Zillow Indoor Dataset: Annotated Floor Plans With 360^o Panoramas and 3D Room Layouts

Steve Cruz*†[1], Will Hutchcroft*[2], Yuguang Li[2], Naji Khosravan[2], Ivaylo Boyadzhiev[2], Sing Bing Kang[2]
[1]University of Colorado Colorado Springs [2]Zillow Group
scruz@uccs.edu {willhu,yuguangl,najik,ivaylob,singbingk}@zillowgroup.com

Abstract

We present Zillow Indoor Dataset (ZInD): A large indoor dataset with 71,474 panoramas from 1,524 ***real*** *unfurnished homes. ZInD provides annotations of 3D room layouts, 2D and 3D floor plans, panorama location in the floor plan, and locations of windows and doors. The ground truth construction took over 1,500 hours of annotation work. To the best of our knowledge, ZInD is the largest* ***real*** *dataset with layout annotations. A unique property is the room layout data, which follows a real world distribution (cuboid, more general Manhattan, and non-Manhattan layouts) as opposed to the mostly cuboid or Manhattan layouts in current publicly available datasets. Also, the scale and annotations provided are valuable for effective research related to room layout and floor plan analysis. To demonstrate ZInD's benefits, we benchmark on room layout estimation from single panoramas and multi-view registration.*

1. Introduction

In computer vision, 3D scene understanding from 2D images has received considerable attention due to its critical role in research areas such as robotics and mixed reality. Work has been done on automatic room layout estimation from single 360^o panoramas [52, 56, 50, 42, 56, 32], which typically depend on either a relatively small set of real images, e.g., 2,295 RGB-D panoramic images in MatterportLayout [56], or the use of synthetic data due to the scale needed for training, e.g., around 196k images in Structure3D [54]. Also, there are techniques for reconstructing floor plans from a sequence of such panoramas ([6, 34]), or RGB-D video [26]. In regards to real estate, approaches on floor plan generation from user-specified constraints [45, 16, 30] make use of architectural floor plan datasets. We believe there is a need for more properly annotated data of indoor layouts and complete floor plans.

Many real and synthetic datasets have been released. Current real datasets [52, 56] are limited both in size and variation due to challenges in *capturing* and *annotating* indoor spaces, both of which require *non-trivial logistics, resources* and *privacy concerns*. Synthetic datasets [54, 22] address the scale issue, but introduce a domain gap, which complicates generalization to real scenes.

In this paper, we introduce Zillow Indoor Dataset (ZInD)[1], a dataset containing large numbers of annotated 360^o panoramas, room layouts, and floor plans of *real unfurnished* residential homes[2]. For each panorama, its location within the floor plan and its associated 3D room layout are given (Figure 1). Also, locations of windows and doors are provided.

ZInD contains 1,524 homes, with a total of 71,474 panoramas, 21,596 room layouts, and 2,564 floor plans. We believe our dataset can facilitate research on room layout estimation, floor plan reconstruction from multiple panoramas, and floor plan analysis that include image features (e.g., scene graphs). To showcase ZInD's attributes, we ran experiments with state-of-the-art (SOTA) methods for room layout estimation and multi-view registration.

2. Related Work

In this section, we provide an overview of publicly available datasets. They are categorized in Table 1 based on the type of annotations, size, and attributes.

Depth and Semantic Annotations. Most current datasets focus on providing depth and semantic annotations. For example, Stanford 2D-3D-S [2] contains 552 RGB panoramic images collected from 6 large-scale indoor environments, including offices, classrooms, and other open spaces. The dataset provides annotations of depths, surface normals, semantic annotations, global XYZ images, and camera information. Another dataset, Matterport3D [7], has over ten thousand RGB-D panoramic images collected from 90 building-scale scenes with surface reconstructions, camera

*Equal contribution.
†This work was done when Steve Cruz was an intern at Zillow.

[1]ZInD and scripts are at https://github.com/zillow/zind
[2]Using unfurnished homes alleviates privacy concerns.

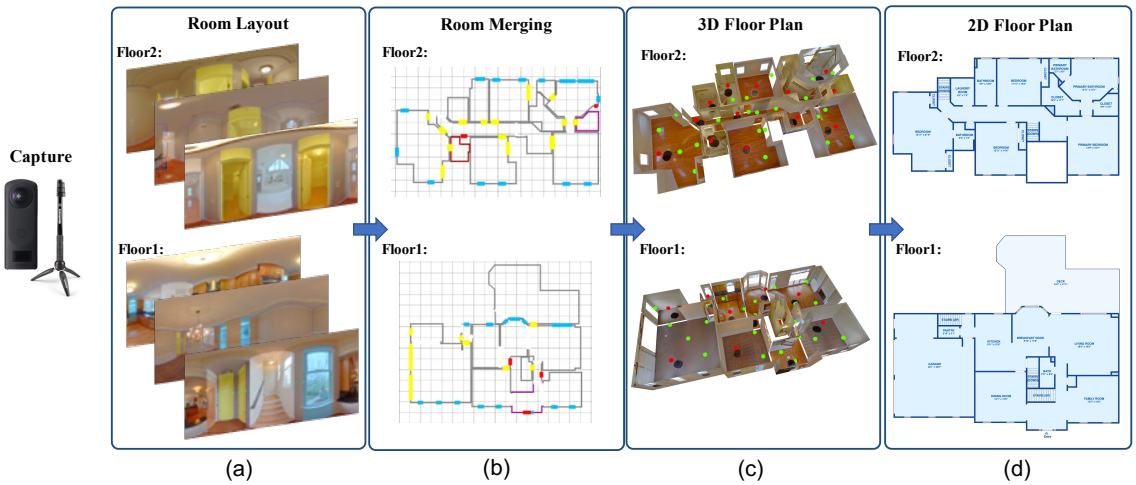


Figure 1. The Zillow Indoor Dataset dataset provides visual data covering *real* world distribution of *unfurnished* homes, including *primary* 360 panoramas with annotated room layouts, windows and doors, merged rooms, *secondary* localized panoramas, and final 2D floor plans. From left-to-right: (**a**) captured RGB panoramas with layout, windows and doors *annotations*, (**b**) merged layouts, (**c**) 3D textured mesh, where **red dots** indicate *primary* and **green dots** indicate *annotated secondary* panoramas, (**d**) final 2D floor plan "cleanup" *annotation*.

