

Mobile Robot Oriented Large-Scale Indoor Dataset for Dynamic Scene Understanding

Yi-Fan Tang[†], Cong Tai[†], Fang-Xing Chen[†], Wan-Ting Zhang, Tao Zhang, Xue-Ping Liu, Yong-Jin Liu, Long Zeng*

Project website: <https://jackyzengl.github.io/THUD-Robotic-Dataset.github.io/>

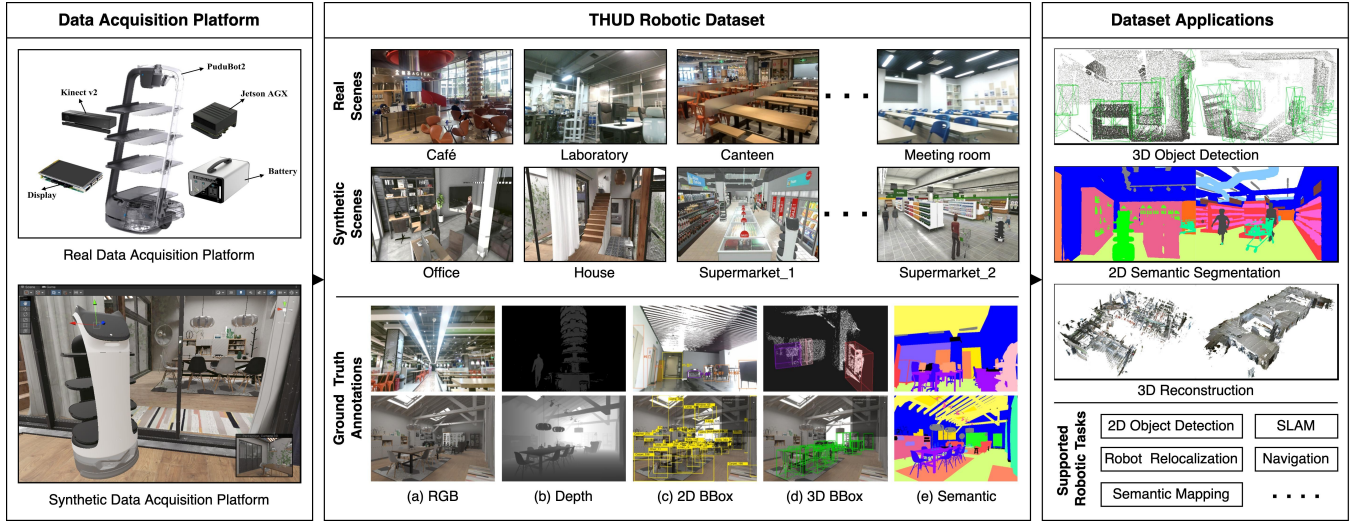


Fig. 1: THUD robotic dataset, first column: real and synthetic data acquisition platforms; second column: scene classes and annotation types; third column: applications supported by THUD.

Abstract—Most existing robotic datasets capture static scene data and thus are limited in evaluating robots’ dynamic performance. To address this, we present a mobile robot oriented large-scale indoor dataset, denoted as THUD (Tsinghua University Dynamic) robotic dataset, for training and evaluating their dynamic scene understanding algorithms. Specifically, the THUD dataset construction is first detailed, including organization, acquisition, and annotation methods. It comprises both real-world and synthetic data, collected with a real robot platform and a physical simulation platform, respectively. Our current dataset includes 13 large-scale dynamic scenarios, 90K image frames, 20M 2D/3D bounding boxes of static and dynamic objects, camera poses, and IMU. The dataset is still continuously expanding. Then, the performance of mainstream indoor scene understanding tasks, e.g. 3D object detection, semantic segmentation, and robot relocalization, is evaluated on our THUD dataset. These experiments reveal serious challenges for some robot scene understanding tasks in dynamic scenes. By sharing this dataset, we aim to foster and iterate new mobile robot algorithms quickly for robot actual working dynamic environment, i.e. complex crowded dynamic scenes.

Index Terms—mobile robot, RGB-D dataset, dynamic indoor scenes

I. INTRODUCTION

Mobile robots are widely used in various scenarios, e.g. restaurant and supermarket, which are typical dynamic environments with many moving objects. However, most existing

mobile robotic datasets mainly capture data of static scenes, which cannot support well in training and evaluating mobile robots for their practical working status, particularly for large-scale indoor scenes. To overcome this limitation, we have constructed a dataset specifically designed for training mobile robot to perform various tasks in large-scale dynamic indoor scenes.

