TO-Scene: A Large-scale Dataset for Understanding 3D Tabletop Scenes

Mutian Xu[1*], Pei Chen[1*], Haolin Liu[1,2], and Xiaoguang Han[1,2†]

[1] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen
[2] The Future Network of Intelligence Institute, CUHK-Shenzhen
{mutianxu, peichen, haolinliu}@link.cuhk.edu.cn,
{hanxiaoguang}@cuhk.edu.cn

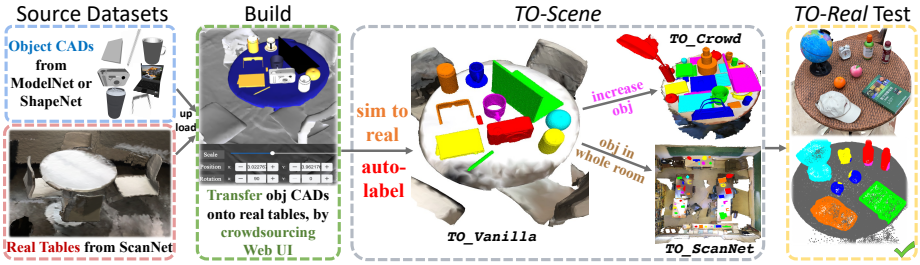


Fig. 1: Overview of data acquisition framework and TO-Scene dataset. We firstly transfer object CADs from ModelNet [43] and ShapeNet [5] onto real tables from ScanNet [10] via crowdsourcing Web UI. Then the tabletop scenes are simluated into real scans and annotated automatically. Three variants (TO_Vanilla, TO_Crowd and TO_ScanNet) are presented for various scenarios. A real-scanned test data TO-Real is provided to verify the practical value of TO-Scene.

Abstract. Many basic indoor activities such as eating or writing are always conducted upon different *tabletops* (e.g., coffee tables, writing desks). It is indispensable to understanding tabletop scenes in 3D indoor scene parsing applications. Unfortunately, it is hard to meet this demand by directly deploying data-driven algorithms, since 3D tabletop scenes are rarely available in current datasets. To remedy this defect, we introduce TO-Scene, a large-scale dataset focusing on **t**ablet**o**p scenes, which contains 20,740 scenes with three variants. To acquire the data, we design an efficient and scalable framework, where a crowdsourcing UI is developed to transfer CAD objects from ModelNet [43] and ShapeNet [5] onto tables from ScanNet [10], then the output tabletop scenes are simulated into real scans and annotated automatically.
Further, a tabletop-aware learning strategy is proposed for better perceiving the small-sized tabletop instances. Notably, we also provide a *real*

* M. Xu and P. Chen contribute equally.
† Corresponding author.

scanned test set TO-Real to verify the practical value of TO-Scene. Experiments show that the algorithms trained on TO-Scene indeed work on the realistic test data, and our proposed tabletop-aware learning strategy greatly improves the state-of-the-art results on both 3D semantic segmentation and object detection tasks. Dataset and code are available at https://github.com/GAP-LAB-CUHK-SZ/TO-Scene.

Keywords: 3D tabletop scenes, efficient, three variants, tabletop-aware learning

1 Introduction

Understanding indoor scenes is a fundamental problem in many industrial applications, such as home automation, scene modeling, virtual reality, and perception assistance. While this topic spawns the recent development of 3D supervised deep learning methods [47,52,27], their performance directly depends on the availability of large labeled training datasets. Thanks to the progress of 3D scanning technologies and depth sensors (e.g., Kinect [51]), various 3D indoor scene datasets arised [36,35,46,38,21,34,2,10,4]. The most popular indoor scene benchmark among them, ScanNet [10], is consisted of richly annotated RGB-D scans in real world, which is produced by a scalable data acquisition framework.

Albeit the great advance on 3D indoor datasets allows us to train data-hungry algorithms for scene understanding tasks, one of the most frequent and widely-used setups under indoor environments is poorly investigated – the scene focusing on *tabletops*.

In indoor rooms, for satisfying the basic daily requirements (such as eating, writing, working), humans (or robots) need to frequently interact with or face to different tabletops (e.g., dining tables, coffee tables, kitchen cabinets, writing desks), and place, catch or use various tabletop objects (e.g., pencils, keyboards, mugs, phones, bowls). Thus, perceiving and understanding tabletop scenes is *indispensable* in indoor scene parsing applications. Unfortunately, it is hard to meet this demand by directly deploying the 3D networks, since existing indoor scene datasets lack either adequate samples, categories or annotations of tabletop objects (illustrated in Table 1), from where the models is not able to learn the corresponding representations. Therefore, it is substantially meaningful to build a dataset focusing on the tabletop objects with sufficient quantities and classes.

We attempt to construct such a dataset and present our work by answering the below questions:

(**1**) *Assumption – What is the sample pattern supposed to be?* Starting from our previous motivation, the dataset is expected to meet the practical demand in real applications. Thus, the sample is assumed to look similar with the real scanned tabletop scene, which is a bounded space filled by a table with multiple objects above it, and surrounded by some background furniture. This requires the model to perceive both the individual objects and their inter-relationships, while assisted by the context information learned from finite indoor surroundings.

(**2**) *Acquisition – How to build such a dataset with decent richness, diversity, and scalability, at a low cost?* Building large-scale datasets is always challenging, not only because of the laborious collection of large amounts of 3D data, but also the non-trivial annotation. As for our work, it is undoubtedly burdensome to manually place objects above real tables, then scan and label 3D data. Instead, we design an efficient and scalable framework (Fig. 1) to overcome this difficulty. We firstly develop a Web UI (Fig. 2) to help novices transfer CAD objects of ModelNet [43] and ShapeNet [5] to suitable tables extracted from ScanNet [10] scenes. The UI makes it possible to enlarge our dataset in the future. After that, we simulate the synthetic objects into real-world data, expecting that the model trained on our dataset can work on real scanned data. Last, an automatic annotation strategy is adopted to produce point-wise semantic labels on each reconstructed object's meshes, based on its bounding box that directly gained from the CAD model. So far, the complete acquisition pipeline enables us to construct a vanilla dataset as mentioned in (**1**), which contains 12,078 tabletop scenes with 60,174 tabletop instances belonging to 52 classes, called **TO_Vanilla**.

(**3**) *Enrichment – Can we create more variants of the data to bring new challenges into the indoor scene parsing task?* i) In our daily life, the tabletops are sometimes full of crowded objects. To simulate this situation, we increase the instances above each table, producing a more challenging setup called **TO_Crowd**, which provides 3,999 tabletop scenes and 52,055 instances that are distinguished with TO_Vanilla. ii) Both TO_Vanilla and TO_Crowd assume to parse the tabletop scenes. Nevertheless, some real applications require to parse the *whole* room with all furniture including tabletop objects in one stage. To remedy this, another variants **TO_ScanNet** comes by directly using the tables in TO-Vanilla, but the complete scans of rooms that accommodate the corresponding tables are still kept. It covers 4663 scans holding around 137k tabletop instances, which can be treated as an *augmented ScanNet* [10]. Combining three variants, we introduce **TO-Scene**.

(**4**) *Strategies – How to handle the open challenges in our TO-Scene?* The tabletop objects in TO-Scene are mostly in smaller-size compared with other large-size background furniture, causing challenges to discriminate them. To better perceive the presence of tabletop instances, we propose a tabletop-aware learning strategy that can significantly improve upon the state-of-the-art results on our dataset, by jointly optimizing a tabletop-object discriminator and the main segmentation or detection targets in a single neural network.

(**5**) *Practicality – Can TO-Scene indeed serve for real applications?* To investigate this, we manually scan and annotate three sets of data, that corresponds to the three variants of TO-Scene. We denote the whole test data as **TO-Real**, which provides 197 *real* tabletop scans with 2373 objects and 22 indoor room scans holding 824 instances. Consequently, the models trained on TO-Scene get promising results on our realistic test data, which suggests the practical value of TO-Scene.

Here, the contributions of this paper are summarized as:

Table 1: Overview of 3D indoor scene datasets. "kit" indicates kitchen, "obj" denotes object, "bbox" means bounding boxes, "point segs" is point-wise segmentation. Our large-scale TO-Scene with three variants focuses on tabletop scenes, and is efficiently built by crowdsourcing UI and automatic annotation.

Dataset	Tabletop	♯Scenes	Collection	Annotation
SUNCG [39]	✗	45k	synthetic, by designers	dense 3D
SceneNet [19]	✗	57	synthetic, by designers	dense 3D
OpenRooms [25]	✗	1068	synthetic, by designers	dense 3D
3D-FRONT [14]	✗	19k	synthetic, by designers	dense 3D
NYU v2 [36]	✗	464	scan, by experts	raw RGB-D [41]
SUN 3D [46]	✗	415	scan, by experts	2D polygons
S3DIS [2]	✗	265	scan, by experts	dense 3D [17]
ScanNet [10]	✗	1513	scan, by crowdsourcing UI [10]	dense 3D
WRGB-D [23]	✓, 5 sorts small obj	22	scan, by experts	point segs, 2D polygons
GMU Kit [16]	✓, 23 sorts kit obj	9	scan, by experts	dense 3D
TO-Scene (ours)	**✓, various tables, 52 sorts obj**	**21k, 3 variants with augmented [10]**	**effortless transfer** by our-own **crowdsourcing UI**	**dense 3D: bboxes of [5,43] + auto point segs**

- TO-Scene – To the best of our knowledge, the first large-scale dataset primarily for understanding tabletop scenes, with three different variants.
- An efficient and scalable data acquisition framework with an easy-to-use Web UI for untrained users.
- A tabletop-aware learning strategy, for better discriminating the small-sized tabletop instances in indoor rooms.
- A real scanned test set – TO-Real, with the same three variants as TO-Scene, for verifying the practical value of TO-Scene.
- Experiments demonstrate that the networks running on TO-Scene work well on TO-Real, and our proposed tabletop-aware learning strategy greatly improves the state-of-the-arts.
- TO-Scene and TO-Real, plus Web UI are all open source.