poses, and 2D and 3D semantic segmentation. Also, over 12 indoor environments are covered (e.g., bedroom, office, bathroom and hallway).

Gibson [47] has data from 572 buildings with a total of of 1447 floors to provide an environment for active agents. The Replica dataset [41] provides photorealistic 3D reconstructions of 18 indoor scenes with a dense mesh, high resolution HDR textures, per-primitive semantic class, instance information, and planar mirror and glass reflectors. NYU-Depth V2 [29] contains 1,449 RGB-D images with surface normal and semantic labels. Another RGB-D dataset, SceneNN [17], provides 100 scenes represented by triangle meshes with per-vertex and per-pixel annotations.

There are datasets with more densely sampled RGB-D data. For example, ScanNet [11] features RGB-D videos from 1,513 scenes of 707 unique indoor environments with 2.5 million frames. Included with the data are 3D CAD models, annotations of 3D camera poses, surface reconstructions, and semantic segmentation. Meanwhile, SUN3D [48] contains 415 RGB-D image sequences with camera pose and object labels.

Synthetic. Due to challenges in generating large datasets with real scenes, there is growing interest in synthetic datasets. For example, Structured3D [54] provides synthetic images with 3D structure annotations. The dataset contains 21,835 rooms in 3,500 scenes and more than 196k photorealistic 2D renderings of rooms. 3D structure annotations along with 2D renderings of scenes were automatically extracted from many professional interior designs. In contrast to datasets mostly working with cuboid rooms, Structured3D introduced a wide variety of room layouts.

InteriorNet [22] is another synthetic dataset extracted from professional interior designs and renderings at video frame rate. The dataset has one million furniture CAD models and 22 million interior layouts. SUNCG [40] contains synthetic 3D scenes (over 45k different scenes) with dense volumetric annotations. House3D [46] is a virtual dataset consisting of 45k indoor scenes equipped with a diverse set of scene types, layouts, and objects sourced from SUNCG. By converting SUNCG into an environment, House3D adds the flexibility for customization to other applications. Based on CAD, SceneNet [22] has over one million models and 20 million rendered images.

While synthetic models are getting more photorealistic, there remains a domain gap. This makes it difficult to extract optimal performance when testing on real inputs. Techniques for domain adaptation [10, 5] should help, but there is still no good substitute for real data.

ZInD fills this important gap by providing a reasonably large-scale dataset with panoramic images of *real unfurnished* homes to facilitate research on global structure understanding. To the best of our knowledge, our dataset has the largest number of panoramic images of real rooms compared to datasets with real indoor scenes. There are benefits to both synthetic and real data; our goal is to complement synthetic datasets by giving insights and sampling the distribution of real homes, thus providing a new benchmark.

Indoor Visual Localization. Image-based localization is of interest for indoor navigation or augmented reality due to the unavailability of GPS signals. There are far fewer datasets for indoor scenes than those for outdoors. Some examples include 7-Scenes [14] (RGB-D data), Tiara et.al [43] (277 RGB-D panoramic images), and RISE [38] (13,695 panoramas).

Only Floor Plans. Datasets in this category can be 2D or 3D. 2D floor plan generation typically relies on user specification of the floor plan, e.g., floor plan boundary [45], room counts and their connectivity [16], and bubble diagram that defines room connectivity [30]. One floor plan dataset is RPLAN [45] (used in [16]), which consists of 120,000 floor plans from real-world residential buildings in Asia. Another is LIFULL HOME[3], which contains five million real floor plans (seen in [30]). Yet another is CubiCasa5K [19], with 5,000 floor plan images having ground-truth annotations encoded in SVG vector graphics format. The annotations include object categories that appear in the floor plan.

	Name	Data type	# of images	Layouts	Applications
Real	Gibson [47]	RGB-D pano	572 scenes	–	O U D
	Replica [41]	CAD	18 scenes	C, M, N-M	O U D F
	SceneNN [17]	RGB-D	100	–	O U D
	CRS4/ViC[4]	RGB pano	191	C, M, N-M	O U L F
	SUN3D [48]	RGB-D	415	–	O U D
	PanoContext [52]	RGB pano	514	C	O U L
	2D-3D-S [2]	RGB-D pano	552	–	O U D
	LayoutNet [55]	RGB-D pano	1,071	C	O U D L
	NYU-Depth V2 [29]	RGB-D	1,449	–	O U D
	ScanNet [11]	RGB-D video	1513	–	O U D
	MatterportLayout [56]	RGB-D pano	2295	C, M	O U D L
	Matterport3D [7]	RGB-D pano	10,800	–	O U D
	ZInD (**Ours**)	RGB pano	71,474	C, M, N-M	S L F
Synthetic	SunCG [40]	CAD	45, 622	C, M, N-M	O U D F
	Structured3D [54]	CAD	196k	C, M, N-M	O U D S F
	SceneNet [22]	CAD	20M	C, M, N-M	O U D F
	InteriorNet [22]	CAD	22M	C, M, N-M	O U D F

Table 1. Overview of publicly available 3D indoor datasets. **O** (object detection), **U** (scene understanding), **S** (structured 3D modeling), **L** (layout estimation), **F** (floor plan), **D** (depth), **C** (cuboid), **M** (Manhattan), **N-M** (non-Manhattan).