The overview of our mobile robot oriented large-scale indoor dataset, named THUD (Tsinghua University Dynamic), is shown in Fig. 1. First, the dataset is acquired in 13 dynamic scenarios, i.e. 8 real-world and 5 synthetic scenes, with moving pedestrians, mobile robots, shopping cart, and other dynamic objects. The data are collected via a real robot platform and a physical simulation platform, both containing different levels of dynamic complexity. This is useful in evaluating real mobile robots’ performance in restaurant-like scenarios, which have vary dynamic complexity in different time period. Then, to provide rich information, the collected data is annotated with dense ground-truth labels, including over 90K frames (each with a RGB image and a depth map), over 20M 2D/3D object detection bounding boxes, semantic segmentation annotations, and camera poses. This makes our dataset applicable to both static and dynamic mobile robot indoor tasks.

Thus, we evaluated and tested THUD on some representative algorithms from static and dynamic mobile robot

[†]These authors contributed equally to this work.

*Corresponding author.

TABLE I: RGB-D Datasets Comparison

Annotation type	Dataset	Data type	Year	# Labels	# Annotations per frame	# Object classes	Dynamic objects ^a	Tasks ^b			
								2D	3D	SS	RC
2D	B3DO [1]	Real	2011	849 frames	2~5	50+	×	✓	×	×	×
	NYU-Depth v2 [2]	Real	2012	1,449 frames	30~40	894	×	×	×	✓	×
	SUN3D [3]	Real	2013	8 scans	10~15	-	×	×	×	✓	✓
	Stanford 2D-3D-S [4]	Real	2017	70,496 frames	10~15	13	×	×	×	✓	×
	SceneNet RGB-D [5]	Synthetic	2017	5M frames	20~30	255	×	×	×	✓	✓
3D	SUN RGB-D [6]	Real	2015	10k frames	20~30	800	×	×	✓	✓	✓
	ScanNet [7]	Real	2017	2.5M frames	10~15	21	×	×	✓	✓	✓
	SUN-CG [8]	Synthetic	2017	500k frames	5~15	84	×	×	×	✓	✓
	Matterport 3D [9]	Real	2017	194,400 frames	5~15	40	×	×	×	✓	✓
	InteriorNet [10]	Synthetic	2018	20M frames	20~30	158	×	×	×	✓	✓
	ARKitScenes [11]	Real	2022	5,047 scans	5~10	17	×	×	×	×	×
	ScanNet++ [12]	Real	2023	1,858 scans	20~30	1,000+	×	×	×	✓	✓
	THUD(Ours)	Synthetic&Real	2023	90k frames	150~200	91	✓	✓	✓	✓	✓

^aDynamic objects: Includes moving robots, walking people, and people moving with shopping carts.

^b2D: 2D Object Detection; 3D: 3D Object Detection; SS: Semantic Segmentation; RC: 3D Reconstruction.

indoor tasks, i.e. 3D object detection, semantic segmentation, and robot relocalization tasks. The results demonstrate that algorithms designed for different tasks exhibit varying degrees of performance degradation in scenes include dynamic objects, especially in robot relocalization. We can expect our dataset or its expanding version will support other static and dynamic indoor mobile robot tasks well, e.g. scene reconstruction, semantic map construction, robot navigation, object tracking, semantic segmentation, trajectory prediction, domain adaption, etc.

By utilizing the THUD dataset, we hope to provide a valuable resource for researchers and developers in the field of mobile robotics, fostering the development and performance improvement of mobile robot systems. By sharing this dataset, we aim to encourage more researchers to participate, facilitating continuous innovation and progress in mobile robot studies.

II. RELATED WORK

Scene understanding is one of the most fundamental topics in robot environment perception and includes many common tasks in computer vision, such as: 3D object detection, object classification, semantic segmentation, pose estimation, spatial layout estimation, CAD model retrieval and alignment, etc. With the development of more lightweight and low-cost RGB-D sensors, personal users can get convenient access to images and depth data of indoor scenes. Therefore, many related works on RGB-D datasets have emerged to meet the evolving needs of robot perception algorithms.

Some popular datasets are summarized in Table I. Based on their annotation types, those RGB-D datasets are coarsely divided into two categories: annotated in 2D domain and annotated in 3D domain.

Since effective and intensive 3D annotations is hard to acquire, some works label RGB-D images with 2D annotations, which indirectly serve as ground truth for 3D tasks. Berkeley 3D Object Dataset [1] has 2D bounding box annotations on RGB-D images, NYU Depth v2 [2] includes 2D semantic segmentation from short RGB-D videos with 1449 selected

frames tagged, SUN3D [3] dataset is composed of 415 RGB-D video sequences in 254 scenes. Stanford 2D-3D-Semantics dataset[4] utilizes the iGibson simulation environment to provide large scale virtual scenes with 2D texture, geometric as well as semantic information. SceneNet RGB-D dataset [5] contains 5000k images and different kinds of 2D annotations.

3D annotation is challenging, yet some works have still made significant contributions. SUN RGB-D [6] contains 10,335 RGB-D images with dense 2D/3D annotations, including 2D polygons, 3D bounding boxes with accurate object orientation and 3D room layout. ScanNet [7] contains a total of 1,513 video sequences with annotations of 3D camera poses, surface reconstruction and semantic segmentation, and partially aligned CAD models. SUN-CG [8] provides 45k virtual scene layouts and 500k rendered images with single-view RGB, depth maps and semantic segmentation maps. Matterport 3D [9] contains 194,400 RGB-D images to generate panoramas with surface reconstruction, camera position and 2D/3D semantic segmentation annotations. InteriorNet [10] is rendered entirely in virtual home scenes and contains 15k sequences. ARKitScenes [11] improves the resolution of ground truth geometry from laser scans. ScanNet++ [12] is a new dataset that contains 460 high-resolution 3D reconstructions of indoor scenes with dense semantic and instance annotations.

Drawing references from both 2D and 3D annotation works, current datasets lack dynamic objects and pedestrians, which are highly common in real-world scenes[13]–[20] and pose sufficient challenges for robot-related research. These datasets commonly exhibit issues such as poor data quality, small annotation volumes, limited label types, noisy annotations, and unreasonable virtual scene layouts. Current research often requires testing, training, and deployment on real hardware in virtual scenarios, which the current datasets cannot fully satisfy.

A. Contributions and Advantages of THUD

- Our dataset provides data annotated with dynamic instances for large-scale indoor scenes, which is closer

to mobile robots' real working environment, posing significant challenges to dynamic robot tasks.

- Our dataset supports training and testing for various robotic scene understanding tasks (object detection, semantic segmentation, robot relocalization, scene reconstruction, etc.)
- Our dataset contains both real and synthetic annotated data, which can satisfy the testing of mobile robot algorithms in different scenes.

III. DATASET CONSTRUCTION

A. Data Acquisition

1) Real Data Acquisition:

Acquisition platform: Data collection relies on the PUDUbot2 & Kinect V2 joint collection platform, as depicted in Fig. 2 PUDUbot2 is an advanced mobile service robot that, when in developer mode, can record real-time robot poses (translation distance and Euler angles). The depth and RGB images of real data are collected through the Kinect V2.

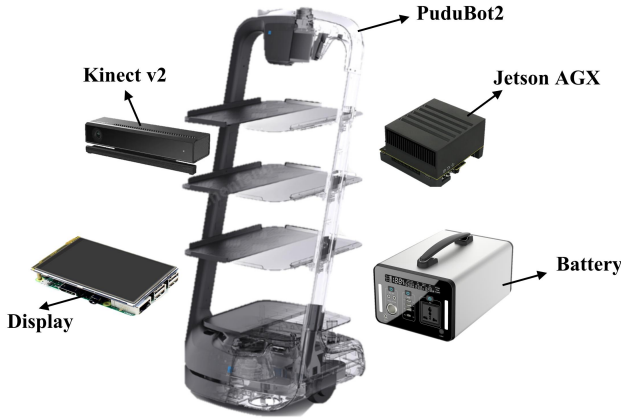


Fig. 2: PUDUbot2 & Kinect V2 joint collection platform

Acquisition scenarios: The data collection for this dataset is primarily focused on the field of mobile robotics. In order to better serve this domain, the real-world data was collected in common service robot environments on the campus of Tsinghua University. As shown in Fig.1 column 2nd. These include eight scenes: a corridor of laboratory, the lobby of teaching building, department meeting rooms, departmental visiting laboratories, a cafeteria, a campus dining hall, elevator area, and shops.

One of the unique features of THUD is its dynamic nature. To showcase different levels of dynamic complexity in the real-world scenarios, data collection was conducted at various time intervals. For example, Fig. 3 illustrates the data collection process in the canteen scene at different time periods. Within different data sequences in the same scene, there are sequences with low, moderate and high density of pedestrians. These sequences with varying levels of dynamic complexity in the same environment will help users evaluate the algorithm's capabilities and better align with the real-world situations that service robots encounter at different

times. Another notable feature of this dataset is the inclusion of various challenging labels, such as elevators, glass objects, and other item categories. These items have a significant impact on robot perception during actual testing.

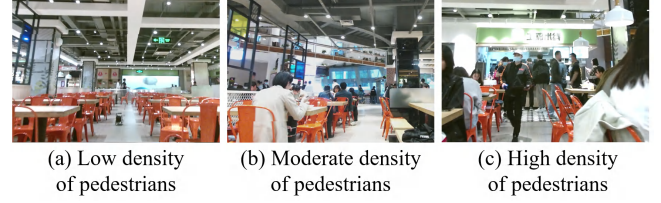


Fig. 3: Scenes with varying levels of dynamic complexity

Acquisition method: The real data collection system consists of two independent systems: pose acquisition and image acquisition. The acquisition of robot pose data primarily relies on the v-SLAM localization at the top of PUDUbot2, supplemented by the odometry of the steering wheel. The data is computed at a frame rate of 40fps.

The image data includes depth and RGB images. We have developed ROS scripts for synchronous acquisition of RGB and depth images. The depth images in the released version have been denoised using the Self-Supervised Deep Depth Denoising method [21]. Due to the inherent instability of Kinect V2, the actual frame rate for capturing RGB and depth images fluctuates between 15-30fps.

2) Synthetic Data Acquisition:

Acquisition platform: The platform is designed to simulate the realistic working environment for mobile robots, allowing for the collection of virtual sensor data such as RGB images, depth maps, robot pose, and IMU data. The platform has been developed using open source tools from Unity3D and has undergone secondary development to meet the specific requirements of the synthetic data acquisition for mobile robots. The Unity3D physics engine is utilized to simulate the physical movements of the mobile robot, and simultaneously publish the collected synthetic data and labels through ROS2. The platform architecture is shown in Fig. 4. The ultimate aim is to provide a comprehensive synthetic data acquisition platform that can be used to train and test mobile robot indoor scenes understanding algorithms.

Acquisition scenarios: To acquire more realistic and effective data, the scenes were designed based on the actual working environment of mobile robots, taking into account two main aspects: dynamic scenarios with moving obstacles and special scenarios that pose potential hazards for robots in the reality. As illustrated in Fig. 5. Dynamic scenarios were created by setting up pedestrians in the scenes, some special cases were also considered, such as running children, people pushing shopping carts, other moving robots in the scene. The scenes also take into account pedestrians of different ages and clothing styles for men, women, and children. For special scenarios, the scene settings focus on realistic objects that tend to pose difficulties for robotic tasks, such as stairs, windows, glass doors, etc.

Acquisition method: The synthetic data is generated by creating virtual sensors in Unity3D scenes. Based on the type

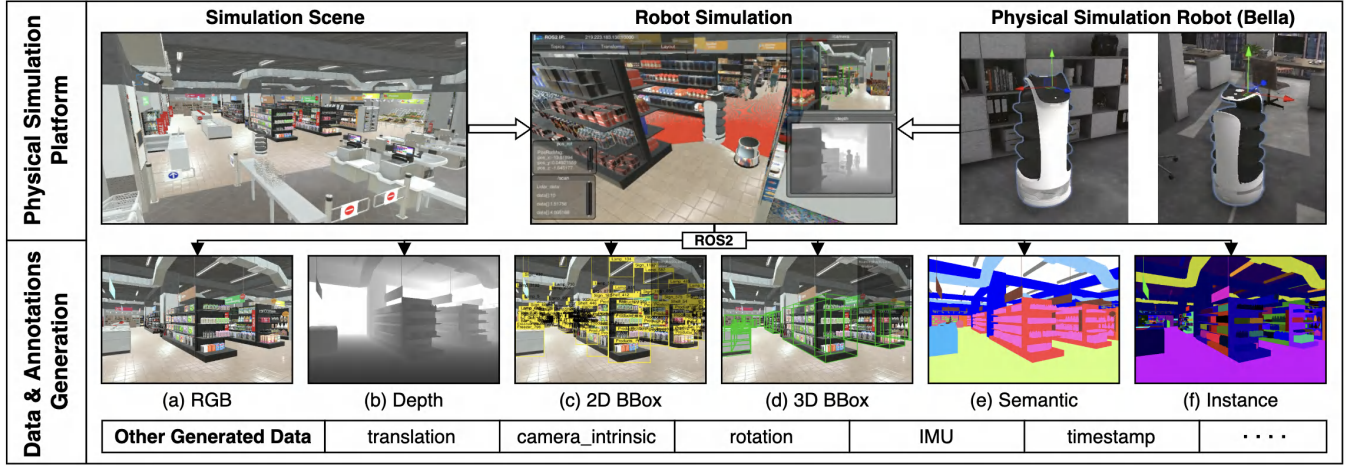


Fig. 4: Mobile robot synthetic data acquisition platform.



Fig. 5: Dynamic and special scenarios

and position of sensors on the real mobile robot, the camera for the physical simulation robot is configured accordingly, with a height of 1.2m and a pitch angle of 0 degrees. As the robot moves through the scenes, *RGB images*, *depth maps*, *robot poses*, and *IMU* data are captured. The resolution of the RGB images and depth maps is 730×530 , and the robot pose is represented by its *xyz* coordinates in the scene and its rotation angle around the *z*-axis.

B. Ground Truth Annotation

1) *Synthetic Data Annotation*: The annotation of the simulated data was automatically generated through our custom-built data acquisition platform. We integrated the robot model of *Bella* into the virtual scenes for data collection, enabling the robot to navigate and capture data from its own perspective. During the data collection process, we simultaneously extracted vital information, including 2D and 3D bounding boxes, semantic and instance segmentation images, robot's pose and IMU information. These annotations were then synchronized with the captured RGB images and depth maps, ensuring a one-to-one correspondence.

Importantly, the annotation information extracted within our data acquisition platform represents ground truth within the virtual scenes. This ensures a high level of precision and accuracy in the generated annotations, allowing us to

rapidly generate large quantities of meticulously annotated data within a short time.

2) *Real Data Annotation*: To annotate the real-world data with 2D and 3D object information, we first preprocessed RGB images and depth maps by applying automated algorithms to predict the 2D&3D positions and bounding boxes of objects, such as Faster-RCNN[22] and Votenet[23]. However, the algorithms for 3D objects detection are not very effective. Subsequently, we performed manual review and verification to correct and refine the annotated bounding boxes, ensuring the accuracy and completeness of the detected objects. For the 2D semantic and instance segmentation task, we similarly employed a semi-automated annotation method, manual review and correction were also conducted by annotators to ensure accurate pixel-wise classification and semantic labeling.

In summary, our dataset construction process leveraged a combination of automated and semi-automated annotation methods to enhance efficiency and accuracy in data annotation. For providing high-quality training and evaluation data for robots indoor scene understanding tasks.

C. Data Statistics

Our dataset comprises a total of 90,175 annotated frames, consisting of 84,984 frames from synthetic data collection and 5,191 frames from real-world data collection. Each frame has undergone intensive annotation, resulting in a total of over 20M labels, with over 1.2M labels for dynamic objects such as pedestrians, robots, and shopping carts. On average, each frame contains 176 data labels.

These labels encompass four annotation types: 2D bounding boxes, 3D bounding boxes, semantic segmentation, and instance segmentation, with their respective proportions as shown in Fig. 6(b). The dataset encompasses 8 real and 5 synthetic large-scale indoor scenes, with average area exceeding $300m^2$. There are 91 different object categories within these scenes, and their distribution, along with the counts of each annotation type, is illustrated in Fig. 6(a)(c).

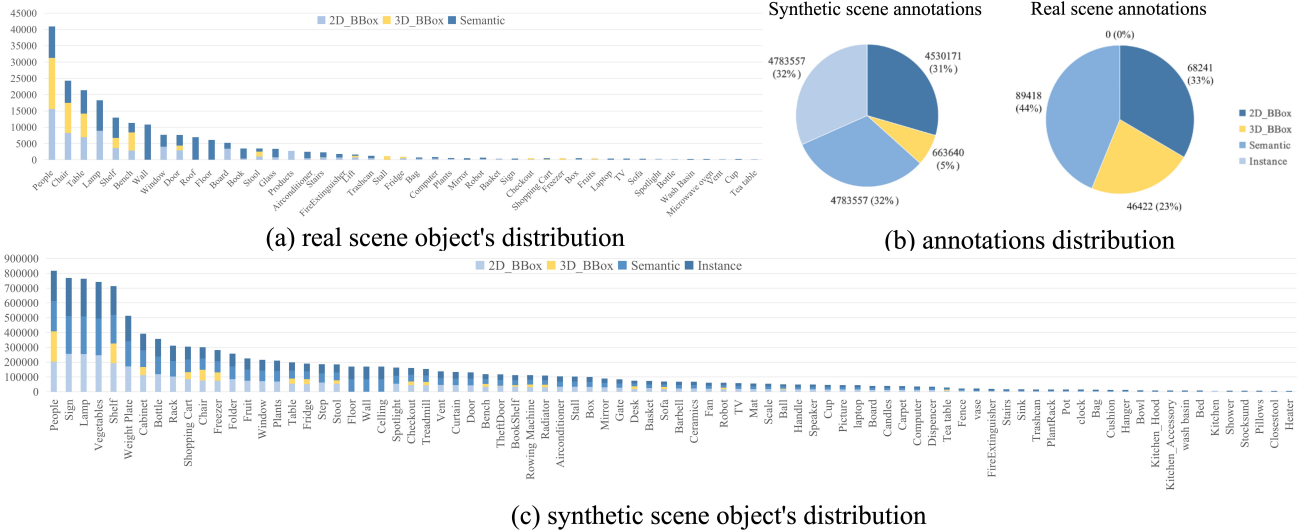


Fig. 6: Statistics of annotations in our dataset.

IV. EVALUATION ON THUD ROBOTIC DATASET

To verify whether our dynamic dataset is necessary and useful for mobile robots, three tasks, i.e. 3D object detection, semantic segmentation, and robot relocalization, are tested and analyzed. Important results and conclusion are given here and their analysis are elaborated in *project website*.

A. 3D Object Detection

3D object detection methods estimate a 3D bounding box and a pose of each object presented in a scene [26]–[28]. To compare the performance difference, three representative indoor 3D object detection algorithms which have good performance on both dynamic and static objects, i.e. F-PointNet [24], ImVoteNet [23] and DeMF [25], are selected. These algorithms were evaluated on the real-world canteen scene and the synthetic supermarket scene to assess their mean Average Precision (mAP) for static and dynamic objects. The experimental results are depicted in Table II. In the supermarket scene, the selected 5 static objects include *Chair*, *Table*, *Shelf*, *Cabinet*, *Fridge*. While the 3 dynamic objects encompass *Shopping cart*, *Robot*, *People*. In the canteen scene, the 7 static objects are *Chair*, *Table*, *Shelf*, *Bench*, *Door*, *Stairs*, *Stool*, and the dynamic object is only *People*. Moreover, we used the average number of dynamic objects per frame as a metric to assess the level of scene dynamic complexity, the supermarket and canteen scene dynamic complexity is 0.94 and 3.34, respectively.

TABLE II: Experiments on 3D Object Detection

Scene	Method	Dynamic Objs (mAP)	Static Objs (mAP)
Supermarket	F-PointNet [24]	7.89	8.92
	ImVoteNet [23]	17.49	17.29
	DeMF [25]	34.51	38.24
Canteen	F-PointNet [24]	18.16	36.67
	ImVoteNet [23]	26.72	43.37
	DeMF [25]	28.56	45.43

We can see two important conclusion from the test results in Table II. First, the results indicate that different algorithms have varying degrees of decrease in accuracy when detecting dynamic objects as compared to static objects. Second, it can be observed that having a larger density of dynamic objects lead to more significant accuracy difference between static and dynamic objects. But this could also be influenced by the object categories.

B. Semantic Segmentation

Semantic segmentation methods assign each pixel in an image to its corresponding semantic category, focusing on the geometric structure and differentiation of different objects [33]–[35]. However, robot indoor semantic segmentation faces several challenges, e.g. complex indoor scenes, occlusions, lighting variations, as well as the diversity of object shapes and sizes. Four RGB-D semantic segmentation algorithms, i.e. ACNet [29], RedNet [30], ESANet [31], and SA-Gate [32], are selected and tested on the aforementioned synthetic supermarket scene and real-world canteen scene. 31 and 19 static and dynamic objects with semantic labels are added to the supermarket and canteen scene, respectively. Objects without labels are set to the void value with corresponding pixels (0, 0, 0).

The accuracy of different semantic segmentation methods is measured by calculating the MIOU (Mean Intersection over Union).

TABLE III: Experiments on Semantic Segmentation

Scene	Method	Backbone	MIOU(%)
Supermarket	ACNet [29]	3xR50	74.83
	RedNet [30]	2xR34	76.92
	ESANet [31]	2xR34	78.42
	SA-Gate [32]	2xR101	83.19
Canteen	ACNet [29]	3xR50	51.85
	RedNet [30]	2xR34	59.83
	ESANet [31]	2xR34	65.97
	SA-Gate [32]	2xR101	58.34

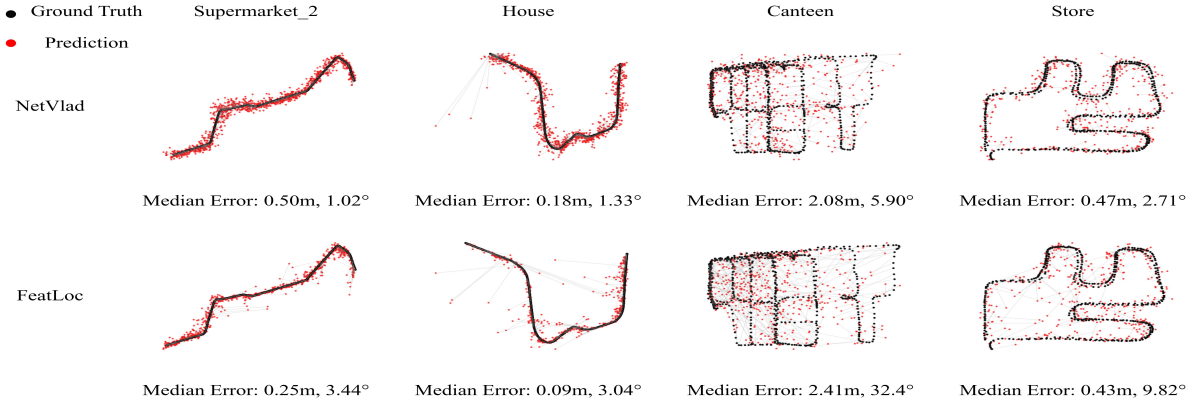


Fig. 7: Test result visualization for robot relocalization

Union) during testing and the results are shown in Table III. We can obtain three conclusions. First, the accuracy on real-world canteen scene is much lower than synthetic supermarket scene. Second, the MIOU results are comparable to their original published results, though tested on different datasets. Third, to obtain a quantitative comparison on dynamic objects, we also train and test the ESANet [31] algorithm in the supermarket scene with and without dynamic objects, the MIOU of the test is 78.42% and 79.63% respectively. There is no significant difference, which is coincident with our intuition. In addition, as shown in Fig. 8, different methods still have varying segmentation ability on dynamic object, such as the people and shopping carts.

C. Robot Relocalization

Robot relocalization refers to the process of accurately determines and adjusts its position based on the environmental perception information and its own localization capabilities, during its movement [36]–[38]. Existing methods can be coarsely classified into global- and local-feature based methods.

We conducted extensive experiments on THUD using a representative global- and local-feature based methods, i.e. NetVlad [39] and FeatLoc [40], respectively. We randomly selected two sequences as testing and the results on both synthetic and real-world scene are visualized in Fig. 7.

In addition, to analyze the effect of the number of dynamic objects (primarily pedestrians) in various scene images, we define a metric to represent a scene’s dynamic complexity, computed as the average number of dynamic pedestrians per frame across the entire dataset. As shown in Fig. 9, with

the scene’s dynamic complexity increasing, the methods’ accuracy decreases and FeatLoc [40] has a higher drop in accuracy compared to NetVlad [39], attributed to dynamic objects introducing disturbances and noise in the extraction of scene image features.

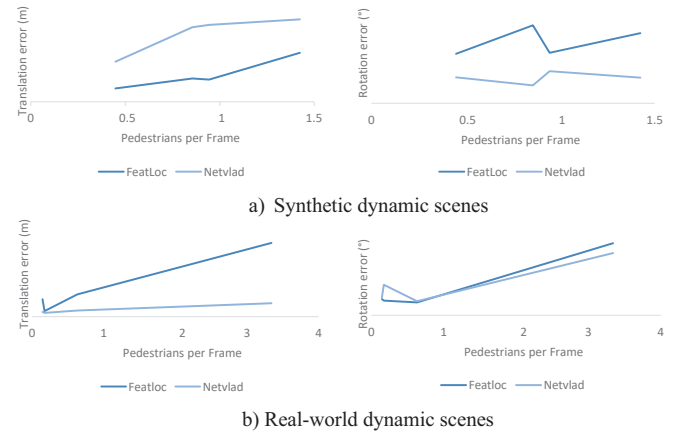


Fig. 9: Trans. & rot. error with different dynamic complexity

V. CONCLUSIONS

In this paper, we introduced a mobile robot oriented large-scale indoor dataset for dynamic scene understand tasks. It consists of both real and synthetic data to support both static and dynamic robotic task. In the near future, we are continuously expanding and iterating this dataset, to cover more robot dynamic tasks, e.g. robot navigation, object tracking, trajectory prediction, etc. The dataset initial version is shared at: <https://jackyzengl.github.io/THUD-Robotic-Dataset.github.io/>.

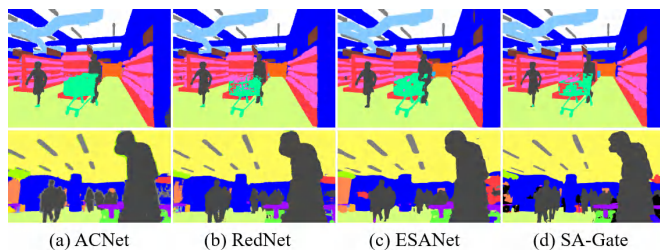


Fig. 8: Comparison of semantic segmentation methods

REFERENCES

- [1] A. Janoch, S. Karayev, Y. Jia, *et al.*, “A category-level 3-d object dataset: Putting the kinect to work,” in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1168–1174. DOI: 10.1109/ICCVW.2011.6130382.
- [2] P. K. Nathan Silberman Derek Hoiem and R. Fergus, “Indoor segmentation and support inference from rgb-d images,” in *ECCV*, 2012.
- [3] J. Xiao, A. Owens, and A. Torralba, “Sun3d: A database of big spaces reconstructed using sfm and object labels,” in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 1625–1632.
- [4] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, “Joint 2d-3d-semantic data for indoor scene understanding,” *ArXiv*, vol. abs/1702.01105, 2017.
- [5] J. McCormac, A. Handa, S. Leutenegger, and A. J. Davison, “Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth,” *ArXiv*, vol. abs/1612.05079, 2016.
- [6] S. Song, S. P. Lichtenberg, and J. Xiao, “Sun rgb-d: A rgb-d scene understanding benchmark suite,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 567–576. DOI: 10.1109/CVPR.2015.7298655.
- [7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3d reconstructions of indoor scenes,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2432–2443. DOI: 10.1109/CVPR.2017.261.
- [8] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, “Semantic scene completion from a single depth image,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 190–198. DOI: 10.1109/CVPR.2017.28.
- [9] A. X. Chang, A. Dai, T. A. Funkhouser, *et al.*, “Matterport3d: Learning from rgb-d data in indoor environments,” *2017 International Conference on 3D Vision (3DV)*, pp. 667–676, 2017.
- [10] W. Li, S. Saeedi, J. McCormac, *et al.*, “InteriorNet: Mega-scale multi-sensor photo-realistic indoor scenes dataset,” *ArXiv*, vol. abs/1809.00716, 2018.
- [11] A. Dehghan, G. Baruch, Z. Chen, *et al.*, “ArkitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data,” in *NeurIPS Datasets and Benchmarks*, 2021.
- [12] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, “ScanNet++: A high-fidelity dataset of 3d indoor scenes,” *ArXiv*, vol. abs/2308.11417, 2023.
- [13] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, *ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes*, en, *arXiv:1702.04405 [cs]*, Apr. 2017.
- [14] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, *Joint 2D-3D-Semantic Data for Indoor Scene Understanding*, en, *arXiv:1702.01105 [cs]*, Apr. 2017.
- [15] B.-S. Hua, Q.-H. Pham, D. T. Nguyen, M.-K. Tran, L.-F. Yu, and S.-K. Yeung, “SceneNN: A Scene Meshes Dataset with aNNotations,” en, in *2016 Fourth International Conference on 3D Vision (3DV)*, Stanford, CA, USA: IEEE, Oct. 2016, pp. 92–101, ISBN: 978-1-5090-5407-7. DOI: 10.1109/3DV.2016.18.
- [16] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, “A benchmark for rgb-d visual odometry, 3d reconstruction and slam,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1524–1531. DOI: 10.1109/ICRA.2014.6907054.
- [17] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, *SUNCG? Semantic Scene Completion from a Single Depth Image*, en, *arXiv:1611.08974 [cs]*, Nov. 2016.
- [18] S. Song, S. P. Lichtenberg, and J. Xiao, “SUN RGB-D: A RGB-D scene understanding benchmark suite,” en, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, Jun. 2015, pp. 567–576, ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298655.
- [19] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 573–580. DOI: 10.1109/IROS.2012.6385773.
- [20] J. Valentin, A. Dai, M. Niessner, *et al.*, “Learning to navigate the energy landscape,” in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 323–332. DOI: 10.1109/3DV.2016.41.
- [21] V. Sterzentsenko, L. Saroglou, A. Chatzitofis, *et al.*, “Self-supervised deep depth denoising,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019.
- [22] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017. DOI: 10.1109/TPAMI.2016.2577031.
- [23] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas, “Imvotenet: Boosting 3d object detection in point clouds with image votes,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4403–4412. DOI: 10.1109/CVPR42600.2020.00446.
- [24] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum pointnets for 3d object detection from rgb-d data,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927. DOI: 10.1109/CVPR.2018.00102.
- [25] H. Yang, C. Shi, Y. Chen, and L. Wang, “Boosting 3d object detection via object-focused image fusion,” *ArXiv*, vol. abs/2207.10589, 2022.

- [26] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *ArXiv*, vol. abs/2010.04159, 2020.
- [27] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, *TokenFusion:multimodal token fusion for vision transformers*, Jul. 15, 2022. arXiv: 2204.08721[cs].
- [28] Z. Wang and K. Jia, *Frustum ConvNet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection*, Aug. 14, 2019. arXiv: 1903.01864[cs].
- [29] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgb-d semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1440–1444.
- [30] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation," *ArXiv*, vol. abs/1806.01054, 2018.
- [31] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Groß, "Efficient rgb-d semantic segmentation for indoor scene analysis," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13 525–13 531, 2020.
- [32] X. Chen, K.-Y. Lin, J. Wang, *et al.*, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *European Conference on Computer Vision*, 2020.
- [33] F. Cao and Q. Bao, "A survey on image semantic segmentation methods with convolutional neural network," in *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2020, pp. 458–462.
- [34] Y. Hu, Z. Chen, and W. Lin, "Rgb-d semantic segmentation: A review," in *2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2018, pp. 1–6.
- [35] C. Wang, C. Wang, W. Li, and H. Wang, "A brief survey on rgb-d semantic segmentation using deep learning," *Displays*, vol. 70, p. 102 080, 2021.
- [36] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera re-localization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.
- [37] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 883–890.
- [38] S. Chen, X. Li, Z. Wang, and V. A. Prisacariu, "Dfnet: Enhance absolute pose regression with direct feature matching," in *European Conference on Computer Vision*, Springer, 2022, pp. 1–17.
- [39] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, 2018. DOI: 10.1109/TPAMI.2017.2711011.
- [40] T. B. Bach, T. T. Dinh, and J. H. Lee, "Featloc: Absolute pose regressor for indoor 2d sparse features with simplistic view synthesizing," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022.