NadirFloorNet: reconstructing multi-room floorplans from a small set of registered panoramic images

Giovanni Pintore
CRS4 & National Research Center in HPC,
Big Data and Quantum Computing, Italy

Uzair Shah
HBKU, Quatar

Marco Agus
HBKU, Quatar

Enrico Gobbetti
CRS4 & National Research Center in HPC,
Big Data and Quantum Computing, Italy

Abstract

We introduce a novel deep-learning approach for predicting complex indoor floor plans with ceiling heights from a minimal set of registered 360° images of cluttered rooms. Leveraging the broad contextual information available in a single panoramic image and the availability of annotated training datasets of room layouts, a transformer-based neural network predicts a geometric representation of each room's architectural structure, excluding furniture and objects, and projects it on a horizontal plane (the Nadir plane*) to estimate the disoccluded floor area and the ceiling heights. We then merge and process these Nadir representations on the same floor plan, using a deformable attention transformer that exploits mutual information to resolve structural occlusions and complete room reconstruction. This fully data-driven solution achieves state-of-the-art results on synthetic and real-world datasets with a minimal number of input images.*

1. Introduction

The automatic 3D reconstruction of indoor structures from purely visual input is a very active visual computing research topic [28]. The focus is on creating quick and efficient methods to reconstruct the permanent architectural structure of common environments, like homes, offices, and public buildings [14], by exploiting their regularities to aid reconstruction from incomplete or noisy data [28]. In this context, 360° photography is one of the most prevalent capture methods [25, 33, 41].

In this work, we address the challenge of reconstructing 3D floorplans from a minimal set of mutually registered 360° images of cluttered rooms. This scenario is representative of common practices in the real estate industry, where millions of casual users generate and share virtual tours by capturing about a single image per room with consumer-grade cameras, exploiting a small amount of mutually visible space, e.g., through open doors, for registration [13, 18, 35].

In this context, the use of 360° images allows mutual registration even with a minimum number of shots compared with the use of perspective images [15]. However, the amount of visible features and consequently the three-dimensional information present is very sparse [18]. Leveraging this sparse data to automatically generate floorplans could enable a range of important applications, including enhanced navigation during virtual tours, improved price estimation, and integration with building information models [7]. However, this minimalistic capture approach poses significant challenges due to the sparsity and ambiguity of available information (Sec. 2).

Recent deep-learning approaches have obtained important successes in 3D inference tasks by exploiting available large-scale building datasets of panoramic images coupled with annotated layouts, both synthetic [56] and real [7] (Sec. 2). However, while good performance may be achieved for pixel-wise depth inference (e.g., [1, 46]), 3D room layout reconstruction from monocular input is still limited by the need to extrapolate large portions of the invisible structure, which can be occluded not only by objects but by the structure itself [33], forcing the incorporation of heavy priors and heuristic post-processing [16, 45, 60]. Moreover, inference at the individual room level may produce inconsistent global results, especially in invisible areas, leading to the need for specialized integrated model computation methods [28]. However, state-of-the-art multi-room reconstructors usually require the availability of reliable density maps for their analyses [5, 53]. These maps, accumulating the occurrence of 3D points projected onto the floorplan, are built from dense point clouds that are hard

to generate with sufficient precision using single-view inference from purely visual data.

We overcome these challenges using a bottom-up approach, in which information on the architectural room structure is inferred for each image before being used, globally, to predict the multi-room shape from multi-view fusion. Leveraging the availability of annotated datasets of room layouts and exploiting the vertical walls assumption, we train a transformer-based neural network to predict, from a single panoramic image of a cluttered room, the room's geometric structure – excluding furniture and objects – and its projection onto a horizontal plane (the *Nadir plane*). This process yields, for each room, an estimate of ceiling heights and a map of the disoccluded floor area (the *Nadir shape*). The Nadir shapes of all rooms within an environment are then integrated into a consistent, unified floor plan using a deformable attention transformer, which employs mutual information to resolve structural occlusions. Our main contributions are the following:

- We introduce a transformer-based neural network that predicts, from an equirectangular image of a cluttered room, the depth of the emptied scene and its projection onto the Nadir Plane (Sec. 4). It combines gated convolution for residual feature extraction with gravity alignment for feature compression. By transforming and projecting the uncluttered Euclidean depth and incorporating indoor-specific transformations and loss functions, we obtain a metrically consistent regularized and self-completed Nadir shape, only assuming vertical walls rather than imposing heavy constraints, such as MWM or AWM (Manhattan or Atlanta World) [27, 52]. Focusing on learning to infer the uncluttered scene proves more effective than analyzing the depth of the cluttered scene for structural predictions. While the Nadir shape is not a full room layout, its accuracy is comparable to that of dedicated layout prediction methods, but without requiring additional post-processing or heuristic completion [45].
- We join all the Nadir shapes in the same Nadir floorplan, exploiting a deformable transformer network [57] to process inter-relations, completing self-occluded parts, and refining room shapes (Sec. 5). In contrast to previous work, we do not take a dense multi-view occurrences map as input to the fusion step [5, 41, 53], but we start from the self-completed *Nadir shape* information extracted at the single image level to drive the solution of the final floorplan with ceiling height (Sec. 6.3), supporting reconstructions with a single image per room.
- We propose a fully data-driven, end-to-end solution, termed *NadirFloorNet*, that integrates into a single deep network architecture per-image Nadir shape inference with global multi-room floor plan reconstruction. *NadirFloorNet* directly maps a set of panoramic equirectangular images, each with its associated reference frame, to a complete floor plan with room heights, without any intermediate processing steps (Sec. 3). Although NadirFloorNet functions as a single network during inference, the Nadir shape prediction module is trained first and acts as a pre-trained module during the training of the floor plan module. Given the differentiability of the entire network, we support future fine-tuning, allowing all combined network parameters to be learnable.

We evaluated our method on large-scale synthetic and real benchmarks [7, 56]. Our results (Sec. 6) demonstrate that *NadirFloorNet* outperforms current state-of-the-art approaches starting from a minimal set of panoramic images.

2. Related work

Indoor reconstruction is a broad research topic. We analyze only the closely related approaches, referring the reader to established and recent surveys for wider coverage [28, 33].

Depth and layout estimation from panoramic images. 360° images are increasingly used for data-driven depth estimation in indoor spaces for their ability to capture the full surroundings. Their exploitation includes adapting perspective methods through spherical convolutions [24, 43, 44, 47, 58], joint processing in mixed equirectangular and cube-map projections spaces [49], leveraging perspective views sampled on panoramic images before combining depth maps using transformers [1, 20, 34], as well as exploiting gravity-aligned features for direct processing in equirectangular space [29, 30, 45]. Our approach emphasizes the structural nature of the desired output [17] by seeking to obtain the depth of the permanent structure deprived of the various objects in the room, combining and extending image inpainting and structural reconstruction concepts [29]. Instead of computing per-pixel depth data, structural extraction from single panoramas can also be obtained by layout estimation methods targeting the production of seamless 3D boundary surfaces in case of self-occlusions. Prominent examples include *LayoutNet* [59], which predicts the corner probability map and boundary map directly from a panorama, and *HorizonNet* [45], which simplifies the layout as three 1D vectors before extracting the 2D layout by fitting MWM segments to the estimated corner positions, as well as a breed of transformer-based approaches that project the equirectangular input image to planar surfaces [16, 27, 32, 50, 52, 55]. All these methods, however, require heavy preprocessing, such as detection of main MWM directions from vanishing lines analysis and related image warping [19, 54, 60], or complex layout post-processing, such as MWM regularization of detected features [37, 45, 52, 59]. In this work, we propose, instead, a fully data-driven approach for estimating the Nadir plane footprint and 3D cues, not from the cluttered RGB image but from its uncluttered depth, appropriately transformed

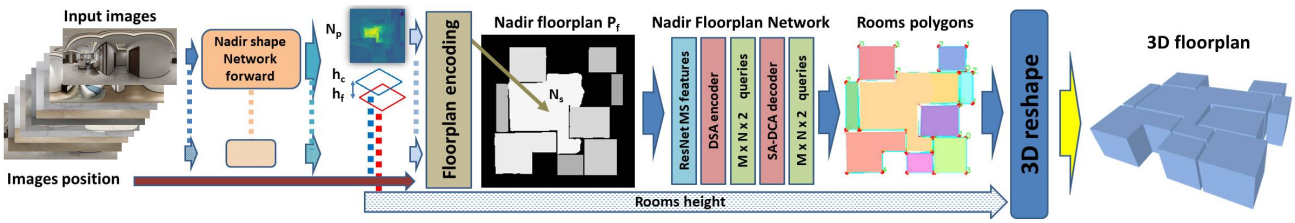


Figure 1. **Overview**. For each input panoramic image, a Nadir shape network forward pass infers a Nadir shape and the heights of floor-ceiling planes (see Sec. 4). These representations are then merged into a fixed-size Nadir floorplan P_f, exploiting the available mutual registration. The Nadir floorplan is processed through the Nadir floorplan Network (see Sec. 5) to predict the 2D polygonal multi-room floorplan. Finally, a 3D floorplan is recovered by extruding the 2D plan using the room heights recovered by the Nadir shape network.

(Sec. 4), deferring to multi-view computation the definition of consistent layout details also in invisible areas, without imposing MWM priors or requiring the hallucination of non-visible shapes. Our method only requires overlap if needed for mutual registration, enabling reconstruction with most points seen only once – typically using one image per room registered via a few feature points from open doors. Unlike multi-view approaches that rely on multiple images per room, even if only a few [3, 39, 40, 51], we support a single image per room by estimating monocular uncluttered depth before merging.

Data-driven floorplan reconstruction. Early approaches combined low-level image processing with geometric reasoning and energy minimization solvers to extract room layouts [2, 9, 14, 22, 26, 38]. Many recent solutions adopt a hybrid approach where low-level primitives detected by neural networks are assembled by optimization techniques into the final models. These include methods that detect room corners and then generate wall segments through integer programming [21], methods that detect room segments with Mask R-CNN [12] and reconstruct individual room polygons by sequentially solving shortest path problems [4, 23], approaches that uses Monte-Carlo Tree-Search to select room proposals [41], and diffusion techniques that generate plausible room arrangements by combining graph neural networks with constrained diffusion [10, 36]. In contrast to the hybrid solutions, several recent approaches employ, like ours, an end-to-end deep-learning architecture. In particular, HEAT [5] integrates the DETR deformable transformer [57], into a bottom-up pipeline that first detects corners and then classifies edge candidates connecting corners. Also based on DETR [57], RoomFormer [53] predicts floorplans from a dense point cloud using a single-stage, end-to-end trainable neural network. Different from previous data-driven approaches [4], RoomFormer encodes the floorplan as a variable-size set of polygons, which are variable-length sequences of ordered vertices. SLIBO-Net [42] focuses on improving RoomFormer's semantic and local geometric quality by incorporating additional MWM priors and post-processing steps. More recently, PolyDiffuse [6] refines polygonal reconstructors from point cloud density maps through a conditional generation procedure. We can consider it an orthogonal work that could be used as a post-processor for our baseline, similar to how it was applied to RoomFormer [53]. We also exploit deformable transformers [57] encoding floor plans with a variable sequence of polygons and vertices [42, 53], but start from pure, sparse imagery input, without needing dense coverage, dense point clouds or MWM-based post-processing. Extreme registration approaches [13, 18, 35], like ours, take as input sparse panoramic images, but exploit them for relative pose estimation with minimal (or no) overlaps. They can offer a complement for the initial camera registration.

3. Overview

Fig. 1 summarizes our reconstruction approach, focusing on the forward prediction pass for a clearer illustration. We refer the reader to the supplementary material for network and training details. We take as input a small set of spatially registered and gravity-aligned equirectangular images. For each input image, the *Nadir shape* network predicts a regularized probability map of the walkable floor area N_p and the floor-ceiling planes distance from the observer, respectively h_f and h_c (Sec. 4). We then exploit the relative pose of the input images to encode all Nadir shapes in a common, fixed-size, floor plan map P_f. The recovered metric scaling is stored separately to reshape the final 3D floorplan. Given this joined representation P_f, named *Nadir floorplan*, an encoder-decoder transformer-based network processes inter-relations, completes self-occluded parts, and refines room shapes as 2D polygons (Sec. 5). Finally, a 3D floor plan is generated by combining ceiling-floor heights, metric aspect, and 2D polygons.

4. Nadir shape generation

The Nadir shape generation is performed, for each input image, by the Nadir shape network (Fig. 1). The network takes as input an equirectangular image I_c of a cluttered room, and outputs in the forward pass a regularized prob-

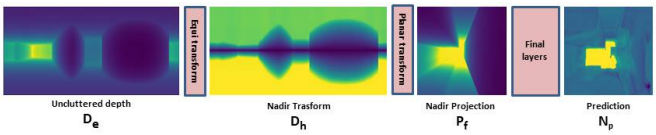


Figure 2. **Nadir shape network last layers**. Starting from the intermediate uncluttered depth D_e, the fully differentiable transform and projection recover the Nadir projection P_f. Finally, the last autoencoding layers return the Nadir Shape N_p.

ability map of the free floor area N_p with the floor-ceiling planes distances h_f and h_c. The same network also outputs the equirectangular depth of the emptied scene D_s as an intermediate result, which is used in loss computation and weight update during the backward pass of training (Fig. 2). Internally, the network predicts an attention mask, segmenting the input image into cluttered or uncluttered zones. A *residual-gated* network extracts features from the input image and its mask. Features are then encoded as a fixed-size sequence of vertically-compressed features, leveraging the indoor structural nature of the desired output, and then processed by a transformer to decode an *uncluttered* depth map of the scene. The depth map is then transformed and projected to recover the Nadir shape.

Attention mask and feature extraction. Starting from the input equirectangular image I_c, we predict an attention mask M_e through a pre-trained, lightweight network [31] (see supplementary material for implementation details), roughly segmenting I_c into cluttered or empty zones to bootstrap the gated-residual encoder. We concatenate the attention mask M_e with the input image I_c and we extract multi-layer features (i.e., four layers with different depths and spatial sizes) through a residual-gated encoder (see supplementary materials for implementation details). In contrast to canonical residual schemes [11], often adopted for generic depth estimation, our approach, for each convolution layer, introduces a dynamic gating approach to process the masked scene efficiently and to complete missing parts, corresponding to structural bounding surfaces of the room occluded by the clutter.

Gravity-aligned-features transformer. To make optimum use of the pixel-wise information of the equirectangular image, our network internally predicts an equirectangular clutter-free depth map D_s, by introducing a specific transformer-based approach. In contrast to a standard vision transformer baseline [8], which would require the conversion of feature maps into a sequence of pixel-wise patches, our approach aims to encode the feature maps into a sequence by accounting for the indoor nature of the scene and its spherical representation. Thus, as in other depth estimation works [29, 46], starting from the assumption that gravity plays an important role in the design and construction of interior environments, we assume that world-space vertical and horizontal features have different characteristics. Such an assumption is even stronger in our case since our goal here is to recover only the structure (i.e., walls, ceiling, and floor), without the need to model clutter details, which typically contain the most free-form data. We perform, thus, an anisotropic, contractive encoding that reduces the vertical direction (i.e., gravity direction) while keeping the horizontal direction unchanged, generating gravity-aligned features (GAF). We apply such compression to each encoded feature map. Finally, compressed features are reshaped to the same size and joined in a flattened GAF feature, $L_s = (l_0 \dots l_s)$, as a sequence of s feature vectors of dimension l (i.e., s horizontal size of the less deep feature map - $s = 256$ and $l = 1024$ for a 512×1024 input). Such a sequence L_s contains both local and non-local geometric features, which are exploited to recover missing depth samples through a multi-head self-attention transformer [48]. Once passed to the transformer, sequence decoding is very fast, through convolutions, upsampling modules, and ELU activations, until we reach the target output resolution.

Nadir shape recovery. To recover the Nadir shape for each camera, we need to efficiently translate the information from the emptied depth map D_e to the Nadir plane. Since recognizing man-made structures is not immediate in spherical space, the network processes 3D data by storing at each pixel of the equirectangular map, rather than the depth D_e, the distance to the horizontal plane passing through the camera location, i.e., the height relative to the eye. This is accomplished by scaling each value in D_e, which corresponds to an azimuth angle θ (along w) and polar angle ϕ (along h), by $\|\sin\phi\|$ to obtain the map D_h. This representation D_h (Fig. 2, D_h), under the vertical walls assumption, better highlights the room structure, although still in equirectangular format. The height value is also exploited to recover additional 3D cues, such as floor and ceiling distances from the camera, to complete the 3D layout of the rooms. To recover the footprint in the Nadir plane, which corresponds to the predicted floor, exploiting the Vertical Walls prior [28], we project the equirectangular depth map D_h to the Nadir plane, ${P_f}^{wp \times wp}$ perpendicular to the Z-axis, so that P_f represents information belonging to the bottom hemisphere of D_h (Fig. 2, P_f). This transformation and projection operation is fully differentiable and integrated into the network. However, because such an intermediate representation P_f directly comes from the predicted emptied depth, much structural information, especially on the edges, may be missing or noisy. Therefore, the final layers of the network consist of an UNet-based (see supplementary material for details) autoencoder, which takes the P_f representation and output a regularized probability map N_p (see Fig. 2 and Fig. 1). Projecting onto the floor, which is assumed to be a single plane, lends itself better to the merging of multiple footprints, also in the case of rooms with differ-

ent ceiling heights. In contrast, projecting onto the ceiling, as done by traditional layout recovery methods working on cluttered input images [27, 52] would necessitate a consistent horizontal plane across the entire multi-room environment, limiting the solution space.

Nadir shape network training. We train the Nadir shape network supervising the training with single room layouts from real-world and synthetic annotated floorplans (Sec. 6.2). We do not need additional annotations for this training task, but we use layouts from the main floorplan reconstructor sets (i.e., adopting the same train/valid/test split - Sec. 5). To directly supervise the Nadir shape prediction and the layout heights, we adopt, respectively, $\mathcal{L}_l$, the binary cross entropy with logits loss for the predicted probability map N_p Sec. 4 and$\mathcal{L}_h$, the $L1$ distance error for the predicted average ceiling-floor distances. Furthermore, starting from the assumption that indoor structure and depth are interrelated [17], we enforce the training by exploiting the intermediate clutter-free depth D_e (see Sec. 4). To recover ground truth data for D_e, we render such depth from the annotated layout. We adopt then several specific losses for such depth. The robust *Adaptive Reverse Huber Loss (BerHu)* $\mathcal{L}_d$; the Structural Similarity Index Measure (SSIM) $\mathcal{L}_{dss}$, which measures the preservation of highly structured signals with strong neighborhood dependencies. Furthermore, to enforce the indoor nature of the output, we include more specific indoor structural losses, respectively the normal consistency loss L_n and the gradient of normals consistency loss L_g, which supervise the smoothness of walls and sharpness of edges. Assuming vertical walls, we restrict the normal map to the horizon [50], computing the cosine similarity to get the normal loss [16] $\mathcal{L}_n = \frac{1}{W}\sum_{i=1}^{W}(-n_i \cdot \overline{n_i})$, where W is the width of the equirectangular image, n_i is the predicted normal, $\overline{n_i}$ is the ground truth normal. To supervise the turning of corners, we compute the loss term as $\mathcal{L}_g = \frac{1}{W}\sum_{i=1}^{W}(g_i - \overline{g_i})$, where as $g_i = arccos(n_{i-1} \cdot n_{i+1})$ is the angle between consecutive normals. As a result, the Nadir shape network is trained by combining indoor depth and layout losses $\mathcal{L}_{sv} = \lambda_d\mathcal{L}_d - \lambda_{dss}\mathcal{L}_{dss} + \lambda_l\mathcal{L}_l + \lambda_h\mathcal{L}_h + \lambda_n\mathcal{L}_n + \lambda_g\mathcal{L}_g$ (see supplementary for lambda values).

5. Nadir floorplan reconstruction

The Nadir shapes recovered for each image are joined and encoded in the Nadir floorplan (Sec. 3). We cast the floorplan reconstruction as a prediction problem of a polygon set, where each polygon represents a room modeled as an ordered sequence of vertices. It should be noted that, differently to pure prediction from an unstructured point cloud [5, 41, 53], here we start from a partial room segmentation, so the main goal is to process inter-relations, complete self-occluded parts, and refine room shapes.

To do this, we scale and fit our Nadir floorplan, which is in metrically-scaled dimensions, into a fixed-size map of dimension ${F_n}^{wf \times wf}$ (Fig. 1), according to common, available annotations [5, 41]. To join and encode the Nadir shapes in this representation, we transform each probability map N_p into a binary mask N_s using a softmax, displace N_s exploiting its camera position, then encode each N_s mask with a progressive integer value (Fig. 1, Floorplan encoding).

The floorplan prediction problem is then cast as $Q \in \mathbb{R}^{M \times N \times 2}$ queries [53] and solved using a deformable transformer [57]. We set M to the maximum number of rooms and N to the maximum number of vertices for each room ($M = 20, N = 40$ in our experiments), and associate each vertex with a binary value, indicating whether it is valid or not. The prediction is implemented through a network (called Nadir Floorplan network in Fig. 1) which includes multi-scale features extraction and flattening, deformable self-attention (DSA) for transformer encoding, and self-attention and deformable cross-attention (SA-DSA) for decoding.

Nadir Floorplan network. Similarly to the Nadir shape encoder (Sec. 4), the first layers extract multi-scale features (four levels) from the Nadir floor map F_n using a plain *ResNet* scheme [11] without gated convolutions, and flatten them to a feature sequence to serve as input to a six layers transformer encoder. Each encoder layer consists of a multi-scale deformable *self-attention* module and a simple feed-forward network [57] (DSA in Fig. 1). Similarly to encoding, decoding is performed by six stacked layers, but, differently from the encoder, each layer consists of a self-attention module, a multi-scale deformable cross-attention module, and a feed-forward network (SA-DSA in Fig. 1). In the decoder, we adopt a combination of content and positional queries [53], so that the decoder performs self-attention on all corner-level queries, regardless of the room they belong to. This design not only allows the interaction between corners of a single room but also enables interaction among corners across different rooms, thus combining information from different polygons to better solve structural occlusions. The decoder output is passed to a shared feed-forward network to predict binary class labels c for each query, indicating its validity as a corner. Finally, the decoder outputs a $M \times N \times 2$ polygons map and a $M \times N \times 1$ classification map, from which a set of valid 2D polygons is extracted [41].

Floorplan network training. The NadirFloor network predicts room logits and corners. Its training is supervised with the same strategy and losses employed by *RoomFormer* [53], with the important difference that we take as input the registered representations inferred for each room by the pre-trained Nadir shape prediction network rather than occurrence maps from point clouds. The transformer decoder outputs M polygons of N vertices (including non-

valid ones, mapped to 0) while the ground truth contains an arbitrary number of polygons with an arbitrary number of vertices. As in *RoomFormer* [53], we adapt Deformable-DETR *Hungarian matcher* [57] to find the optimal match between prediction and ground truth. Note that, since the NadirFloor and NadirShape networks are concatenated and form, together, a fully differentiable end-to-end network mapping a set of panoramic images to a floorplan, it is possible to further improve quality through a fine-tuning pass, where the parameters of the NadirShape network are also made learnable. We plan to explore this path in future work.

6. Results

Our approach is implemented with *PyTorch* and has been tested on a large variety of synthetic and real-world indoor scenes. In the following, we report the costs of inference and training, as well as qualitative and quantitative results on public datasets. Additional details and results are provided in the additional material.

6.1. Training and inference cost

The computational complexity of both neural network modules is very low. The Nadir shape prediction network, including the initial attention mask prediction (Sec. 4), has $31.81M$ parameters with $98.97Gflops$. The average inference time for a 512×1024 equirectangular input image is $67ms$ on an NVIDIA RTX2060 GDD6GB laptop (e.g., building a 20-room Nadir floorplan takes less than 1.5 seconds). While the architecture consists of a single differentiable network, for practical reasons, we train the two modules separately, also saving intermediate results (i.e., individual Nadir shapes and empty depths) for analysis. We trained the Nadir shape network with the Adam optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an adaptive learning rate from 0.0001, on an NVIDIA RTX A5000 (24GB VRAM) with a batch size of 8, with average training time 613 ms/image. The weights are: $\lambda_d = 1.0; \lambda_{ss} = \lambda_l = 0.5; \lambda_h = \lambda_n = \lambda_g = 0.1$. The Nadir floorplan reconstruction module was trained for 500 epochs following the same settings of RoomFormer (RF) [53]. The average inference time for a 256×256 floorplan map is $20ms$ on a NVIDIA RTX2060 GDD6GB laptop.

6.2. Datasets

To train and test our approach, we used public datasets for which ground truth annotated floor plans and registered panoramic images were available. In addition to Structured3D (S3D) [56], adopted by most related state-of-the-art methods and comprising 3,500 synthetic houses with diverse floor plans covering both MWM and non-MWM layouts, we also employed ZinD [7], one of the largest real-world indoor dataset with 3D multi-room layout annotations, containing 71,474 panoramas from 1,524 real homes. To have a consistent comparison, we converted ZInD to S3D's format, following the authors' original instructions [7, 56]. It should be noted that several ZInD annotations and scenes are not parsed and converted correctly to S3D's format, so have been discarded for the experiments.

6.3. Reconstruction performance

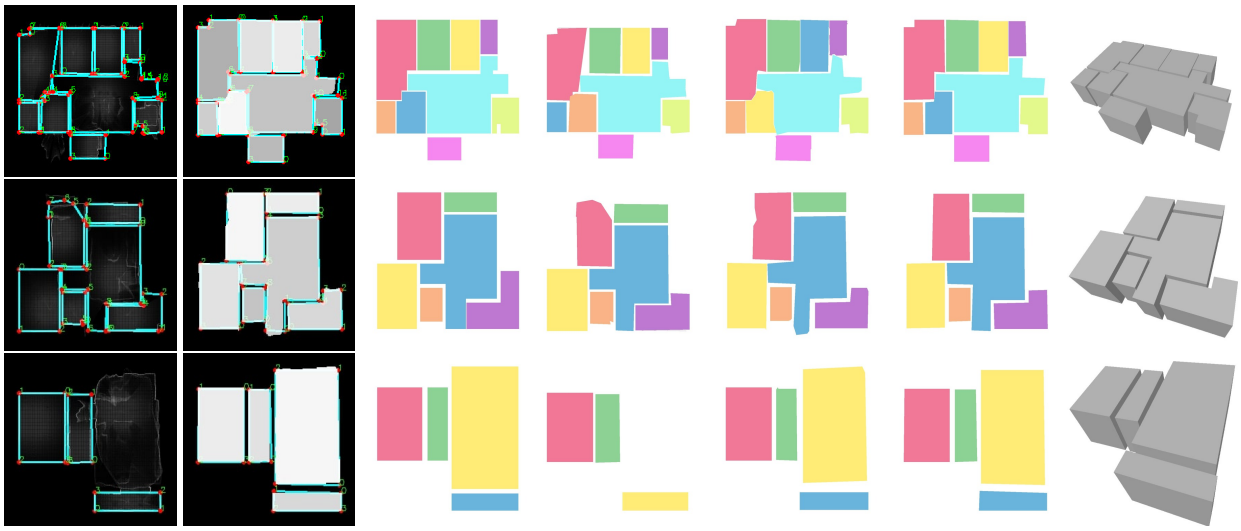
Tab. 1, Fig. 3, and Fig. 4 show quantitative and qualitative results on S3D [56] and ZInD [7] floorplans. As in many floorplan reconstruction works [5, 41, 53], for each ground truth room, we find the best-matching reconstructed room among all predictions in terms of IoU, reporting its Precision, recall, and F1 scores.

Method	Set	Room			Corner			Angle		
		Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
RF	S3D	0.892	0.876	0.884	0.704	0.681	0.692	0.643	0.623	0.633
Our NSO	S3D	0.939	0.930	0.928	0.553	0.542	0.647	0.424	0.522	0.466
Our full	**S3D**	**0.943**	**0.939**	**0.940**	**0.834**	**0.801**	**0.817**	**0.746**	**0.717**	**0.731**
RF	ZInD	0.811	0.756	0.788	0.712	0.615	0.660	0.528	0.444	0.505
Our NSO	ZInD	0.786	0.733	0.758	0.652	0.538	0.627	0.505	0.365	0.424
Our full	**ZInD**	**0.824**	**0.767**	**0.794**	**0.765**	**0.626**	**0.689**	**0.594**	**0.488**	**0.536**

Table 1. **Floorplan performance.** We show our quantitative performance (Our full - in bold) on Structured3D (S3D) [56] and ZInD [7], compared to other representative state-of-the-art works (RoomFormer (RF) [53]) adapted to the same 360° images input. We also include the performance of our method just spatially registering individual Nadir shapes (NS) - Nadir shapes only).

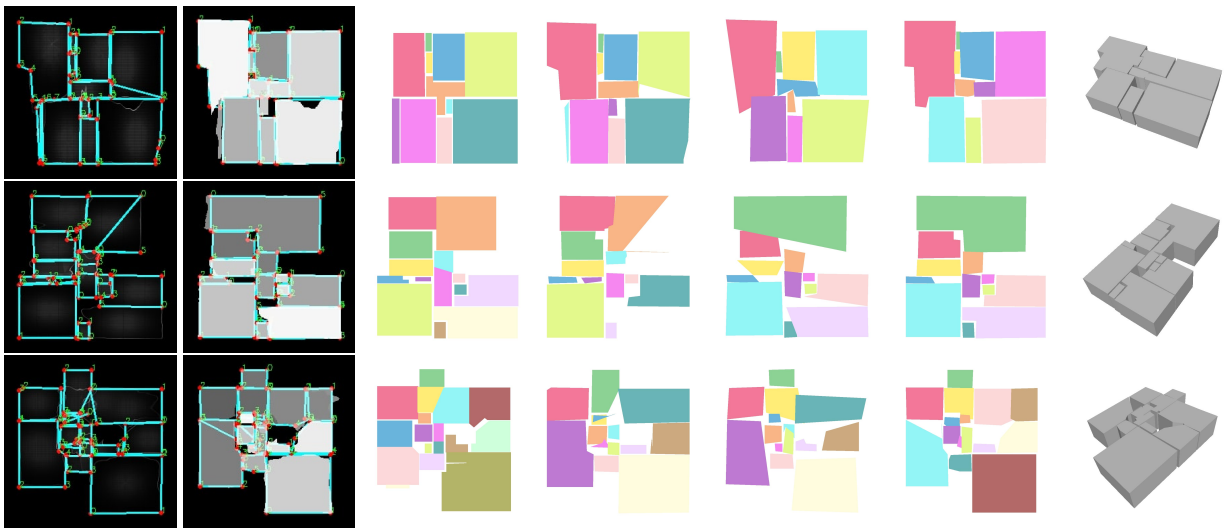
As discussed in Sec. 1 and Sec. 2, current deep-learning floorplan reconstruction methods segment a 2D occupancy map recovered from a dense point cloud, typically generated by dense coverage with regular cameras. Only a few particular methods take panoramic images as direct input, but to find their registration under extreme conditions rather than producing full floor plans [13, 35]. Both benchmarks used here present a challenging situation for reconstructing an articulated multi-room environment from just panoramic data, since only about one image/room is available for S3D, and about 1.5 images/room for ZInD.

To fairly evaluate our method against alternatives, we assume that no 3D data is available, and we predict the occupancy maps required by other methods by predicting the depth of the room using monocular panoramic depth inference ,followed by vertical projection and accumulation proposed in previous works [53]. We experimentally determined (see supplementary material) that when trained with original scenes rather than uncluttered ones, our Nadir shape network has a depth estimation performance in line with other state-of-the-art solutions. We thus selected our depth estimator, trained on the cluttered original input images, and combined it, to form a representative baseline, with RoomFormer (RF) [53], which is a state-of-the-art approach for floor plan segmentation from point cloud input.



(a) RF [53] occ. map | (b) Our Nadir map | (c) GT [56] | (d) RF [53] | (e) Our NSO | (f) **Our full** | (g) **Our 3D**

Figure 3. **Qualitative performances on Structured3D (S3D) [56] dataset**. Each row shows: RoomFormer (RF) [53] point cloud occupancy map (RF occ. map), our Nadir floorplan map (Our Nadir map), ground truth floorplan (GT), RoomFormer prediction (RF), our floorplan prediction just composing the Nadir shapes (our NSO - Nadir shapes only), our full floorplan prediction (Our full), our 3D prediction with predicted metric information (Our 3D).



(a) RF [53] occ. map | (b) Our Nadir map | (c) GT [7] | (d) RF [53] | (e) Our NSO | (f) **Our full** | (g) **Our 3D**

Figure 4. **Qualitative performance on ZInD dataset [7]**. For each row we show: RoomFormer (RF) [53] point cloud occupancy map (RF occ. map), our Nadir floorplan map (Our Nadir map), ground truth floorplan (GT), RoomFormer prediction (RF), our floorplan prediction just composing the Nadir shapes (our NSO - Nadir shapes only), our full floorplan prediction (Our full), our 3D prediction with predicted metric information (Our 3D).

The same solution is adopted as a baseline by recent orthogonal works [6, 42].

We fully retrained RF on both Structured3D (S3D) [56] and ZInD [7] predicted point clouds, obtaining the performance illustrated in Tab. 1, Fig. 3 and Fig. 4. With both datasets, our approach (Tab. 1, *Our full*) achieves the best performance in all metrics. Furthermore, Tab. 1 analyses the performance of a version of our method without the floorplan integration network (Tab. 1, *Our NSO* (i.e., Nadir Shape Only). The floorplan in this case (NSO) is the sepa-

rate polygonization of the contours of each Nadir shape, and the joining of them in the same reference frame thanks to the available mutual registration of each shape, without further optimization. This comparison numerically highlights the interaction between Nadir shapes in the Floorplan network to improve the reconstruction and completion of rooms. This difference is evident qualitatively in Fig. 3 and Fig. 4. In both figures, each row shows: a comparison between a predicted point cloud occupancy map (RF [53] occ. map) and our Nadir floorplan map (Our Nadir map Sec. 4); the ground truth floorplan (GT), the compared method prediction (RF [53]), our floorplan prediction just composing the Nadir shapes (our NSO - Nadir shapes only), our full-pipeline floorplan prediction (Our full), our 3D prediction with predicted metric information (Our 3D) and metric aspect. The proposed scenes highlight how much of the error in the compared solutions comes from the deterioration of structural information in the occupancy map (RF occ. map), as opposed to the Nadir floorplan, where, instead, planarity, sharpness, and other structural features are already visible from single-view processing (our NSO - Nadir shapes only). These differences become more pronounced in the real-world case (Fig. 4), where the contribution of disocclusion and regularization from multi-view integration (Sec. 5) also becomes more evident (Our full).

6.4. Ablation study

We further analyze our design choices in the ablation study in Tab. 2). The first two rows (NS and UNC+NS) summarize the performance of our pipeline without the floorplan integration network (Sec. 5), highlighting the effect of performing the uncluttering of the scene (Sec. 4). In the first case (NS), the Nadir shapes are obtained from cluttered scenes (i.e., intermediate depth D_e includes clutter), while the second case (UNC+NS), adopting the uncluttering layers, corresponds to the *Our NSO* setup of Tab. 1. Similarly, in the third and fourth rows (NS+NF and UNC+NS+NF), we show our full pipeline and the effect of uncluttering on the final result, where the fourth configuration (UNC+NS+NF) is our full method (i.e., our full in Tab. 1).

Method	Room			Corner			Angle		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
NS	0.902	0.897	0.822	0.540	0.526	0.602	0.412	0.516	0.424
UNC+NS (NSO)	0.939	0.930	0.928	0.553	0.542	0.647	0.424	0.522	0.466
NS+NF	0.940	0.936	0.932	0.788	0.768	0.712	0.546	0.688	0.698
UNC+NS+NF	**0.943**	**0.939**	**0.940**	**0.834**	**0.801**	**0.817**	**0.746**	**0.717**	**0.731**

Table 2. **Ablation study.** NS: spatially registering of individual Nadir shapes from cluttered depth maps; UNC+NS: NS from uncluttered depth maps; NS+NF: our full pipeline but without depth uncluttering; UNC+NS+NF: our full pipeline with uncluttered depth shapes.

7. Conclusion and future works

Our novel deep-learning pipeline proves capable of reconstructing multi-room environments from a minimum number of registered images (i.e., about one per room). Our approach combines indoor priors with single panorama analysis to obtain a clutter-free geometric representation of single views and transformer-based resolution of major structural occlusions while fusing multiple views. Although we achieve state-of-the-art results for reconstruction with minimal data, some limitations remain, and several aspects can be extended in the future.

Some failure cases, which arise when some of our priors are not met, are shown in the supplementary material. Such examples include cases, for instance, when structural parts, such as stairs, become dominant in the scene, generating ambiguity in both the uncluttering process and the geometric reconstruction, or when the image includes outdoor parts and the assumption of an indoor environment fully bounded by vertical walls lapses. Another point that we expect to develop more in the future is ceiling modeling since our method can already return a heightmap of the whole floorplan, segmented into rooms. However, the current annotated datasets available for floorplans do not provide an unambiguous representation of the ceiling shapes, forcing other works to impose the AWM [5, 41, 53]. In the future, we expect to exploit and extend the already available annotations (e.g., ZinD [7]) to model 3D floorplans of more complex shapes.

Acknowledgments GP and EG acknowledge the contribution of the Italian National Research Center in High-Performance Computing, Big Data, and Quantum Computing (Next Generation EU PNRR M4C2 Inv 1.4). US and MA acknowledge funding from NPRP-S 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work and are solely the authors' responsibility.

References

[1] Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. HRDFuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proc. CVPR*, pages 13273–13282, 2023. 1, 2

[2] R. Cabral and Y. Furukawa. Piecewise planar and compact floorplan reconstruction from images. In *Proc. CVPR*, pages 628–635, 2014. 3

[3] Wenjie Chang, Yueyi Zhang, and Zhiwei Xiong. Depth estimation from indoor panoramas with neural scene representation. In *Proc. CVPR*, pages 899–908, 2023. 3

[4] Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floor-SP: Inverse cad for floorplans by sequential room-wise shortest path. In *Proc. CVPR*, pages 2661–2670, 2019. 3

[5] Jiacheng Chen, Yiming Qian, and Yasutaka Furukawa. HEAT: Holistic edge attention transformer for structured re-

construction. In *Proc. CVPR*, pages 3866–3875, 2022. 1, 2, 3, 5, 6, 8

[6] Jiacheng Chen, Ruizhi Deng, and Yasutaka Furukawa. Polydiffuse: Polygonal shape reconstruction via guided set diffusion models. *NeurIPS*, 36, 2024. 3, 7

[7] Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. Zillow indoor dataset: Annotated floor plans with 360° panoramas and 3D room layouts. In *Proc. CVPR*, pages 2133–2143, 2021. 1, 2, 6, 7, 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[9] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Reconstructing building interiors from images. In *Proc. ICCV*, pages 80–87. IEEE, 2009. 3

[10] Arnaud Gueze, Matthieu Ospici, Damien Rohmer, and Marie-Paule Cani. Floor plan reconstruction from sparse views: Combining graph neural network with constrained diffusion. In *Proc. CVPR*, pages 1583–1592, 2023. 3

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016. 4, 5

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proc. ICCV*, pages 2961–2969, 2017. 3

[13] Will Hutchcroft, Yuguang Li, Ivaylo Boyadzhiev, Zhiqiang Wan, Haiyan Wang, and Sing Bing Kang. CoVisPose: Covisibility pose transformer for wide-baseline relative pose estimation in 360° indoor panoramas. In *Proc. ECCV*, pages 615–633, 2022. 1, 3, 6

[14] Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. Structured indoor modeling. In *Proc. ICCV*, pages 1323–1331, 2015. 1, 3

[15] San Jiang, Kan You, Yaxin Li, Duojie Weng, and Wu Chen. 3d reconstruction of spherical images: a review of techniques, applications, and prospects. *Geo-spatial Information Science*, 27(6):1959–1988, 2024. 1

[16] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. LGT-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proc. CVPR*, pages 1654–1663, 2022. 1, 2, 5

[17] Lei Jin, Yanyu Xu, Jia Zheng, Junfei Zhang, Rui Tang, Shugong Xu, Jingyi Yu, and Shenghua Gao. Geometric structure based and regularized depth estimation from 360° indoor imagery. In *Proc. CVPR*, pages 889–898, 2020. 2, 5

[18] John Lambert, Yuguang Li, Ivaylo Boyadzhiev, Lambert Wixson, Manjunath Narayana, Will Hutchcroft, James Hays, Frank Dellaert, and Sing Bing Kang. Salve: Semantic alignment verification for floorplan reconstruction from sparse panoramas. In *Proc. ECCV*, pages 647–664. Springer, 2022. 1, 3

[19] David C Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *Proc. CVPR*, pages 2136–2143, 2009. 2

[20] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360° monocular depth estimation via geometry-aware fusion. In *Proc. CVPR*, pages 2801–2810, 2022. 2

[21] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. FloorNet: A unified framework for floorplan reconstruction from 3D scans. In *Proc. ECCV*, pages 201–217, 2018. 3

[22] Aron Monszpart, Nicolas Mellado, Gabriel J Brostow, and Niloy J Mitra. Rapter: rebuilding man-made scenes with regular arrangements of planes. *ACM Trans. Graph.*, 34(4): 103–1, 2015. 3

[23] Nelson Nauata and Yasutaka Furukawa. Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference. In *Proc. ECCV*, pages 711–726. Springer, 2020. 3

[24] Gregoire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P. Breckon. Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360° panoramic imagery. In *Proc. ECCV*, pages 812–830, 2018. 2

[25] Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. Mobile mapping and visualization of indoor structures to simplify scene understanding and location awareness. In *Proc. ECCV Workshops*, pages 130–145. Springer, 2016. 1

[26] Giovanni Pintore, Fabio Ganovelli, Alberto Jaspe Villanueva, and Enrico Gobbetti. Automatic modeling of cluttered multi-room floor plans from panoramic images. *Computers Graphics Forum*, 38(7):347–358, 2019. 3

[27] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. AtlantaNet: Inferring the 3D indoor layout from a single 360° image beyond the Manhattan World assumption. In *Proc. ECCV*, pages 432–448, 2020. 2, 5

[28] 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. *Comput. Graph. Forum*, 39(2):667–699, 2020. 1, 2, 4

[29] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. CVPR*, pages 11536–11545, 2021. 2, 4

[30] Giovanni Pintore, Eva Almansa, Marco Agus, and Enrico Gobbetti. Deep3DLayout: 3D reconstruction of an indoor layout from a spherical panoramic image. *ACM Trans. Graph.*, 40(6):250:1–250:12, 2021. 2

[31] Giovanni Pintore, Marco Agus, Eva Almansa, and Enrico Gobbetti. Instant automatic emptying of panoramic indoor scenes. *IEEE TVCG*, 28(11):3629–3639, 2022. Proc. ISMAR. 4

[32] Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. Deep scene synthesis of atlanta-world interiors from a single omnidirectional image. *IEEE TVCG*, 29, 2023. 2

[33] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. Automatic 3D modeling and exploration of indoor structures from

panoramic imagery. In *SIGGRAPH Asia 2024 Courses (SA Courses '24)*. ACM Press, 2024. 1, 2

[34] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360MonoDepth: High-resolution 360° monocular depth estimation. In *Proc. CVPR*, pages 3762–3772, 2022. 2

[35] