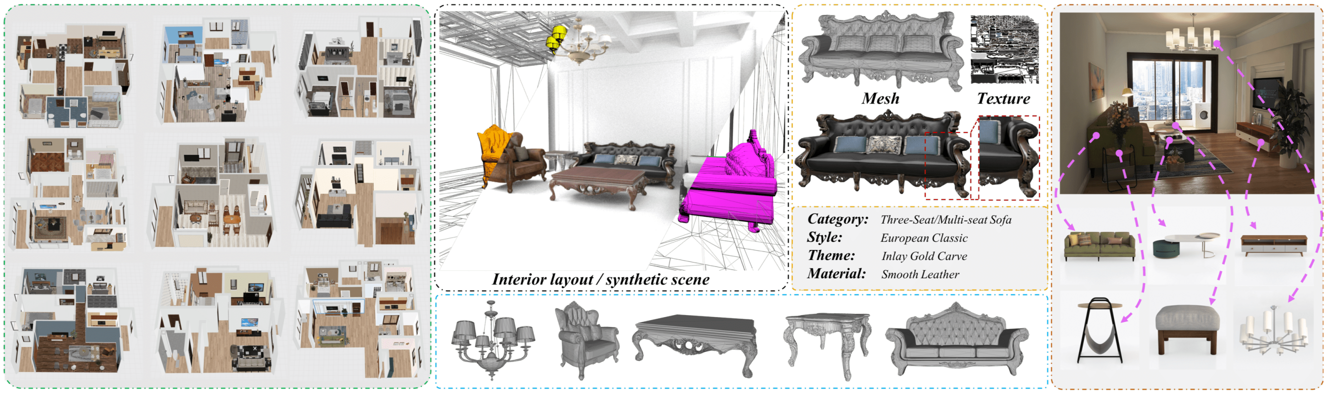
3D-FRONT: 3D Furnished Rooms with layOuts and semaNTics

Huan Fu[1] Bowen Cai[1] Lin Gao[2] Ling-Xiao Zhang[2] Jiaming Wang[1]
Cao Li[1] Qixun Zeng[1] Chengyue Sun[1] Rongfei Jia[1] Binqiang Zhao[1] Hao Zhang[3]

[1]Tao Technology Department, Alibaba Group
[2]Institute of Computing Technology, Chinese Academy of Sciences
[3] GrUVi (Graphics U Vision) Lab, Simon Fraser University

{fuhuan.fh, kevin.cbw, rongfei.jrf, binqiang.zhao}@alibaba-inc.com
{gaolin, zhanglingxiao}@ict.ac.cn haoz@sfu.ca



Professionally designed layouts *19K rooms with high-quality textured 3D furniture objects* *Interior design ideas from experts*

Figure 1: **3D-FRONT** is a new, large-scale, and comprehensive repository of synthetic indoor scenes with professionally and distinctively designed layouts, a large number (18,968) of rooms populated with 3D furniture objects that are stylistically compatible and endowed with high-quality textures. All freely available to the academic community and beyond.

Abstract

We introduce 3D-FRONT (3D Furnished Rooms with layOuts and semaNTics), a new, large-scale, and comprehensive repository of synthetic indoor scenes highlighted by professionally designed layouts and a large number of rooms populated by high-quality textured 3D models with style compatibility. From layout semantics down to texture details of individual objects, our dataset is freely available to the academic community and beyond. Currently, 3D-FRONT contains 6,813 CAD houses, where 18,968 rooms diversely furnished by 3D objects, far surpassing all publicly available scene datasets. The 13,151 furniture objects all come with high-quality textures. While the floorplans and layout designs (i.e., furniture arrangements) are directly sourced from professional creations, the interior designs in terms of furniture styles, color, and textures have been carefully curated based on a recommender system we develop to attain consistent styles as expert designs. Furthermore, we release Trescope, a light-weight rendering tool, to support benchmark rendering of 2D images and annotations from 3D-FRONT. We demonstrate two applications, interior scene synthesis and texture synthesis, that are especially tailored to the strengths of our new dataset.

1. Introduction

The computer vision community has invested much effort into the study of 3D indoor scenes, from 3D reconstruction, visual SLAM, and navigation, to scene understanding, affordance analysis, and generative modeling. With data-driven and learning-based approaches receiving more and more attention in recent years, there has been a steady accumulation of indoor scene datasets [27, 36, 43, 4, 19, 6, 9, 18, 23, 49, 24] to drive the deep learning revolution that has

Dataset	Layout Design	#3DFRs	#CAD models	Model Textures	3D Annotation
NYU-Depth v2 [27]	Real scan	N/A	N/A	No texture	Raw RGB-D
TUM [36]	Real scan	N/A	N/A	No texture	Raw RGB-D
SUN3D [43]	Real scan	254	N/A	No texture	Raw PCD
S3DIS [4]	Real scan	270	N/A	No texture	Raw PCD
3DSSG [38]	Real scan	478	N/A	Rec. from Scan	Raw Mesh
SceneNN [19]	Real scan	100	N/R	Rec. from Scan	Raw Mesh
Matterport3D [6]	Real scan	2,056	N/A	Rec. from Scan	Raw Mesh
ScanNet [9]	Real scan	1,506	296	Rec. from Scan	Raw Mesh
Scan2CAD [5]	Real scan	1,506	3,049	No texture	Mesh
OpenRooms [24]	Real scan	1,068	2,500	Amateur	Mesh
SceneNet [18]	Professional	57	N/R	No texture	Mesh
InteriorNet [23]	Professional	N/A	N/A	No texture	N/A
Hypersim [30]	Professional	N/A	N/A	Per-pixel color	RGB-D
Structured3D [49]	Professional	N/A	N/A	No texture	3D structures
3D-FRONT	Professional	**18,968**	**13,151**	**Professional**	Mesh

Table 1: **Comparison between prominent 3D indoor scene datasets**, where "#3DFRs" represents the number of rooms or scenes populated with 3D furniture objects, "N/A" = "not available", "N/R" = "not reported", "Raw Mesh" denotes machine reconstructed meshes, and "Raw PCD" refers to reconstructed point clouds. For model textures, "Rec. from Scan" is the result of reconstruction from raw RGB-D data, while "Amateur" and "Professional" refer to who designed the textures. The "3D structures" annotatd by Structured3D [49] contain information on primitives including 3D boxes and their relations.

redefined the landscape of indoor scene processing.