3D floor plan generation methods make use of visual data; such data includes a dense set of panoramic images [33, 34, 6], an input stream of RGB-D images [26], panoramic RGB-D images [18], and partial 3D scans for merging [24]. For machine learning-based training, the dataset is typically specific to the method. For example, [26] captures their own dataset for training and verification. Cabral and Furukawa [6] use semantic information in panoramas in addition to structure-from-motion to generate a 3D texture-mapped floor plan.

Panoramic Images and Floor Plans. Datasets with panoramic images and layout information are the closest to ZInD. PanoContext [52] contains 514 panoramic images and provides mostly cuboid layouts of limited scene types, e.g., bedrooms and living rooms. MatterportLayout [56] selected a subset of RGB-D panoramic images from Matterport3D and extended them with general Manhattan layout annotations. This subset contains 2,295 panoramas within closed 3D space, and features only Manhattan 3D layouts. While Realtor360 [50] has over 2,500 indoor panoramas and annotated 3D room layouts, it is currently not publicly available. Also, Layoutnet [55] extended annotations for 2D-3D-S, and in combination with PanoContext, provides over 1,000 panoramas with room layout annotations.

[3] https://www.nii.ac.jp/dsc/idr/lifull

[4] http://vic.crs4.it/download/datasets/

3. Zillow Indoor Dataset

In this section, we introduce Zillow Indoor Dataset. First, we define our terminology and clarify the assumptions in our annotation process. Next, we describe how we acquire the panoramic images before detailing the annotation pipeline for generating the floor plans. Finally, we report important statistics. The different levels of annotations are illustrated in Figure 1.

3.1. Terminology and Assumptions

Camera Height. When generating ZInD, we assume the camera height is constant throughout the capture process for a home (could vary between homes). Our capture protocol for photographers includes this requirement.

Ceilings and Floors. We assume *Atlanta world* [39, 35], where the modeled 3D layout has horizontal floor, ceiling, and vertical walls. However, ZInD has non-flat ceilings as well. Annotators indicate a ceiling to be *flat* (a simple horizontal plane) or *non-flat* (ceiling consists of multiple planes, vaulted or is unfinished). Regardless, the height of the ceiling for each room is associated with its largest surface area. In addition, we do not support 3D modeling of the *non-flat* ceilings (only 3D planar approximation), stairs, and occluders such as cabinets and kitchen appliances.

Primary and Secondary Panoramas. Annotators were asked to select panoramas with the "best" views of entire rooms, (e.g. room centers, as *primary* and others near to walls/doors, where the layout might be harder to annotate, as *secondary*). For big open spaces, they are asked to use their best judgement to partition the space into disjoint parts, with each part represented by a primary panorama. Only primary panoramas are used to generate room layouts and subsequently floor plans. The rest (*secondary* panoramas) are localized, using a semi-automatic human-in-the-loop approach, within the layouts to provide denser spatial data.

3.2. Capture Process

To enable capturing entire home interiors at scale, we opted for sparse 360^o panoramic capture of every room in the home using a panoramic camera paired with an iPhone. To do so, photographers across 20 US cities were hired to do the capture and given specific instructions to ensure uniformity in the capture quality.

Hardware. We use the Ricoh Theta (V and Z1)[5] for panoramic capture, which has high quality image stitching technology and HDR support. Panoramas are captured with

[5] https://theta360.com/en/

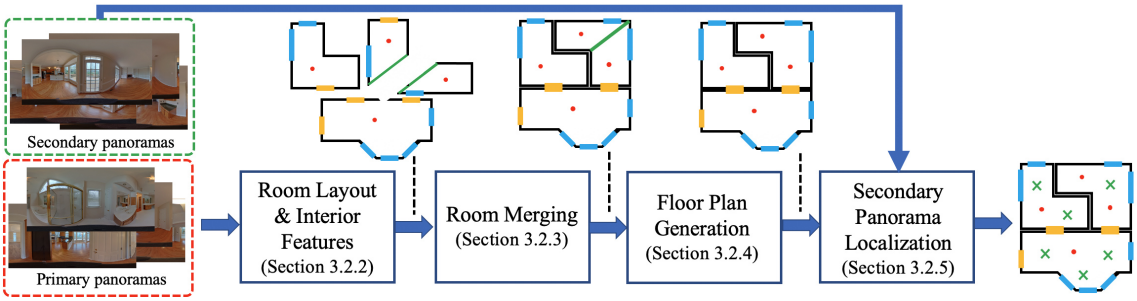


Figure 2. Zillow Indoor Dataset generation pipeline. The processes are all in 2D, but since the ceilings are assumed horizontally flat with known heights, we can infer 3D from 2D floor plans. Please see the text for more details of the pipeline. The **dots** in the floor plan indicate primary panorama locations while the ×'s are the secondary panorama locations. Windows, doors, and openings are color-coded in **blue**, **orange**, and **green**, respectively.

3 auto bracketing exposure configurations, default white balance settings, and IMU-based vertical tilt correction.

Protocol. To ensure uniformity in the quality of capture, photographers were given guidelines on lighting, number of captures per room (based on room size), and not appearing in the camera's view. An iOS application was developed for this purpose. Further, to ensure a fixed camera height during the capture, photographers are asked to keep a fixed tripod height throughout the capture process. To retrieve this scale, we added a procedure of capturing either a printed AprilTag [31] or a standard paper size on the floor.