2 Related Work

3D indoor scene datasets. 3D indoor scene datasets have been actively made over the past few years. NYU v2 [36] is an early real dataset for RGBD scene understanding, which contains 464 short RGB-D videos captured from 1449 frames, with 2D polygons annotations as LabelMe [41] system. SUN3D [46] captures a set of 415 sequences of 254 spaces, with 2D polygons annotation on key frames. The following SUN RGB-D [38] collects 10,335 RGB-D frames with diverse scenes. Yet it does not provide complete 3D surface reconstructions or dense 3D semantic segmentations. To remedy this, Hua et al. [21] introduce SceneNN, a larger

scale RGB-D dataset consisting of 100 scenes. Another popular dataset S3DIS [2] includes manually labeled 3D meshes for 265 rooms captured with a Matterport camera. Later, Dai et al. [10] design an easy-to-use and scalable RGB-D capture system to produce ScanNet, the most widely-used and richly annotated indoor scene dataset, which contains 1513 RGB-D scans of over 707 indoor rooms with estimated camera parameters, surface reconstructions, textured meshes, semantic segmentations. Further, Matterport3D [4] provides 10,800 panoramic views from 194,400 RGB-D images of 90 building scenes. In [39,19,25,14], the synthetic 3D indoor scene data are generated. The recent 3D-FRONT [14] contains 18,797 rooms diversely furnished by 3D objects, surpassing all public scene datasets.

The aforementioned datasets ignore an important data form in indoor scene parsing applications – *tabletop* scenes, which is the basis of our dataset. There are two existing datasets emphasizing small objects that may appear on tables. WRGB-D Scenes [23] includes 22 annotated scene video sequences containing 5 sorts of small objects (subset of WRGB-D Object [23]). GMU kitchen [16] is comprised of multi-view kitchen counter-top scenes, each containing 10-15 hand-hold instances annotated with bounding boxes and in the 3D point cloud. However, their complex data collections and annotations cause the severe limitation on the data richness, diversity and scalability (see Table 1).

3D shape and object datasets. The tabletop objects in TO-Scene are originated from ModelNet [43] containing 151,128 3D CAD models of 660 categories, and ShapeNetCore [5] covering 55 object classes with 51,300 3D models. As for 3D shape datasets, [5,43,22,3,44,45] provide CAD models, while [6,37,29,42,8] advocate the realistic data. [23,33] contain multi-view images of 3D objects, and the recent Objectron [1] is a collection of short object-centric videos. The aforesaid datasets deal with single 3D objects. In contrast, our TO-Scene highlights the holistic understanding and relationship exploration on various tabletop objects under indoor room environments.

Robot grasping and interacting datasets. [24,12,13] contribute to the grasping of tabletop objects. Their annotations of object 6D poses and grasp poses exclude object categories, which are customized for the robot grasping, instead of understanding tabletop scenes. Garcia et al. [15] introduce a hyperrealistic indoor scene dataset, which is explored by robots interacting with objects in a simulated world, but still specifically for robotic vision tasks.

3D indoor scene parsing algorithms. The bloom of the indoor scene datasets opens the chances for training and benchmarking deep neural models in different 3D scene understanding tasks, such as 3D semantic segmentation [32,48,47,52] and 3D object detection [30,50,27]. Our TO-Scene raises a new challenge of discriminating the small-sized tabletop objects in indoor rooms. To tackle this, we propose a tabletop-aware learning strategy with a jointly optimized tabletop-object discriminator, for better perceiving tabletop instances.

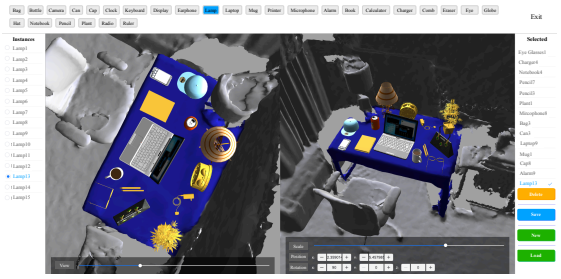


Fig. 2: Our web-based crowdsourcing interface for transferring CADs onto real tabletops. The user chooses an suitable object and click somewhere on Bird's-Eye-View (BEV) (left) for placing it above 3D tables (right).

3 Data Acquisition

Building large-scale 3D datasets is always challenging, not only because of the laborious collection of large amounts of 3D data, but also the non-trivial annotation. This problem is especially severe even for expert users when collecting cluttered tabletop scenes under indoor environments, as in [16,23], which makes it non-trivial to replenish the dataset (i.e., inferior scalablity), and limits the richness and diversity of the dataset.

To solve this issue, we present an efficient and scalable framework for creating TO-Scene, with the ultimate goal of driving the deep networks to have a significant effect on the real world 3D data.

3.1 Transfer Object CADs into Indoor Scenes

In this work, rather than manually placing and scanning the tabletop objects in the real world, we propose to transfer the object CADs from ModelNet [43] and ShapeNet [5] into the tables located in ScanNet [10] rooms.

Source datasets. ShapeNetCore [5] contains 55 object classes with 51,300 3D models and organizes them under the WordNet [28] taxonomy. Despite the tremendous size of ShapeNetCore, its quantity of tabletop objects are not able to meet our requirement. Thus, we also employ ModelNet [43] covering 151,128 3D CAD models of 660 classes. Both of them have the annotations of consistent rigid alignments and physical sizes. Leveraging their richness, we are able to agreeably borrow a large amount of CAD instances, belonging to 52 classes (detailed in Fig. 3) that are commonly seen on different tables in our daily life, as the tabletop objects of TO-Scene.

Another important source is ScanNet [10], which is a richly annotated real-world indoor scene dataset including 2.5M RGB-D images in 1513 scans acquired in 707 distinct spaces, especially covering diverse table instances. We extract the tables from ScanNet to place tabletop objects. Additionally,

since the scanning of a tabletop scenes in real applications will also cover some background furniture, the ScanNet indoor environments around the tables also act as the context in our tabletop scenes.