Existing 3D scene datasets all fall into two broadly categories: acquired (via scanning and reconstruction) vs. designed (i.e., synthetic scenes created by humans). In terms of data volume, the largest repository is ScanNet [9] which consists of 2.5M RGB-D images from 1,513 scanned real scenes acquired by commodity sensors, in 707 distinct spaces. The 3D scenes, including textured 3D objects, were recovered by state-of-the-art 3D reconstruction techniques from the raw scans, which are typically noisy and incomplete. As a result, the reconstructed meshes are often of low quality, both in geometric fidelity and texture quality.

In the world of synthetic 3D indoor scene datasets, the recent exit by SUNCG [34] has left an apparent void in the community. Most recently, Structured3D [49] and OpenRoom [24] have emerged as promising alternatives. In addition to providing professionally designed room layouts, Structured3D [49] aims to provide large-scale photo-realistic scene *images* with rich 3D structure annotations. However, the actual 3D furniture objects populating the scenes are not included in the dataset. OpenRoom [24] replaces detected objects in a set of 1,068 scanned scenes from ScanNet [9] with CAD models from ShapeNet [7]. A major contribution of this dataset is to provide ground-truth annotations of complex material parameters for the CAD objects. However, the dataset has not been released at this point and according to the authors' account, only 2.5K CAD models were annotated with material properties.

In this paper, we introduce 3D-FRONT (3D Furnished Rooms with layOuts and semaNTics), a new, large-scale, and *comprehensive* repository of synthetic 3D indoor scenes. It contains professionally and distinctively designed layouts spanning 31 scene categories (or room types), object semantics (e.g., category, style, and material labels), and a large number (18,968) of rooms populated with 3D furniture objects. Most importantly, these 3D furniture objects are all endowed with high-quality textures, thanks to 3D-FUTURE [14], a recently released dataset of quality 3D furniture used in industrial productions. Furthermore, the selection of furniture objects from 3D-FUTURE to populate the scenes in 3D-FRONT has been inpired by expert interior designs. Specifically, the selection is based on a *recommender system* learned from the expert designs, while taking into account of furniture styles both in terms of geometry and texture. As a result, the furnished rooms in 3D-FRONT consist of *stylistically compatible* objects adhering to the design inspirations.

In Table 1, we present essential information for the current public release of 3D-FRONT and compare to other prominent indoor scene datasets. As we can see, the most compelling feature of our dataset is the large number of 3D furnished rooms, which far surpasses all the other publicly available datasets. Style compatibility, as well as the high texture quality, of the furniture objects in each scene (see middle of Figure 1) is another unique attribute of 3D-FRONT. On top of all these, the total number of rooms with professionally designed layouts is 45,000, in which 18,968 rooms are fully populated with 3D furniture shapes. Last but not least, we share Trescope, a light-weight rendering tool, with the community so that the users of 3D-FRONT can easily capture their desired 2D renderings and annotations to guide their image-driven learning tasks. We will continuously improve 3D-FRONT by providing much enriched texture and 3D geometry contents.

We anticipate that 3D-FRONT, being as comprehensive as it is, will enable and further drive a whole suite of AI-powered and data-driven scene analysis and modeling applications. We demonstrate two applications which cannot be well supported by other publicly available datasets — these applications are best served by having a large number of high-quality textured mesh models with style consistency, a unique feature of 3D-FRONT. One such application is learning to texture 3D objects in indoor scenes. In another, by learning the layout of 3D furniture in each room with [40], we can coherently predict and arrange functional furniture for an empty room.

2. Related Work

Over the past years, a large number of RGB-D benchmarks have been constructed and made publicly available [21, 3, 27, 36, 43, 33, 19, 6, 9, 18, 26, 23, 49, 24, 4, 32, 35, 42, 11, 17, 47, 38, 2]. Current 3D scene datasets are mainly collected based on scanning and reconstruction or human creation. These datasets thus fall into two broadly categories: Acquired vs. Designed.

Acquired Scenes. To construct "Acquired" datasets, researchers capture RGB-D videos, reconstruct the scene meshes, and manually label the frames or the reconstructed scenes. For example, NYU-Depth v2 [27] gathered 464 short RGB-D sequences from different rooms via Kinect, where 1,449 images are selected and labeled with pixel-level annotations. SUN RGB-D [33] collected 10,335 RGB-D images and provided more 2.5D annotations, such as 2D polygons and 3D bounding boxes correspondences, room layouts, and scene categories. These datasets may lack the physical relationship between the frames and the scene space's real 3D structure. To address the issue, SUN3D [43] developed an interactive reconstruction pipeline to recover the 3D scene structures for 254 different spaces in 41 buildings, in which 8 scenes are provided with semantic labels for 3D point clouds and camera poses. SceneNN [19] improved the pipeline by recovering mesh surfaces instead of point clouds for 100 scans. Further, one of the largest "Scanned" datasets, *i.e.,* ScanNet [9], has been established. It reconstructed 1,513 rooms based on 2.5M RGB-D views, and labeled rich 3D annotations, including estimated 3D camera poses, surface reconstructions, semantic segmentation, and 2D-3D alignments.

Since 3D scene reconstruction with fine geometric and textures details is still a challenging problem with the depth cameras on the shelf such as Kinect, the mesh qualities in these scene dataset are usually not as good as the synthetic data. Besides, some of the 3D annotations may be unreliable or imprecise due to the reconstruction error, such as camera pose and 2D-3D alignment.

Designed Scenes. Another type of scene dataset is from the human creation with professional design software as this 3D-FRONT. In addition to 3D-FRONT, there is one synthetic (designed) dataset that shares both the layout and the well-posited 3D CAD mesh models, *i.e.,* SceneNet [18] with providing 59 scenes. Several other synthetic benchmarks share 2D and 2.5D contents based on designed synthetic scenes. For example, InteriorNet [23] released 15k sequences and 5M images, which are rendered from their large-scale scene packages. Further, Structure3D [49] provided 21,835 panoramic images with the corresponding structure annotations, such as panoramic layouts, depth, surface normal. Recently, Li et al. [24] built OpenRooms, a synthetic benchmark based on ScanNet, and planned to share rendered images with their high-quality SVBRDF and spatially-varying lighting. Also, Hypersim [30] presented a photorealistic synthetic dataset for holistic indoor scene understanding, focusing on providing per-pixel depth and disentangled illuminance and reflectance properties over scene images designed by professional artists.

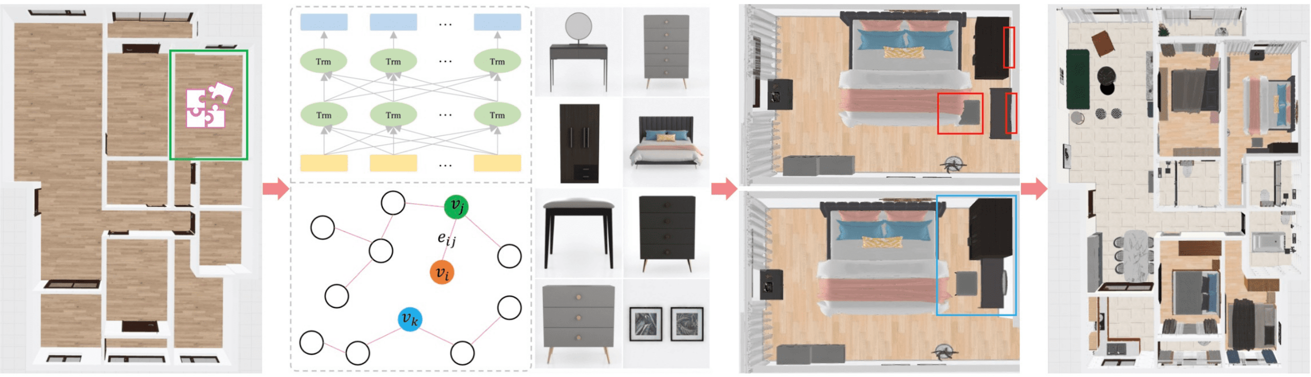
These large-scale synthetic datasets have not made the completed scene packages, including the floorplans' mesh, the large amount of involved CAD models with fine geometric and texture details, and the layout with design ideas, publicly available. In contrast, 3D-FRONT shares everything that is used to construct houses, from real layouts to interior design ideas and involved objects. The holistic repository of indoor scene packages enables a robot to navigate in them. It also allows the researchers to render whatever information they need for new subjects studying.

3. Building 3D-FRONT

Creating a large-scale 3D scene repository is a non-trivial task. Our 3D-FRONT project has been built on a large volume (about 60K) of professionally designed houses and 1M 3D CAD meshes. While we are unable to publish all these meshes, due to copyright restrictions, all the models and learning algorithms employed during the data collection progress have been trained on the large database. As shown in Figures 2, we start from some house collections, create room suites, optimize the layout, verify the created interior designs. In the following, we will detail the pipeline as well as the techniques involved.

3.1. Room Suite Creation

Given a CAD house and its professional design ideas, we automatically create room suites for the scenes. Here, a room consist of the category labels of objects that are suggested to put in, and their positions, orientations, sizes, and



House + Design Ideas *Room Suit Creation* *Layout Optimization* *Verification*

Figure 2: **Pipeline of building 3D-FRONT.** We start from an empty house with professional design ideas, create the room suites, optimize the layouts (e.g., to resolve artifacts highlighed in the red boxes), and finally verify the furnished rooms. Here, the design ideas for a room consist of the category labels of objects that are suggested to put in, and their positions, orientations, sizes, and styles.

styles. Taking a bedroom as an example, we first randomly select a seed object, *e.g.,* a bed, from a 3D model pool according to the suggested size and style of the design ideas. We then recurrently identify the visually matched furniture according to the room suite thus far until the room is filled.

We mainly rely on the Furnishing Suite Composition (FSC) approach in 3D-FUTURE [14] to create visually compatible suites. Specifically, leveraging on the large-scale expert scene designs, we carry out two tasks, *i.e.,* mask prediction and suite compatibility scoring, to model visual compatibility. The first task predicts the masked (removed) furniture given other objects in a suite. And the second task evaluates the compatibility score of the input suite. We utilize a textured image to represent each object (furniture), as shown in Figure 2 (*Room Suite Creation*). The two tasks optimize a visual embedding network (VEN) [31] and two transformer architectures [37, 10], so that the trained VEN can extract informative visual feature for each object. With the learned visual representation and the given attributes, including category, style, color, material, and size, for each object. FSC trains gradient boosting decision trees (GBDT) [13] to infer decision rules based on these information, and post a logistic regression (LR) layer to estimate the comparability scores of the room suites. These two techniques are integrated as the GBDT-LR model.

