. PanoContext-Former: Panoramic total scene understanding with a transformer. *arXiv preprint arXiv:2305.12497*, 2023. 3

[6] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Zexiang Xu, Hong-Xing Yu, Kalyan Sunkavalli, Milos Hasan, Ravi Ramamoorthi, and Manmohan Chandraker. OpenRooms: An end-to-end open framework for photorealistic indoor scene datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[7] Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, and Vittorio Ferrari. CAD-estate: Large-scale CAD model annotation in RGB videos. *arXiv preprint arXiv:2306.09011*, 2023. 1, 2, 3

[8] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems*, 34:12013–12026, 2021. 1

[9] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martın-Martın, Linxi Fan, Guanzhi Wang, Shyamal Buch, Claudia D'Arpino, Sanjana Srivastava, Lyne P Tchapmi, Kent Vainio, Li Fei-Fei, and Silvio Savarese. iGibson, a simulation environment for interactive tasks in large realistic scenes. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, 2021. 3

[10] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE, 2021. 1, 3

[11] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[12] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021. 1, 3

[13] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 1

[14] Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. DeepPanoContext: Panoramic 3D scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 12632–12641, 2021. 1, 3

[15] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. PanoContext: A whole-room 3D context model for panoramic scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–686. Springer, 2014. 3

[16] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. *arXiv preprint arXiv:1908.00222*, 2019. 3

[17] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 3