3.3. Annotation Pipeline

Our annotation pipeline is shown in Figure 2. To generate 3D floor plans from panoramas, we developed an internal tool[6]. Our pipeline starts with pre-processing of panoramas. If necessary, trained annotators then verify and correct automatically generated room layout and interior wall features. Next, the verified room layouts are merged to form a *draft* floor plan using an automatic rank-and-propose algorithm to assist the human-in-the-loop process. An automatic global optimization and refinement step is subsequently applied to the draft floor plan, followed by human-assisted cleanup and verification process to generate the final version of the floor plan.

3.3.1 Pre-processing

Prior to annotation, we refine the IMU-based zenith correction by detecting dominant vertical image lines in all the captured panoramas. This is done using vanishing lines analysis similar to [52, 50, 42, 56]. The computed vanishing points are used to generate upright panoramas, which simplifies alignment between them. For consistency and scalability, we resize all panoramas to 2048×1024.

[6]Due to legal considerations, our tool cannot be made public.

3.3.2 Room Layout and Interior Features

As mentioned in section 3.1, our *primary* panoramas are the ones used to generate the floor plan. Thus, the first step is to generate a room layout from each primary panorama. Our internal tool is similar in spirit to [49], but with important production-level features to enable high throughput. The tool allows annotators to start from an initial automatic room layout prediction and Window, Door and Opening[7] (WDO) detection. All the predictions are based on continuous training and deployment of SOTA models for room layout estimation [42, 32] and object detection [37].

Annotators can also bypass the automation outputs by directly indicating main structural elements (floor, wall, ceiling boundaries, and WDO). The tool shows the evolution of room layout as the panorama is being annotated. Also, the tool allows annotators to enforce Manhattan constraints, or predefined corner angles, for the room layout.

3.3.3 Room Merging

Once all room layouts are generated, the next step is to combine them to produce a *draft* floor plan of the building. This is similar to the task defined in [25]; instead of partial RGB-D scans that can be subject to refinement, we use room layouts that are considered ground truth. There are two types of room merging used in our annotation process: (1) semi-automatic merging of different rooms to form a *draft* floor plan, and (2) automatic merging of what we call partial rooms to remove openings and create complete layout of big spaces.

Merging Different Rooms. Our tool includes an interactive room merging UI, which allows room layouts to "snap" to each other using WDO locations. Annotators may re-

[7]An opening is an artificial construct that divides a large room into multiple parts. See Figure 2 for an example that features openings. Openings allow a large room to be represented by multiple panoramas with no overlapping layouts. Note, openings are later processed for removal.

fine room location by visualizing the reprojection of its layout onto the current reference panorama and allowing pose changes in *image* space.

The initial "snap" alignment is done by automatic rank-and-propose pairing process. The annotated WDO features from the previous stage are used to generate room pair proposals. The ranking is done based on (1) minimizing room intersections, (2) maximizing loop closure, (3) maximizing multi-view alignment of semantic elements, and (4) producing the most axis-aligned floor plans (making use of the precomputed vanishing points).

Merging Partial Rooms. Large and complex spaces often exhibit partial layout occlusion and/or distant geometry relative to the panorama. Such a space is typically represented by multiple non-overlapping panoramas which are connected by openings (Figure 2). Such panoramas represent *partial* room layouts.

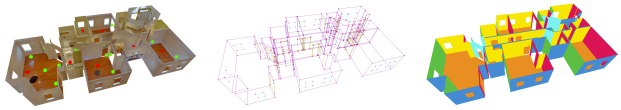
To produce *complete* layouts that support global analysis of large spaces, we merge these partial rooms. First, a graph associating pairs of openings is constructed based on proximity and alignment. Partial polygons are then repeatedly merged through this association graph until no further opening pairs exist.

The resulting set of partial polygons is used to write binary segmentation footprints into a joint floor segmentation of the entire space. Contour extraction and polygon simplification is then applied to extract a final combined layout from the joint segmentation footprint. The result is joint layout shared across multiple panoramas, each with a potentially partial view of a larger and/or complex space (Figure 4). The complete set of oriented room layouts for a home allows the extraction and visualization of textured 3D floor plans, primitives such as lines, junctions, planes and relationships between them (similar to [54]) as seen in Figure 3.

3.3.4 Floor Plan Generation

The next step is to take the draft floor plan from the room merging step and produce a slightly modified version, where the external walls form a closed 1D loop as defined in [27]. Generating the final floor plan is straightforward for the annotator, given all the previous steps previously described. They resolve slight inconsistencies, such as minor wall misalignment due to drift or annotation errors. They also clean up room labels, add missing spaces such as closets and stairs, and indicate unresolved spaces (such as those with high ceilings spanning multiple floors). Furthermore, outdoor spaces, removed from the image set to *reduce privacy concerns*, like patios and balconies are added in the final 2D representation as shown in Figure 1(d).

We developed a set of heuristics to facilitate this process, which we refer to as "cleanup". The cleanup tool takes the



(a) 3D Floorplan (b) Lines, Junctions (c) Planes

Figure 3. Examples of automatically extracted 3D structure annotations available as a result of our 3D floor plan annotation pipeline, similar to [54].

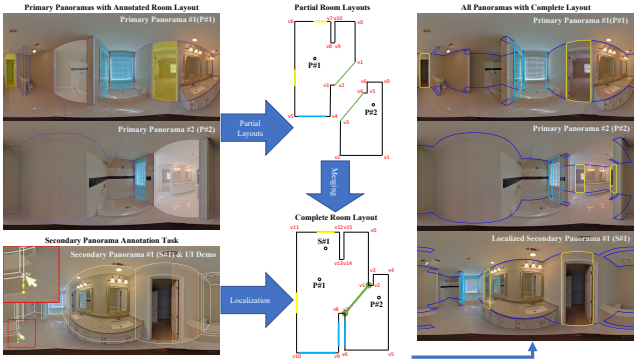


Figure 4. Complete room layout and secondary panorama localization. **Top/Center:** process of computing complete room layouts of complex spaces, based on merged annotations of partial layouts from each panorama's point of view. **Bottom left**: the secondary panorama localization task, where annotators can achieve pixel level accuracy by dragging the corners of an existing 3D layout (inherited from a reference, primary panorama) to snap to 2D image corners. **Right**: the complete layout from the point of view of all 3 panoramas localized on the same shared layout.

annotator's changes and produces a 2D floor plan representation with walls, room type text labels, and window and door locations. The tool makes use of global constraints to group nearby room boundaries into cleaned-up walls, refine their orientation, update room layouts based on wall refinement, and clean up text label placement. Details are given in the supplementary material.

3.3.5 Secondary Panorama Localization

As previously discussed, primary panoramas are used to generate the floor plans. For these panoramas, their poses relative to the floor plan are known. To generate ground truth pose for the secondary panoramas, we use a two-step approach named *LayoutLoc* to generate automatic localization proposal, followed by a human verification and refinement. We evaluate the success rate and accuracy of *LayoutLoc* in Table 6 and further compare it with a standard SfM approach (for the cases of 3 or more panoramas per-space). We found the average pose estimation error of around 8.5cm (and max error bounded by 1m), provides important initial localization that adds a significant amount of automation to

simplify the annotation process.

Our automatic localization generates relative pose estimation by (1) camera pose proposal, followed by (2) pose scoring based on *inferred* layout corners, and windows and doors bounding boxes.

Proposals for pose are generated by estimating room layout [42] corners followed by windows and doors detection [36] for generating matching hypothesis, thus the name *LayoutLoc*. More specifically, proposals are generated based on corner point snapping and image horizontal vanishing point alignment. Each proposal is scored based on how well the re-projected *inferred* layout of one room fits the reference room, and how well *predicted* windows and doors overlap, similar to [9]. We pick the pose that maximizes the mutual layout and windows and doors alignments. Users can accept and refine or select another pair. When a predicted camera pose is reviewed by the annotator, a re-projection-based mechanism is used to refine the secondary panorama location; this is similar to the procedure for merging different rooms in Section 3.3.3. An illustration of the pixel-level refinement process is shown in Figure 4.

We assume camera height is constant throughout the capture process and that the upright orientation has been determined accurately. We solve for the camera horizontal position (x, y) and horizontal orientation θ_z. This reduces the degrees of freedom in the relative pose estimation problem to only three, which substantially increases robustness while reducing pose ambiguity.

3.4. Dataset Statistics

Our dataset involves *unfurnished* residential homes from 20 US cities. The cities span coast-to-coast across 12 states, where the top 5 constitute 75% of the data (8.5%, 13.3%, 16.8%, 16.8%, 19.1%); the remaining 25% are uniformly distributed across another 7 states. In terms of number of floors: 42.8% are single floor, 46.7% are 2-story, and 9.8% are 3-story single family homes. The remaining 0.6% have 4 stories and above, e.g., town-homes. ZInD contains data from 1,524 homes, providing a total of 71,474 panoramas of 26,341 annotator spaces. For each home, we provide a wide range of annotations that include room type (kitchen, bathroom, etc.), camera poses, bounding box annotations for windows and doors, 3D room layouts, and 3D and 2D floor plans with actual dimensions. General statistics are shown in Table 2.

An important aspect of ZInD is that it follows the real world distribution of room layouts. Table 3 shows the statistics of room layout types. Also, we compare the number of room corners to those in other available datasets (Table 4). Notice that our numbers rival or exceed those of Structured3D [54], which is a synthetic dataset with about twice as many homes.

ZInD offers a wide range of room layout complexity, which poses new research challenges in room layout estimation (e.g., for non-Manhattan layout estimation). By providing at-scale real scenes with high quality annotations, we believe the dataset will promote and facilitate these new research directions.

Feature	Total	Avg per home
# panoramas (pri)	31,880	20.90
# panoramas (sec)	39,594	25.98
# floor plans	2,564	1.68
# annotator spaces	26,341	17.28
# rooms	21,596	14.17
# windows	18,658	12.24
# doors	46,657	30.61

Table 2. Statistics for 1,524 homes and 71,474 panoramas. **pri** = primary, **sec** = secondary, "# annotator spaces" refers to spaces identified by annotators (which include closets and hallways), and "# rooms" refers to complete room layouts.

	Cuboid	Manhattan-L	Manhattan-General	Non-Manhattan
# Layouts	11,471	4,273	3,144	2,708

Table 3. Statistics on different room layout types. Since L-shaped layouts are common, we report that separately from others that are also Manhattan. Those that are non-Manhattan typically have room corner angles of 135^o.

# Corners	4	5	6	7	8	9	10+	Total
MatterportLayout	1,211	0	501	0	309	0	274	2,295
Structured3D	13,743	52	3,727	30	1,575	17	2,691	21,835
ZInD	11,544	984	3,493	371	1,335	145	3,724	21,596

Table 4. Comparison of room layout count based on number of room corners.

4. Experiments

To highlight important properties of ZInD, we ran baseline experiments on room layout estimation and multi-view camera registration.

4.1. Layout Estimation

Baseline. To showcase specific challenges of room layouts coming from real homes, we benchmarked with a SOTA room layout estimation method with source code available (training & testing) at the time of submission [42][8]. HNet represents room layouts as three 1D vectors, two vectors representing the position of the floor-wall and ceiling-wall boundaries, and the third representing the existence of wall-wall junctions. When training or testing, we follow the data preprocessing and parameters outlined by [42]. In particular, for this evaluation, we similarly operate on 1024x512 resolution images.

[8] We attempted to contact the authors of AtlantaNet [32] and DuLa-Net [50] to obtain training code for further benchmarking; however, none was available by submission.

Train/Val/Test Splits. When creating our training split, we analyzed the distribution of important data attributes in order to prevent bias and ensure an accurate benchmark. Further, in order to avoid alternate view-points of testing panoramas being included in the training set, we split the dataset at the per-home level. In total, the splits consisted of 10,305, 946, and 889 panoramas for training, validation, and testing, respectively.

For the baseline evaluation, we train and evaluate on a subset of our dataset to closely align with [54]. As described in Section 3, our layout annotations span the full range of complexity of real homes, including large, complex, open spaces, which typically feature room layouts that wrap around corners and have significant self-occlusion. As a result, for the purposes of training and evaluation with HNet, we separated our layout annotations into those that had reasonable expectation of recovery from a single view point ("simple"), and those that do not ("complex").

We define simple annotations as those which do not have any contiguous series of occluded corners. We make this distinction as HNet's Manhattan post-processing is capable of inferring simple occluded corners in order to complete a Manhattan layout, whereas no single perspective method can handle more extensive occlusion. We believe that the annotations for large complex spaces will spur new lines of research, such as multi-panorama layout estimation.

Evaluation Metrics. To align with existing layout estimation literature, we used the standard metrics used by [56, 42, 32], namely, 3D IoU, 2D IoU, and δ_i. In addition to these, we compute matches between predicted and ground truth corners at a threshold of 1% of the training image width to obtain per-image precision and recall, from which we compute the F-score. We propose this metric as an alternative to IoU. Similar to [44] in single-view 3D reconstruction, during training and evaluation of layout estimation models for real-world applications, we found that IoU is often insufficient, especially as the layout complexity increases.

Results. As shown in Table 5, when evaluated on the entirety of the "simple" data test set (*All Data*), HNet demonstrates overall performance comparable to evaluations seen in other works [54, 42]. In particular, high IoU and F-scores are observed for cuboid and general four cornered shapes. Importantly, ZInD extends beyond this distribution to layouts which better represent real homes, such as non-cuboid and non-Manhattan. Specifically, we report metrics for these types.

Further, the typical data assumptions, such as that of Manhattan and Atlanta World, enforce a flat ceiling height; however, this is not indicative of the real world. As such, we report metrics on both flat and non-flat ceilings. Non-flat ceilings, such as vaulted or piece-wise-planar, present a challenge to methods designed to identify single orthogonal ceiling planes exclusively, as they obfuscate the recognition of the desired layout contour. This challenge is demonstrated in the computed metrics, for example, by the significant reduction in F-score for non-flat ceilings. In general, the substantial drop in performance seen for challenging, yet real-world-typical, layouts suggests the need for further research in this area, for which our dataset provides a useful platform.

Config	3D IoU	2D IoU	δ_i	F-score
All Data	85.98	87.57	93.70	80.69
Flat Ceiling	87.43	88.78	94.00	83.98
Non-Flat Ceiling	80.86	83.32	92.64	69.00
4 Corners	88.21	89.68	94.28	87.34
6 Corners	84.94	86.69	95.29	77.87
8 Corners	81.73	83.72	93.78	65.89
10+ Corners	79.82	81.21	94.29	65.24
Odd Corners	82.84	84.34	83.21	69.95
Cuboid	88.24	89.72	94.37	87.47
Manhattan-L	85.06	86.85	96.44	79.99
Manhattan-General	82.28	83.91	95.54	69.34
Non-Manhattan	81.80	83.48	85.46	64.72

Table 5. Evaluation of HNet under different configurations, **(1)** all data, **(2)** ceiling variation, **(3)** # of corners, and **(4)** room layout type. All configurations branch from *All Data*.

4.2. Multi-View Registration

Existing techniques for floor plan generation from 360^o indoor panoramas [6, 34, 33] typically involve using some standard SfM technique for layout and pose estimation. However, wide-baseline SfM and inferred 3D cues from a few indoor images are usually very sparse and noisy [3, 20], even in the context of the increased FoV provided by the 360 camera [35].

ZInD poses an interesting challenge for floor plan recovery from panoramas: homes are *unfurnished* (with significantly less texture), our sampling of panoramas throughout each home is sparse, and there is no guarantee of significant visual overlap between spatially adjacent panoramas. Thus, we evaluate how conventional SfM performs on our dataset. Additionally, we show that our information on layout and window and door positions helps to improve localization (similar to [4, 3, 9]); this localization approach, *LayoutLoc*, was described in Section 3.3.5.

In our evaluation, we sample 13k room layouts of different complexities that contain at least 2 panoramas. Each group (clique) has an average of 2.9 panoramas and an average and maximum pairwise distances of 2m and 5m, respectively. These spatial samplings are comparable to [6, 33].

We used OpenMVG [28], an open-source SfM framework that has a mode that accommodates upright 360^o cameras [1, 8, 21]. We ran incremental SfM [28] as one baseline

Type	$\#_c$	$\#_s$	$\bar{\#}_p/c$	SfM			*LayoutLoc*			
				$\%_{>2}$	$\bar{x}$[cm]	$\bar{s}$[cm]	$\%_{>2}$	$\%_{=2}$	$\bar{x}$[cm]	$\bar{s}$[cm]
Overall	13158	25531	2.94	0.555	**3.29**	0.83	**0.933**	**0.905**	8.50	11.28
Living Room	1365	3443	3.53	0.632	**3.65**	0.93	**0.965**	**0.931**	9.62	11.01
Basement	143	284	2.99	0.412	**1.05**	0.08	**0.851**	**0.802**	11.73	13.41

Table 6. Localization accuracies for SfM and *LayoutLoc*. $\#_c$: numbers of cliques, $\#_s$: numbers of secondary panoramas. $\bar{\#}_p/c$: average number of panoramas per clique. $\%_{>2}$: success rate of localized panoramas for clique sizes greater than 2. $\%_{=2}$: for clique sizes of 2. Note that SfM is incapable of localizing panoramas of clique size 2 (6k out of 13k samples) due to scale ambiguity. The 2 room types presented have the highest and lowest $\%_{>2}$ score for *LayoutLoc*. $\bar{x}$ and $\bar{s}$: mean and standard deviation of spatial error from the estimated camera position to the ground truth camera position in cm.

(for all cliques of size at least 3); we also ran our rank-and-propose localization algorithm (*LayoutLoc*) for comparison. We further report accuracy of *LayoutLoc* on cliques of size 2: an important property of our approach is that the inferred floor points remove the scale ambiguity due to a known camera height (Section 3.2), which is a fundamental problem in uncalibrated two-view geometry [15].

Results are shown in Table 6. We define the success rate as the percentage of localized panoramas with distance errors smaller than 1m and an angular error less than 1°. This comparison demonstrates that the application of conventional SfM on ZInD would be problematic without using additional semantic hints that are available.

5. Discussion

In this section, we highlight interesting aspects of our dataset as well as its limitations. Further details regarding the iOS application, our internal annotation tool, room merging, *LayoutLoc*, the distribution of room types, and various examples of what ZInD provides can be found in the supplementary material.

ZInD represents real *unfurnished* residential homes. This can address the concern raised in Structured3D [54] about accuracy of room layout annotations in other real datasets, to the extent possible in real scenes. Also, we believe that having *unfurnished* rooms provide a unique opportunity for delving into research problems that require empty rooms as ground truth (e.g., cluttered room prediction [53], emptying indoor spaces [51], or photo-realistic restaging and relighting [13, 12, 23]).

Another interesting aspect of ZInD is the non-standard room layouts, mostly in the form of non-flat ceilings and curved walls. To the best of our knowledge, no other real dataset provides these types of annotations, which may be valuable for further investigations into room layout estimation to include rooms with more challenging layouts.

Within our layout annotations, we include those of open floor plans, where semantic distinctions, such as "dining room", "living room" and "kitchen", are not always geometrically clear. As such, we adopt a definition of rooms as the regions our annotators identify through label placement in the "cleanup" phase. In this way, one open-space layout annotation may contain multiple rooms as regional designations. These type of room labels and locations are included in our released data and can be seen in the cleaned up floor plans Figure 1(d).

A noteworthy issue is that of annotator error. Since ground truth is based on human input, they will not be perfect. However, this has precedence in the many datasets that are also based on manual annotation. In our case, in addition to possible errors in room layout generation, there is the issue of conformity in determining if a room contains a non-flat ceiling height. Here, it becomes a judgment call that may vary across different annotators. If the ceiling is judged to be non-flat, its ceiling height is determined to one that dominates, and hence the 3D room layout is only approximate. About 32.5% of all rooms are tagged to have complex ceilings.

Legal and Privacy Notes. To ensure the academic research community is able to access our data unencumbered, we worked with our legal team to ensure that images captured and the data provided can be publicly utilized. There were extensive discussions that included where the images were captured and there is due diligence in ensuring privacy. Also, we automatically detected people and outdoors, and remove panoramas as appropriate.

6. Conclusion

In this paper, we introduce Zillow Indoor Dataset, a new extensive dataset. It consists of 71,474 panoramas from 1,524 *real unfurnished* residential homes across 20 US cities, 2,564 floor plans, and annotations of 18,648 windows and 46,657 doors. Also, the dataset consists of 21,596 room layouts, with panorama locations within them. Over 1,500 hours of annotation work was spent to create this dataset.

The dataset reflects a realistic distribution of layout complexities that include a significant number that are non-cuboid and non-Manhattan. While the featured homes are unfurnished, we have shown that mapping the (sparse) panoramas to floor plans remains a research challenge. We hope that ZInD will inspire new work on analyzing layouts and floor plans in the context of panoramas, including their semantic content such as windows and doors.

Acknowledgments. We are grateful to Pierre Moulon and Lambert Wixson for discussions on ZInD. In addition, Ethan Wan has been very helpful in processing and validating data in preparation for release.

References

[1] Mohamed Aly, Google Inc, and Jean yves Bouguet. Street view goes indoors: Automatic pose estimation from uncalibrated unordered spherical panoramas. 7

[2] 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. 1, 3

[3] Sid Ying-Ze Bao, A. Furlan, Li Fei-Fei, and S. Savarese. Understanding the 3d layout of a cluttered room from multiple images. *IEEE Winter Conference on Applications of Computer Vision*, pages 690–697, 2014. 7

[4] Yingze Bao and Silvio Savarese. Semantic structure from motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 7

[5] Sai Bi, Kalyan Sunkavalli, Federico Perazzi, Eli Shechtman, Vladimir G. Kim, and Ravi Ramamoorthi. Deep cg2real: Synthetic-to-real translation via image disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[6] Ricardo Cabral and Yasutaka Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 3, 7

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

[8] Sunglok Choi and Jong-Hwan Kim. Fast and reliable minimal relative pose estimation under planar motion. *Image and Vision Computing*, 69:103 – 112, 2018. 7

[9] Andrea Cohen, Johannes Schönberger, Pablo Speciale, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. Indoor-outdoor 3d reconstruction alignment. In *2016 European Conference on Computer Vision (ECCV)*, October 2016. 6, 7

[10] Gabriela Csurka, editor. *Domain Adaptation in Computer Vision Applications*. Springer, 2017. 2

[11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 2, 3

[12] Marc-Andre Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagne, and Jean-Francois Lalonde. Deep parametric indoor lighting estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 8

[13] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-Francois Lalonde. Fast spatially-varying indoor lighting estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[14] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time RGB-D camera relocalization. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179, 2013. 2

[15] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004. 8

[16] Ruizhen Hu, Zeyu Huang, Oliver Van Kaick, Hao Zhang, and Hui Huang. Graph2plan: Learning floorplan generation from layout graphs. *ACM SIGGRAPH*, 2020. 1, 3

[17] 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 *International Conference on 3D Vision (3DV)*, pages 92–101, 2016. 2, 3

[18] Satoshi Ikehata, Hang Yan, and Yasutaka Furukawa. Structured indoor modeling. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[19] Ahti Kalervo, Juha Ylioinas, Markus Häikiö, Antti Karhu, and Juho Kannala. Cubicasa5k: A dataset and an improved multi-task model for floorplan image analysis, 2019. 3

[20] Zhizhong Kang, Juntao Yang, Zhou Yang, and Sai Cheng. A review of techniques for 3d reconstruction of indoor environments. *ISPRS International Journal of Geo-Information*, 9(5):330, May 2020. 7

[21] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Closed-form solutions to minimal absolute pose problems with known vertical direction. In Ron Kimmel, Reinhard Klette, and Akihiro Sugimoto, editors, *Computer Vision – ACCV 2010*, pages 216–229, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. 7

[22] 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. In *British Machine Vision Conference (BMVC)*, 2018. 1, 2, 3

[23] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 8

[24] Cheng Lin, Changjian Li, and Wenping Wang. Floorplan-Jigsaw: Jointly estimating scene layout and aligning partial scans. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[25] Cheng Lin, Changjian Li, and Wenping Wang. Floorplan-jigsaw: Jointly estimating scene layout and aligning partial scans. In *International Conference on Computer Vision (ICCV)*, 2019. 4

[26] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floornet: A unified framework for floorplan reconstruction from 3D scans. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 3

[27] Chen Liu, Jiajun Wu, Pushmeet Kohli, and Yasutaka Furukawa. Raster-to-vector: Revisiting floorplan transformation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[28] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016. 7

[29] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, 2012. 2, 3

[30] Nelson Nauata, Kai-Hung Chang, Chin-Yi Cheng, Greg Mori, and Yasutaka Furukawa. House-GAN: Relational generative adversarial networks for graph-constrained house layout generation. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[31] Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3400–3407, May 2011. 4

[32] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. AtlantaNet: Inferring the 3D indoor layout from a single 360 image beyond the Manhattan world assumption. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 4, 6, 7

[33] Giovanni Pintore, Fabio Ganovelli, Ruggero Pintus, Roberto Scopigno, and Enrico Gobbetti. 3D floor plan recovery from overlapping spherical images. *Computational Visual Media*, 4, November 2018. 3, 7

[34] Giovanni Pintore, Fabio Ganovelli, Alberto Jaspe Villanueva, and Enrico Gobbetti. Automatic modeling of cluttered multi-room floor plans from panoramic images. In *Pacific Graphics*, 2019. 1, 3, 7

[35] Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. State-of-the-art in automatic 3d reconstruction of structured indoor environments. *Computer Graphics Forum*, 39(2):667–699, 2020. 3, 7

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(6):1137–1149, June 2017. 6

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 4

[38] Carlos Sánchez-Belenguer, Erik Wolfart, and Vítor Sequeira. RISE: A novel indoor visual place recogniser. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 265–271, 2020. 2

[39] G. Schindler and F. Dellaert. Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, 2004. 3

[40] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[41] 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. 2, 3

[42] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1047–1056, 2019. 1, 4, 6, 7

[43] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 2

[44] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? *CoRR*, abs/1905.03678, 2019. 7

[45] Wenming Wu, Xiao-Ming Fu, Rui Tang, Yuhan Wang, Yu-Hao Qi, and Ligang Liu. Data-driven interior plan generation for residential buildings. *ACM SIGGRAPH Asia*, 2019. 1, 3

[46] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3D environment. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[47] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[48] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *International Conference on Computer Vision (ICCV)*, pages 1625–1632, 2013. 2, 3

[49] Shang-Ta Yang, Chi-Han Peng, Peter Wonka, and Hung-Kuo Chu. Panoannotator: A semi-automatic tool for indoor panorama layout annotation. In *SIGGRAPH Asia Posters*, pages 1–2, December 2018. 4

[50] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single RGB panorama. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3, 4, 6

[51] Edward Zhang, Michael F. Cohen, and Brian Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016)*, 35(6), 2016. 8

[52] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3D context model for panoramic scene understanding. In *European Conference on Computer Vision (ECCV)*, pages 668–686, 2014. 1, 3, 4

[53] Yunhan Zhao, Shu Kong, Daeyun Shin, and Charless Fowlkes. Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[54] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 5, 6, 7, 8

[55] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2051–2059, 2018. 3

[56] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. 3D Manhattan room layout reconstruction from a single 360^o image. *arXiv preprint arXiv:1910.04099*, 2019. 1, 3, 4, 7