FSC first adopts the visual embedding extracted from VEN to perform a primary ranking, then employs the trained GBDT-LR model to re-rank the selected candidates for online recommendation. We improve the primary ranking stage by considering graph auto-encoder techniques [8]. In detail, we define an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, and learn a graph auto-encoder (GAE) for visual compatibility prediction following [8]. The graph nodes are all the involved objects in the designed house database. Each node is represented with a feature vector extracted from VEN. Each edge's weight is equal to 1 if the two objects are visually appealing, and 0 otherwise. With the graph, we first learn a graph convolutional network (GCN) [20] as an encoder to propagate neighborhood information to obtain new representations, depending on the connections. Then, we adopt a fully connected layer as a decoder to reconstruct the weight matrix. When building 3D-FRONT, we use the trained models to perform recommendation from the furniture shapes shared by 3D-FUTURE [14].

3.2. Layout Optimization and Verification

We observed that with the room suites constructed using the techniques described so far, placing objects into the corresponding rooms according to their suggested positions and orientations, various layout artifacts still remained. For example, a bed may overlap with its nearby nightstand in the 3D space. Other examples are highlighted in red boxes in Figure 2. One of the main reasons is that it is difficult to find a visually matched furniture based on concurrent room suite that, at the same time, has the same size required by the design ideas — there is potential conflict between style and size compatibilities. To this end, we apply the layout optimization algorithm proposed in [41].

Specifically, we start from the initially created designs, and slightly modify the object positions in the room suites in order to satisfy several layout constraints in [41], including pairwise distance, focal point distance, distance to wall, accessibility, and collision. These constraints were constructed based on statistics of the design rules from our synthetic house database. Since the intial layouts often provide a good starting point, we only optimize the defined energy function in up to 50 iterations. On average, the optimization only takes 10s for each room.

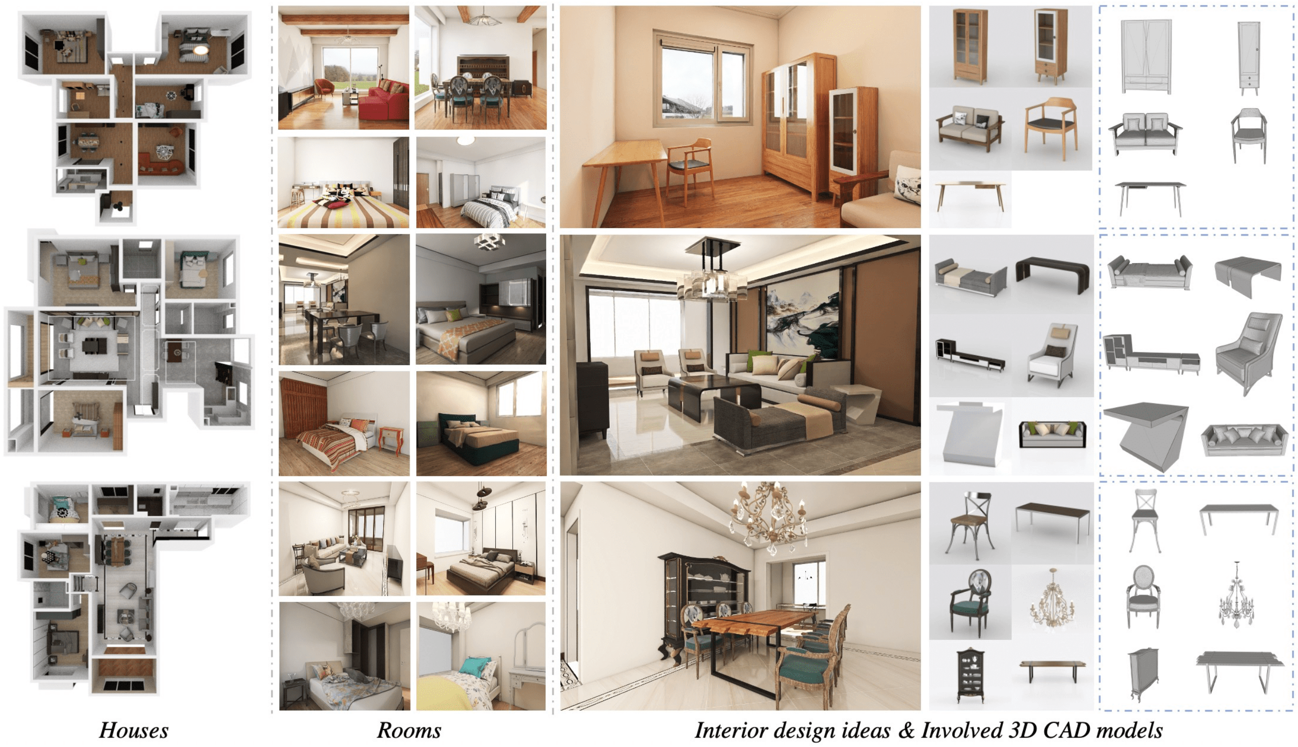


Figure 3: **House Examples in 3D-FRONT.** The left column shows the top-down views of three houses. The middle column presents several rooms contained in these houses, including bedrooms, living rooms, dining rooms, *etc*. The interior design ideas at the right column summarize the textured objects involved in the rooms and their high-quality 3D CAD models.