User interface. For allowing untrained users to create large-scale datasets, we develop a trivial-to-use Web UI to ease the object transfer process. As shown in Fig. 2, the Web UI shows up with a BEV (Bird's-Eye-View) of a table on the left and a 3D view of the surroundings on the right. The table are randomly selected from ScanNet with CADs picked from ModelNet and ShapeNet each time. The operation is friendly for untrained users. Specifically, when placing an object, the user does *not* need to perceive the *height* for placing it in 3D. The only action is just clicking at somewhere on the table in BEV, then the selected object will be automatically placed on the position as the user suggests, based on an encapsulated precise calibration algorithm. The flow of data through the system is handled by a tracking server. More details of UI can be found in the supplementary material.

The UI makes it possible to enlarge our dataset, endowing TO-Scene with decent scalability. Besides, similar transfer between different types of datasets can be achieved by revising the UI for creating new datasets or various purposes.

Transfer rules. Notably, two implicit rules during transfer are achieved:

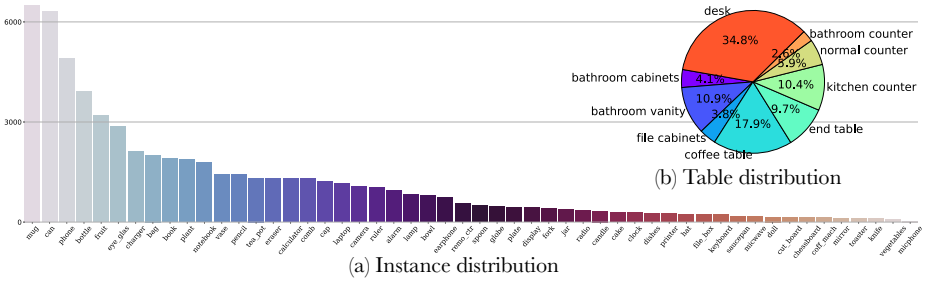
(1) *Rationality.* The function of a table is supposed to *match* with the objects above it. For instance, a mug will more likely to be appeared on coffee tables, while a pencil is possibly placed on writing desks. To achieve this, we set the UI to present objects only from the categories that fit the selected table, and the table label as well as its surroundings are shown to help users figure out the table function. Besides, users are guided to place objects according to their commonsense knowledge instead of reckless operations.

(2) *Diversity.* As mentioned in (1), the tables are randomly picked by UI with various instances shown to fit the table functions. Additionally, we employ around 500 users who may face with diverse tabletop scenes in their daily life, from different professions (e.g., teacher, doctor, student, cook, babysitter) and ages (20∼50). After they finish, about 200 new users will double-check the samples, when they can rearrange, add or delete the objects.

Storage. Each time the user finishes a transfer, the UI will generate a file recording the file names of CADs and rooms, table ID, as well as the calibration parameters. This makes our dataset parameterized and editable, promoting us to store the results in a memory-saving way.

3.2 Simulate Tabletop Scenes into Real Scans

We realize a possible domain mismatch between the original synthetic tabletop objects and the real scanned data, which may result that the deep algorithms trained on TO-Scene are not able to directly work on real-world data. To avoid this, it is necessary to simulate the CAD objects into realistic data.



(b) Table distribution

(a) Instance distribution

Fig. 3: Semantic annotation statistics of TO_Vanilla.

According to the real-world data collection procedures, we firstly utilize Blender [9,11] to render CAD objects into several RGB-D sequences, via emulating the real depth camera. For instance, when a table is against walls, our simulated camera poses will only cover the views in front of tables. Generally, the objects are visible from different viewpoints and distances and may be partially or completely occluded in some frames. Then, the rendered RGB-D frames are sent to TSDF [49] reconstruction algorithms for generating the realistic 3D data.

So far, the whole process brings the occlusions and reconstruction errors, simulating the noise characteristics in real data scanned by depth sensors. As a matter of course, the data domain of tabletop objects are *unified* with the tables and background furniture that are extracted from real scanned ScanNet [10]. Table 5 demonstrates that the model trained on our dataset can be generalized to realistic data for practical applications.

3.3 Annotate Tabletop Instances

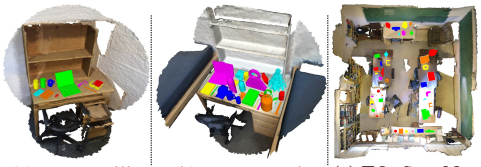
Agreeably, the bounding box annotations (i.e., the center coordinates of boxes, length, width, height, and semantic label) of tabletop objects are naturally gained from their CAD counterparts. Next, since the bounding box of an instance delineates an area covering its owning points, this promotes us to get the point-wise annotations straightforwardly and automatically.

Following above procedures, we build a dataset consisting of 12,078 tabletop scenes, with 60,174 tabletop object instances from 52 common classes. Fig. 4 (a)) shows a sample of this vanilla dataset, denoted as **TO_Vanilla**, which is also the foundation of our dataset. Fig. 3 shows the statistics for the semantic annotation of the major tabletop objects and the used table categories in TO_Vanilla.

3.4 Data Enrichment

We construct another two variants upon TO_Vanilla, for benchmarking existing algorithms under more real scenarios with new challenges.