We further manually verify the created designs and remove the unsatisfied ones to ensure dataset quality. To facilitate the reviewing step, we develop a light-weight renderer Trescope that enables the reviewers to browse the synthetic houses online in an interactive manner. Note that, Trescope supports offline benchmark rendering on local machines for 3D-FRONT. The renderer will be shared so that users of 3D-FRONT can capture their desired renderings such as images, depth, normal, and segmentation.

4. Validation and Assessment

In this section, we offer several means to validate and assess the way our dataset was built and the quality and utility of the data. Applications are discussed in Section 5.

Evaluation of recommender system. We collected 8K room designs and their design logs from the online design platform[1] of Alibaba Topping Homestyler for our evaluation. We discuss several metrics, including Area Under The Curve (AUC) [12], 1-N Average Rank (1-N Avg Rank), N-1 Average Rank (N-1 Avg Rank), 1-N Hit@10, and 1-N Hit@20. These metrics are calculated based on experts' online logs. To explain these measurements, we take a room suite (Bed, Nightstand, Chair) ⇔ (A, B, C) as an example, where a designer chooses the objects A, B, and C in order. 1-N Avg Rank means that we recurrently perform recommendations (A) → Nightstand and (A, B) → Chair, respectively, and compute the average rank (B and C). Here, Nightstand and Chair are the required categories, and B and C are the specific objects. N-1 denotes that we recommend each object given the other two. Hit@K calculates the TopK recall accuracy. For (A, B) → Chair, a correct recommendation in TopK means that C ranks less than K. We refer to the supplementary material for more details about the recommendation process and these metrics.

The qualitative scores are reported in Table 2. Generally, incorporating GAE [8] with the original FSC [14] would yield improvements on all metrics. We point out that both FSC and its improved version (FSC+GAE) can generate high-quality room suites, though it seems that the performance numbers of 1-N Hit@10 (33.6%∼36.1%) and 1-N Avg Rank (41.6∼37.3) are not significant. But it should note that our 3D pool contains more than 1M models. The vast collection makes the visual compatibility inspired recommendation task extremely challenging, though we have filtered out invalid items in the retrieval sequences according to the fine-grained category labels. It's also worth to mention that, after layout optimization and verification, our

[1] https://www.shejijia.com/

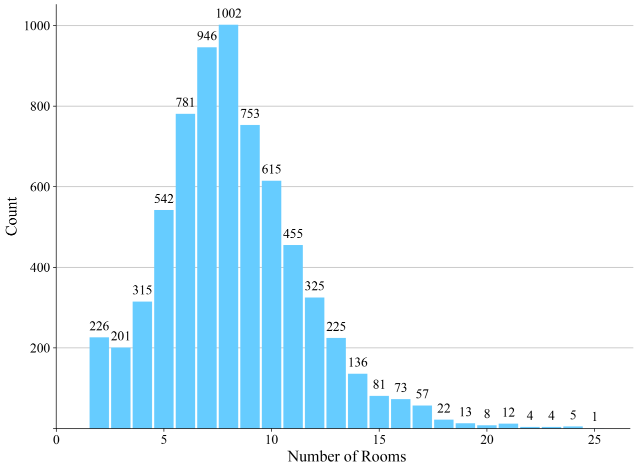


Figure 4: **Statistics of room numbers per house**. 3D-FRONT contains 6,813 distinct houses constructed by 44,427 rooms. There are 6.5 rooms per house on average.

Metrics		FSC [14]	FSC + GAE [8]
↑	AUC	0.766	**0.772**
	1-N Hit@10	33.6%	**36.1%**
	1-N Hit@20	61.3%	**64.3%**
↓	1-N Avg Rank	41.6	**37.3**
	N-1 Avg Rank	26.7	**24.1**

Table 2: **Evaluating the Pipeline.** ↑: higher is better. ↓: lower is better. We perform recommendation based on a extremely large 3D pool (about 1M models). When calculating these scores, invalid items in the retrieval sequences have been filtered out based on fine-grained category labels.

AI created designs (room suites + professional design ideas) have been used for VR shopping by eCommerce merchants. The rate of our high-quality designs (or customer preferred designs) is 88%, while it is only 71% for designs from junior designers. The comparison may not fair for junior designers since we reuse expert design ideas. It shows the good quality of the scene designs in 3D-FRONT.

User study. We conduct a series of user studies, on Amazon Mechanical Turk (AMT), to assess the quality of the data provided by 3D-FRONT, in comparison with SUNCG [34]. The quality criteria considered include those related to scene layouts (in terms of plausibility, design quality, and richness of texture) and individual objects (in terms of texture quality and preferability), as well as style compatibility. We refer to the supplementary material for more details on each study. As for the user study setting, we randomly sampled 90 pairs of scenes and 30 pairs of 3D models from 3D-FRONT and SUNCG based on scene type and model category. Each pair was labeled by 20 master-level annotators in AMT. Thus, the scene and model scores are calculated using 1,800 and 600 feedback, respectively.

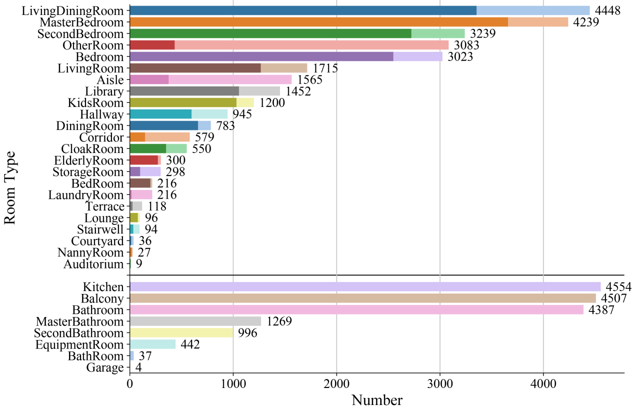


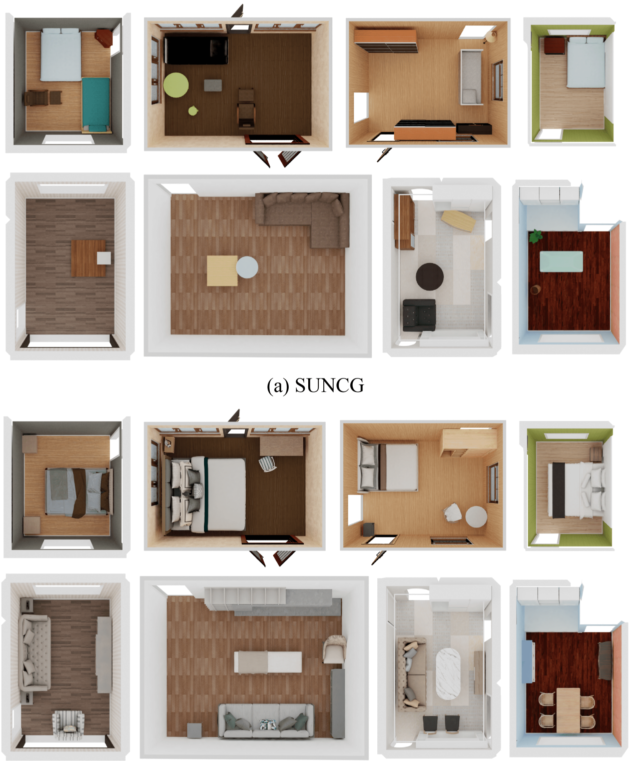
Figure 5: **Distribution of the room scenes available in 3D-FRONT, organized by type.** There are 44,427 rooms in total. A large percentage of rooms (indicated by dark color) in the top part are diversely furnished (18,968). These rooms, such as bedrooms, living rooms, dinging rooms, and study rooms, are the activity spaces where people tend to spend most of their times living indoors.

	Questions	3D-FRONT
Scene	Plausible Layout	**62.5%**
	Design Quality	**69.2%**
	Richer Texture	**70.0%**
	Style Compatibility	**65.4%**
3D Model	Richer Texture	**65.4%**
	Preferable	**61.5%**

Table 3: **User studies on data quality: 3D-FRONT *vs.* SUNCG.** The reported percentages indicate how many users on AMT chose scenes/models from 3D-FRONT when presented questions regarding the quality criteria.

From the scores reported in Table 3, we see that for *each* quality criterion assessed, the majority of Turkers (between 60% and 70%), preferred data presented by 3D-FRONT. We believe that higher-quality datasets would not only lead to improved performance of algorithms which are trained on these datasets, but also enable new applications. It should also be evident that most, if not all, applications in the computer vision and graphics community which had utilized SUNCG, would also be well supported by 3D-FRONT.

Properties of 3D-FRONT. One of the most desirable features of our dataset is that it *publically* shares all the essential data that would enable the modeling of *high-quality* indoor scene, from layout semantics down to stylistic and texture details of individual objects. While the layout ideas are directly sourced from professional designs, the interior designs are transferred from expert creations followed by a post verification process. Figure 3 shows some additional house examples from our dataset.



(a) SUNCG

(b) 3D-FRONT

Figure 6: **Interior Scene Synthesis.** Several scenes produced by a state-of-the-art network trained on SUNCG (a) and 3D-FRONT (b), respectively. The results were synthesized from *randomly* chosen empty rooms. In each set, the first row is for bedrooms and the second row for living rooms. The 3D-FRONT results tend to show a richer variety of objects and more plausible scene layouts. A user study on AMT shows the majority of Turkers (64.8%) prefers scenes synthesized by the 3D-FRONT model.

3D-FRONT enables a variety of AI-powered tasks related to 3D scenes, including data-driven designing studies, such as floorplan synthesis, interior scene synthesis, and scene suites compatibility prediction, that other scene datasets do not support adequately. It also benefits the study of 3D scene understanding subjects, such as SLAM, 3D scene reconstruction, and 3D scene segmentation.

Figures 4 and 5 reveal some statistics of our dataset, with more that can be found in the supplementary material. Further, we assign camera viewpoints to furnished the scenes and release Trescope, a light-weight rendering tool compatible with 3D-FRONT. These would allow users of 3D-FRONT to easily render images and annotations to support their 2D vision studies. We refer to the supplementary material for how we generate camera viewpoints for rooms. Last but not least, we will continuously improve 3D-FRONT by adding more features. A certain plan is to share more enriched texture and 3D geometry contents.

	MMD-CD $\downarrow$	MMD-EMD $\downarrow$	COV-CD $\uparrow$	COV-EMD $\uparrow$
SUNCG [34]	0.3642	1.1490	45.65	46.72
3D-FRONT	**0.3371**	**1.1049**	**50.01**	**52.91**

Table 4: **Evaluting diversity of scenes synthesized by models trained on 3D-FRONT vs. SUNCG.**

5. Applications

We present two applications, interior scene synthesis and object texturing in scene contexts, to demonstrate the utility of our dataset. This only represents a small sampler of applications that can benefit from 3D-FRONT.

5.1. Interior Scene Synthesis

The past several years have seen an explosion of interest in studying interior scene synthesis. As demonstrated in [40, 29, 22, 39, 48, 15, 45, 46], automatically synthesizing plausible rooms would benefits various applications like virtual reality and augmented reality. The main goal of current scene synthesis methods is to coherently predict and arrange functional furniture shapes. The extensive professional layout designs provided by 3D-FRONT may be immensely valuable to support the development of learning-based methods for this synthesis task.

Our demonstration uses the state-of-the-art neural scene synthesis method of Wang et al. [40], where each 3D scene is represented in an orthographic top-down view, which constitutes depth, room mask, wall mask, object mask, and orientation. Their method trains a deep convolutional neural network to iteratively capture scene priors, so as to decide whether to add a next object, what category of object to add and where, and finally insert an instance of that object category with estimated rotation into the scene. Following [40], we conduct our experiment on two scene types, *i.e.,* bedroom (Bedroom, MasterBedRoom, and SecondBedRoom) and living room (LivingRoom and LivingDiningRoom), and remove the rooms whose width or length is larger than 6 meters. As a result, we obtain 6,230 bedrooms and 645 living rooms, with 6,070 / 485 rooms for training and 160 / 160 rooms for evaluation. We refer to [40] for more details on training and test settings.

We evaluate diversity of the synthesized results using converge (COV) and minimum matching distance (MMD) [1] measured by Chamfer Distance (CD) or Earth-Mover Distance (EMD) between scenes synthesized by models trained on 3D-FRONT and on SUNCG, respectively. The results were generated from empty rooms in the combined test set of 3D-FRONT and SUNCG. For each synthesized scene, we randomly sample 100K points and calculate these metrics against the ground truth. Recall that lower MMD and higher COV indicate better synthesis ability of a method. Quantitative comparisons in Table 4 show the dataset advantage of 3D-FRONT over SUNCG. A qualitative comparison is shown in Figure 6.

In addition, we conduct a user study on AMT where Turkers were asked to choose scenes synthesized by [40], from randomly choosen empty rooms, that are deemed to be more "plausible"; see supplementary material for details. From the user feedback, we find that layouts synthesized by the model trained on 3D-FRONT were chosen 64.8% of the time (vs. 35.2% for SUNCG). All these results strongly demonstrate the utility of our new dataset, over other alternatives, for the important scene synthesis task.

5.2. Texturing 3D Models in Indoor Scenes

The quality, richness, and compatibility of object textures in an indoor scene can greatly enhance its realism. The textured 3D models available from 3D-FRONT fulfill these very characteristics, and we expect our dataset to benefit the development of many data-driven scene texture synthesis algorithms. In comparison to texturing a single 3D object [28, 16, 25], doing the same to an object in the context of an indoor scene must take into account that scene context to ensure both quality and visual compatibility.

We extend a recent generative model for textured *meshes*, TM-Net [16], to the 3D scene texturing task. TM-Net represents 3D shape parts with their structural deformable boxes, thus enables to generate part-level structural texture atlases for the given untextured 3D shapes. When applying it to the 3D scene configuration, we enforce the texture coherence between 3D *objects* by randomly choosing a shape in the scene, extracting its texture's VGG feature, and finally using the feature to guide the generation of other objects' textures in the training setting. After training the generative models, we synthesize texture for a random shape, and use it as a condition for other objects' texture generation to keep the consistency.

We conduct a simple experiment to validate the advantage of 3D-FRONT, as training data for TM-Net, for the generation of chair textures given table textures as a condition or guidance. Comparisons are made to ShapeNet [7], which also contains textured 3D models and can serve as the training data. Figure 7 presents some qualitative results where the table-chair settings were sampled from dining rooms. TM-Net trained on 3D-FRONT tends to generate richer and more diverse textures, as can be verified by both a quantitative test and a user study. Specifically, the model trained on 3D-FRONT yields a LPIPS [44] score of 0.289, which outperforms its ShapeNet counterpart, which has a score of 0.215, where we recall that LPIPS is a measure of the diversity of generated textures. Our user study on AMT, where users were asked to select which generated textures were "richer", also shows that results by TM-Net trained on 3D-FRONT were selected 61.1% of the time (vs. 38.9% for ShapeNet); see supplementary material for more details.

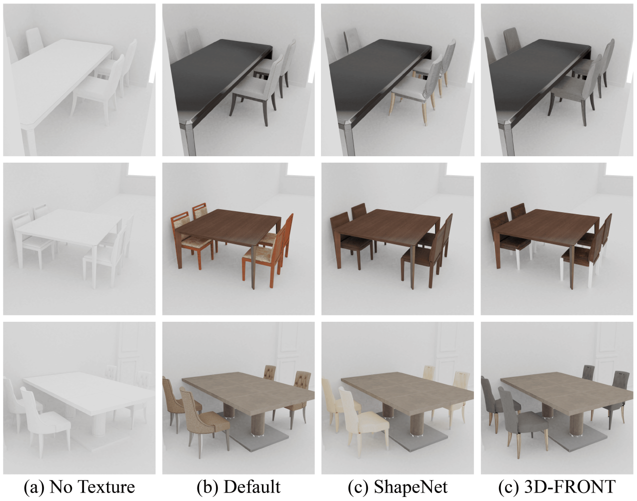


Figure 7: **Texturing 3D models in indoor scenes.** The default textures (b) were provided by the 3D-FRONT dataset. In (c) and (d), we show chair textures generated by TM-Net, conditioned on given textures for the table. The network was trained on ShapeNet (c) or 3D-FRONT (d).