Crowded objects. The tabletops in our daily life are often full of crowded objects with more inter-occlusions, To simulate this challenge, we reload



(a) TO_Vanilla (b) TO_Crowd (c) TO_ScanNet

Fig. 4: Three variants in TO-Scene.

Table 2: Train/Test Split.

	Scenes		Instances	
	♯Train	♯Test	♯Train	♯Test
TO_Vanilla	10003	2075	49982	10192
TO_Crowd	3350	649	43390	8665
TO_ScanNet	3852	811	114631	23032

some object-table sudo-mappings outputs when building TO_Vanilla (i.e., the tabletops after object transfers, yet not rendered or reconstructed), and employ novices to place much more tabletop CADs above each table. Then we render and reconstruct the new crowded set via the same way as before. Consequently, more occlusions with reconstruction flaws are introduced (see Fig. 4 (b)), yielding a more challenging setup of tabletop scenes, indicated by **TO_Crowd**. It covers 3,999 tabletop scenes and 52,055 instances.

Parse whole room in one stage. Previous TO_Vanilla and TO_Crowd assume to only parse tabletop scenes, but many real applications require to process the *whole* room with all furniture including tabletop objects in one stage. To make up this situation, we keep the complete scans of ScanNet [10] rooms in each data sample, from which the tables of TO_Vanilla are extracted. We maintain the semantic label on original room furniture from ScanNet. As a result, another variant **TO_ScanNet** is presented (see Fig. 4 (c)), which requires algorithms to comprehensively understand both tabletop objects and room furniture. TO_ScanNet can be regarded as an *augmented ScanNet*.

3.5 TO-Scene Dataset.

Finally, **TO-Scene** is born by combining TO_Vanilla, TO_Crowd and TO_ScanNet. The train/validation split statistics of three variants are summarized in Table 2, with the per-category details shown in the supplementary material. Our TO-Scene contains totally 16,077 tabletop scenes with 52 common classes of tabletop objects. The annotation includes vertex semantic labels, 3D bounding boxes, and camera poses during rendering, which opens new opportunities for exploring the methods in various real-world scenarios.

Note: The stated three variants are currently organized separately for different uses. One may combine either of them to explore more research prospects. Furthermore, this paper just presents the current TO-Scene snapshot, we will keep replenishing our dataset and provide extra *private test set* for benchmarking in the future. More data samples can be found in the supplementary material.

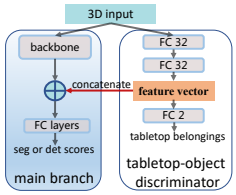


Fig. 5: Tabletop-aware learning.

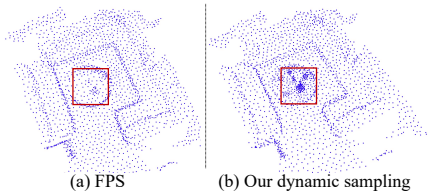


Fig. 6: FPS V.S. dynamic sampling.

4 Tabletop-aware Learning Strategy

For demonstrating the value of TO-Scene data, we focus on 3D semantic segmentation and 3D object detection tasks, towards understanding scenes from both point-wise and instance-level.

In TO-Scene, since tabletop objects are mostly in smaller-size compared with other large-size background furniture, it is naturally difficult to discriminate tabletop instances, especially for TO_ScanNet with lots of big furniture. Additionally, existing 3D networks mostly apply conventional downsampling such as farthest point sampling and grid sampling to enlarge the receptive field. Nevertheless, after being sampled by these schemes, the point densities of small tabletop objects are conspicuously sparser than big furniture (see Fig. 6 (a)), which hurts the perceiving of tabletop objects.

To handle these issues, our idea is to guide the network aware of the presence of tabletop objects, via adding a tabletop-object discriminator that is essentially a binary classifier. The loss is jointly optimized as the sum of the tabletop-object discriminator and segmentation (or detection) loss, which can be written as: $L_{total} = L_{seg\,or\,det} + \lambda L_{dis}$, where λ is the weight. For segmentation, the $0-1$ ground truth (gt) for L_{dis} can be directly gained from the point-wise semantic labels indicating if a point class belongs to tabletop objects, and L_{dis} is a cross-entropy loss. Yet the gt for detection only comes from bounding boxes. For fairly learning the discriminator, we employ a *soft* point-wise gt label that is a normalized per-point distance between each point and the center of a gt tabletop object bounding box, and compute the mean squared error as L_{dis}.

In this work, two operations are derived by tabletop-object discriminator:

(1) As shown in Fig. 5, the feature vector before the last fully connected layer of tabletop-object discriminator is concatenated with the features extracted by the main segmentation or detection branch, so that the predictions of the main branch are driven by the tabletop belongings information.

(2) A dynamic sampling strategy is proposed, where the points with higher tabletop-object discriminator score (i.e., points of tabletop objects) are more likely to be sampled (see Fig. 6 (b)). We replace the original sampling during the feature extraction in all backbone networks with our dynamic sampling.

In the practice, tabletop-object discriminator is implemented through a few fully-connected layers, assisted via max-pooling on K Nearest Neighbor (KNN) point features for fusing contextual information. Our joint learning concept can

Table 3: Segmentation mIoU (%).

Method	TO_Vanilla	TO_Crowd	TO_ScanNet
PointNet [31]	49.31	44.89	36.74
PointNet^{++} [32]	65.57	61.09	53.97
PointNet^{++} + FV	68.74	64.95	57.23
PointNet^{++} + DS	67.52	63.28	56.97
PointNet^{++} + FV + DS	69.87	65.15	58.80
PAConv [47]	75.68	71.28	65.15
Point Trans [52]	77.08	72.95	67.17
Point Trans + FV	79.01	75.06	69.09
Point Trans + DS	77.84	73.81	68.34
Point Trans + FV + DS	**79.91**	**75.93**	**69.59**

Table 4: Detection mAP@0.25 (%).

Method	TO_Vanilla	TO_Crowd	TO_ScanNet
VoteNet [30]	53.05	48.27	43.70
VoteNet + FV	59.33	50.32	48.36
VoteNet + DS	58.87	58.05	52.92
VoteNet + FV + DS	60.06	58.87	56.93
H3DNet[50]	59.64	57.25	52.39
Group-Free 3D [27]	61.75	59.61	49.04
Group-Free 3D + FV	62.26	59.63	53.66
Group-Free 3D + DS	62.19	59.69	55.71
Group-Free 3D + FV + DS	**62.41**	**59.76**	**57.57**

be promoted for tackling similar problems that requires to parse the objects with large size variance. The experimental results are listed in Sec. 5.

5 Benchmark on TO-Scene

For making the conclusions solid, extensive experiments are conducted on 3D semantic segmentation and 3D object detection tasks.

5.1 Common Notes

Below are common notes for both two tasks. The implementation details can be found in the supplementary material.

(1) We follow the original training strategies and data augmentations of all tested methods from their papers or open repositories.

(2) In Table 3 and Table 4, "FV" means applying feature vector of tabletop-object discriminator, "DS" indicates our dynamic sampling strategy.

5.2 3D Semantic Segmentation

Pre-voxelization. A common task on 3D data is semantic segmentation (i.e. labeling points with semantic classes). We advocate to pre-voxelize point clouds, which brings more regular structure and context information, as in [18,7,26,40,52]. We set the voxel size to $4mm^3$ for matching the small sizes of tabletop objects. After voxelization, every voxel stores a surface point with object class annotation. Then we randomly sample 80,000 points from all voxels in a scene for training, and all points are adopted for testing.

Networks. We benchmark PointNet [31], PointNet++ [32], PAConv [47] and Point Transformer [52]. PointNet++ and Point Transformer are chosen as the backbones to plug our tabletop-aware learning modules.

Results and analysis. Following the state-of-the-arts [20,47,52], we use mean of classwise intersection over union (mIoU) as the evaluation metrics. As we can

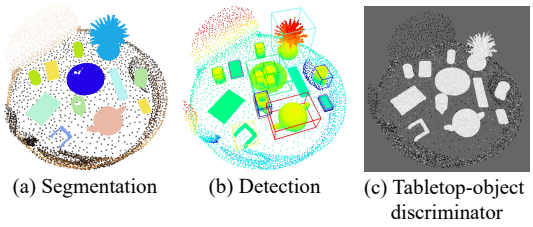


Fig. 7: Benchmark result.

Fig. 8: TO-Real test set.

see from Table 3, the state-of-the-arts learned from the training data are able to perform well in the test set based on the geometric input. Moreover, our tabletop-aware learning modules stably improve the model performance, especially when they are applied together.

5.3 3D Object Detection

Data-preprocessing. Understanding indoor scenes at the instance level is also important. We follow the original pre-processing schemes of the selected state-of-the-arts.

Networks. We run VoteNet [30], H3DNet [50] and Group-Free 3D [27] on TO-Scene. VoteNet and Group-Free 3D are picked as the backbones to integrate our tabletop-aware learning strategies.

Results and analysis. For the evaluation metrics, we use mAP@0.25, mean of average precision across all semantic classes with 3D IoU threshold 0.25, following the state-of-the-arts. As shown in Table 4, the deep networks achieve good results based on the geometric input. Our tabletop-aware learning methods again significantly improve the model performance.

A sample of segmentation, detection result and the predicted tabletop-belongings is visualized in Fig. 7. We can see that the networks successfully segment or detect the objects with tabletop-awareness. Both the segmentation and detection results show that TO_Crowd is more challenging than TO_Vanilla, and TO_ScanNet is most difficult. More result visualizations, the result of each class, and the ablations of tabletop-aware learning strategies are presented in the supplementary material.

6 Real-world Test

The ultimate target of our dataset is for serving the real applications. For a clearer picture of this goal, the below steps are performed.

Data. Since there is no real dataset that perfectly match with the three variants of TO-Scene, the first thing is to acquire real-world data. We employ expert

users to manually scan and annotate **TO-Real** including three sets of data (see Fig. 8), which are respectively denoted as Real_Vanilla, Real_Crowd, Real_Scan. Specifically, Real_Vanilla contains 97 tabletop scenes, while Real_Crowd includes 100 tabletop scenes with crowded objects. In Real_Scan, 22 scenes are scanned with both big furniture and small tabletop objects. The categories in TO-Real cover a subset of our TO-Scene objects (see Table 5).

Implementations. We train Point Transformer [52] for semantic segmentation and VoteNet [30] for detection on different variants of TO-Scene, and *directly* test on the corresponding TO-Real counterparts, without any fine-tuning.

Results and Analysis. Table 5 enumerates the segmentation mIoU and detection mAP of each class, where we find ***two important phenomenons***:

(1) Under both Vanilla and Crowd settings that *specifically* for parsing tabletop objects, the variance between the results tested on TO-Scene and TO-Real is stably acceptable. For detection task, the test mAP (highlighted in red bold) of some categories on TO-Real is even better than it on TO-Scene. This definitely proves the practical value of the *tabletop*-scenes data in TO-Scene.

(2) The more interesting case lies on whole-room Scan (emphasized in gray shadow). As for the big furniture categories (highlighted in blue), the result variance of these categories are obviously undesirable. Note that these categories in TO-Scene all come from ScanNet [10]. Additionally, the results of tabletop object classes *particularly degrade* under whole room Scan setup, which is probably influenced by ScanNet big furniture that are simultaneously learned with tabletop objects during training on TO_ScanNet.

Here we discuss some possible reasons. Scanning big furniture around large room spaces is physically hard to control, causing various unstable noise (point density, *incompleteness*), and instance *arrangements/layouts* greatly change geographically. These unavoidable factors possibly cause the poor model generalization on big furniture across various data collections. As for tabletop objects, we guess the domain gap is small because scanning tabletops is physically controllable, bringing less unstable noise.

More details of TO-Real and the result visualizations are illustrated in the supplementary material.

7 Discussion and Conclusion

This paper presents TO-Scene, a large-scale 3D indoor dataset focusing on tabletop scenes, built through an efficient data acquisition framework. Moreover, a tabletop-aware learning strategy is proposed to better discriminate the small-sized tabletop instances, which improve the state-of-the-art results on both 3D semantic segmentation and object detection tasks. A real scanned test set, TO-Real, is provided to verify the practical value of TO-Scene. One of the variants of our dataset, TO-ScanNet, includes totally 4663 scans with 137k instances, which can possibly serve as a platform for *pre-training* data-hungry

Table 5: Per-category test results on TO-Scene/TO-Real.

Class in Real Data	Segmentation mIoU (%)			Detection mAP@0.25(%)		
	Vanilla	Crowd	Scan	Vanilla	Crowd	Scan
Big furniture:						
wall	-	-	76.8/10.0	-	-	-
floor	-	-	94.9/46.5	-	-	-
cabinet	-	-	59.5/19.3	-	-	78.5/3.8
chair	-	-	80.4/33.6	-	-	92.3/55.7
sofa	-	-	75.3/39.2	-	-	98.7/62.2
table	-	-	71.9/23.5	-	-	83.4/16.3
door	-	-	55.6/19.2	-	-	69.5/10.1
window	-	-	59.5/5.9	-	-	57.1/1.5
bookshelf	-	-	64.0/10.2	-	-	69.6/10.5
picture	-	-	20.8/21.8	-	-	16.7/18.5
counter	-	-	58.5/6.2	-	-	86.0/8.7
desk	-	-	62.8/4.1	-	-	95.2/3.3
curtain	-	-	58.3/4.6	-	-	76.6/41.7
refrigerator	-	-	61.8/1.2	-	-	93.9/21.3
sink	-	-	58.1/34.0	-	-	67.9/70.0
Tabletop object:						
bottle	88.3/62.5	87.9/70.3	85.3/23.6	64.3/**67.3**	72.4/49.3	91.4/33.9
bowl	89.5/61.1	87.5/51.2	85.7/1.8	75.5/45.7	77.0/69.2	90.4/49.2
camera	89.0/74.4	85.3/77.9	81.2/1.2	81.0/**91.7**	74.7/50.5	95.2/3.1
cap	92.1/64.2	87.8/63.1	85.4/28.0	87.1/62.8	88.2/71.7	97.8/40.0
keyboard	86.0/75.7	89.3/77.1	-	51.2/**67.4**	37.7/36.1	-
display	93.6/81.4	91.4/80.6	-	93.9/92.14	84.8/80.2	-
lamp	86.0/72.6	90.0/78.4	72.2/10.9	84.7/84.2	85.6/67.5	98.4/61.6
laptop	94.3/81.6	96.9/66.2	95.5/52.7	90.3/62.8	96.8/69.2	97.7/80.1
mug	94.2/48.8	96.4/72.0	92.0/31.2	81.8/77.3	90.1/67.0	94.6/32.1
alarm	66.7/52.2	66.0/54.0	51.5/2.7	59.2/32.5	48.9/38.9	94.7/9.5
book	77.5/58.6	67.3/41.5	68.6/28.2	60.0/57.2	62.7/**69.8**	92.8/11.8
fruit	94.1/47.4	88.7/42.1	81.6/16.6	76.1/55.9	77.9/44.9	88.8/31.1
globe	96.0/85.1	98.0/75.4	96.4/17.6	87.9/**88.6**	95.9/91.1	99.9/31.1
plant	93.9/65.0	96.9/65.6	92.0/21.0	87.7/**89.2**	89.7/80.7	97.2/13.3

algorithms in 3D tasks towards individual shape-level or holistic scene-level.

Acknowledgements. The work was supported in part by the National Key R&D Program of China with grant No. 2018YFB1800800, the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, and by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001). It was also supported by NSFC-62172348, NSFC-61902334 and Shenzhen General Project (No. JCYJ20190814112007258). We also thank the High-Performance Computing Portal under the administration of the Information Technology Services Office at CUHK-Shenzhen.

References

1. Ahmadyan, A., Zhang, L., Ablavatski, A., Wei, J., Grundmann, M.: Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In: CVPR (2021)
2. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: CVPR (2016)
3. Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Niessner, M.: Scan2cad: Learning cad model alignment in rgb-d scans. In: CVPR (2019)
4. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 3DV (2017)
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
6. Choi, S., Zhou, Q.Y., Miller, S., Koltun, V.: A large dataset of object scans. arXiv preprint arXiv:1602.02481 (2016)
7. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: CVPR (2019)
8. Collins, J., Goel, S., Luthra, A., Xu, L., Deng, K., Zhang, X., Vicente, T.F.Y., Arora, H., Dideriksen, T., Guillaumin, M., Malik, J.: ABO: dataset and benchmarks for real-world 3d object understanding. arXiv preprint arXiv:2110.06199 (2021)
9. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), `http://www.blender.org`
10. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
11. Denninger, M., Sundermeyer, M., Winkelbauer, D., Zidan, Y., Olefir, D., Elbadrawy, M., Lodhi, A., Katam, H.: Blenderproc. arXiv preprint arXiv:1911.01911 (2019)
12. Depierre, A., Dellandréa, E., Chen, L.: Jacquard: A large scale dataset for robotic grasp detection. In: IROS (2018)
13. Fang, H.S., Wang, C., Gou, M., Lu, C.: Graspnet-1billion: A large-scale benchmark for general object grasping. In: CVPR (2020)
14. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, J., Li, C., Zeng, Q., Sun, C., Jia, R., Zhao, B., et al.: 3d-front: 3d furnished rooms with layouts and semantics. In: CVPR (2021)
15. Garcia-Garcia, A., Martinez-Gonzalez, P., Oprea, S., Castro-Vargas, J.A., Orts-Escolano, S., Garcia-Rodriguez, J., Jover-Alvarez, A.: The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions. In: IROS (2018)
16. Georgakis, G., Reza, M.A., Mousavian, A., Le, P., Kosecka, J.: Multiview RGB-D dataset for object instance detection. In: 3DV (2016)
17. Girardeau-Montaut, D.: Cloudcompare3d point cloud and mesh processing software. OpenSource Project (2011)
18. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: CVPR (2018)
19. Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., Cipolla, R.: Scenenet: Understanding real world indoor scenes with synthetic data. In: CVPR (2016)

20. Hu, W., Zhao, H., Jiang, L., Jia, J., Wong, T.T.: Bidirectional projection network for cross dimensional scene understanding. In: CVPR (2021)
21. Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.K., Yu, L.F., Yeung, S.K.: Scenenn: A scene meshes dataset with annotations. In: 3DV (2016)
22. Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D., Panozzo, D.: Abc: A big cad model dataset for geometric deep learning. In: CVPR (2019)
23. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: ICRA (2011)
24. Lenz, I., Lee, H., Saxena, A.: Deep learning for detecting robotic grasps. IJRR (2015)
25. Li, Z., Yu, T.W., Sang, S., Wang, S., Song, M., Liu, Y., Yeh, Y.Y., Zhu, R., Gundavarapu, N., Shi, J., Bi, S., Yu, H.X., Xu, Z., Sunkavalli, K., Hasan, M., Ramamoorthi, R., Chandraker, M.: Openrooms: An open framework for photorealistic indoor scene datasets. In: CVPR (2021)
26. Liu, Z., Hu, H., Cao, Y., Zhang, Z., Tong, X.: A closer look at local aggregation operators in point cloud analysis. In: ECCV (2020)
27. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: ICCV (2021)
28. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM (1995)
29. Park, K., Rematas, K., Farhadi, A., Seitz, S.M.: Photoshape: Photorealistic materials for large-scale shape collections. ACM Trans. Graph. (2018)
30. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: ICCV (2019)
31. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR (2017)
32. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS (2017)
33. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: ICCV (2021)
34. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: PiGraphs: Learning Interaction Snapshots from Observations. ACM Trans. Graph. (2016)
35. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: ICCV Workshops (2011)
36. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
37. Singh, A., Sha, J., Narayan, K.S., Achim, T., Abbeel, P.: Bigbird: A large-scale 3d database of object instances. In: ICRA (2014)
38. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR (2015)
39. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: CVPR (2017)
40. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: ICCV (2019)
41. Torralba, A., Russell, B.C., Yuen, J.: Labelme: Online image annotation and applications. IJCV (2010)
42. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, D.T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: ICCV (2019)

43. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: CVPR (2015)
44. Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., Savarese, S.: Objectnet3d: A large scale database for 3d object recognition. In: ECCV (2016)
45. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: WACV (2014)
46. Xiao, J., Owens, A., Torralba, A.: Sun3d: A database of big spaces reconstructed using sfm and object labels. In: ICCV (2013)
47. Xu, M., Ding, R., Zhao, H., Qi, X.: Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In: CVPR (2021)
48. Xu, M., Zhang, J., Zhou, Z., Xu, M., Qi, X., Qiao, Y.: Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In: AAAI (2021)
49. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In: CVPR (2017)
50. Zhang, Z., Sun, B., Yang, H., Huang, Q.: H3dnet: 3d object detection using hybrid geometric primitives. In: ECCV (2020)
51. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia (2012)
52. Zhao, H., Jiang, L., Jia, J., Torr, P., Koltun, V.: Point transformer. In: ICCV (2021)