6. Conclusion and future work

We present 3D-FRONT, a new large-scale dataset of synthetic 3D indoor scenes. Up to now, there have been a variety of 3D scene datasets established to serve different purposes. Some focus on photorealistic renderings of artist-created scenes, possibly with instance segmentations and per-pixel material and illumination ground truth data, while others acquire large volumes of raw scans of the world to drive research in 3D scene reconstruction and modeling. Compared to these efforts, 3D-FRONT offers the largest publicly available collection of professional designed room layouts instanced with high-quality textured CAD meshes.

One of our intentions was to fill a void in the vision and graphics community after SUNCG became unavailable. Yet, our dataset surpasses SUNCG in three aspects: professional vs. amateur layout designs, CAD model quality, and style compatibility. We demonstrate that these distinctive features enable several data-driven applications which were not well supported by other datasets. In the future, we will continuously improve 3D-FRONT by releasing an industrial render engine (AceRay) and providing much enriched texture and 3D geometry contents.

Acknowledgment

We would like to thank our colleges Yiyun Fei, Yu Zheng, Ying Li, Yi Liu, Peng Liu, Lin Ma, Le Weng, Xin Ma, Xiaohang Hu, and Qian Qian for their great help with the data preparation and image rendering. We also appreciate Alibaba Tianchi for managing the dataset so that it can be easily requested and downloaded.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 7

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5664–5673, 2019. 3

[3] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 3

[4] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543, 2016. 1, 2, 3

[5] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2CAD: Learning CAD model alignment in RGB-D scans. In *CVPR*, pages 2614–2623, 2019. 2

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 2, 3

[7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 8

[8] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12617–12626, 2019. 4, 5, 6

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNET: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1, 2, 3

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[11] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. *ACM Transactions on Graphics (TOG)*, 31(6):1–11, 2012. 3

[12] James Fogarty, Ryan S Baker, and Scott E Hudson. Case studies in the use of roc curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005*, pages 129–136, 2005. 5

[13] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. 4

[14] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *arXiv preprint arXiv:2009.09633*, 2020. 2, 4, 5, 6

[15] Qiang Fu, Xiaowu Chen, Xiaotian Wang, Sijia Wen, Bin Zhou, and Hongbo Fu. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017. 7

[16] Lin Gao, Tong Wu, Yu-Jie Yuan, Ming-Xian Lin, Yu-Kun Lai, and Hao Zhang. Tm-net: Deep generative networks for textured meshes. *arXiv preprint arXiv:2010.06217*, 2020. 8

[17] Alberto Garcia-Garcia, Pablo Martinez-Gonzalez, Sergiu Oprea, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Alvaro Jover-Alvarez. The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6790–6797. IEEE, 2018. 3

[18] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Understanding real world indoor scenes with synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4077–4085, 2016. 1, 2, 3

[19] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. SceneNN: A scene meshes dataset with annotations. In *3DV*, pages 92–101, 2016. 1, 2, 3

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[21] Hema S Koppula, Abhishek Anand, Thorsten Joachims, and Ashutosh Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *Advances in neural information processing systems*, pages 244–252, 2011. 3

[22] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)*, 38(2):1–16, 2019. 7

[23] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. *arXiv preprint arXiv:1809.00716*, 2018. 1, 2, 3

[24] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Sai Bi, Zexiang Xu, Hong-Xing Yu, Kalyan Sunkavalli, Miloš Hašan, Ravi Ramamoorthi, et al. Openrooms: An end-to-end open framework for photorealistic indoor scene datasets. *arXiv preprint arXiv:2007.12868*, 2020. 1, 2, 3

[25] Ricardo Martin-Brualla, Rohit Pandey, Sofien Bouaziz, Matthew Brown, and Dan B Goldman. GeLaTO: Generative Latent Textured Objects. In *European Conference on Computer Vision*, 2020. 8

[26] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2678–2687, 2017. 3

[27] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2, 3

[28] Amit Raj, Cusuh Ham, Connelly Barnes, Vladimir Kim, Jingwan Lu, and James Hays. Learning to generate textures on 3d meshes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 32–38, 2019. 8

[29] Daniel Ritchie, Kai Wang, and Yu-an Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6182–6190, 2019. 7

[30] Mike Roberts and Nathan Paczan. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding, 2020. 2, 3

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[32] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning Interaction Snapshots from Observations. *ACM Transactions on Graphics (TOG)*, 35(4), 2016. 3

[33] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 3

[34] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 6, 7

[35] 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

[36] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580. IEEE, 2012. 1, 2, 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[38] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020. 2, 3

[39] 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. 7

[40] Kai Wang, Manolis Savva, Angel X Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3, 7, 8

[41] Tomer Weiss, Alan Litteneker, Noah Duncan, Masaki Nakada, Chenfanfu Jiang, Lap-Fai Yu, and Demetri Terzopoulos. Fast and scalable position-based layout synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 25(12):3231–3243, 2018. 4

[42] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9068–9079, 2018. 3

[43] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pages 1625–1632, 2013. 1, 2, 3

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[45] Song-Hai Zhang, Shao-Kui Zhang, Wei-Yu Xie, Cheng-Yang Luo, Yongliang Yang, and Hongbo Fu. Fast 3d indoor scene synthesis by learning spatial relation priors of objects. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 7

[46] Shao-Kui Zhang, Wei-Yu Xie, and Song-Hai Zhang. Geometry-based layout generation with hyper-relations among objects. *Graphical Models*, page 101104, 2021. 7

[47] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5287–5295, 2017. 3

[48] Zaiwei Zhang, Zhenpei Yang, Chongyang Ma, Linjie Luo, Alexander Huth, Etienne Vouga, and Qixing Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM Transactions on Graphics (TOG)*, 39(2):1–21, 2020. 7

[49] 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. 1, 2, 3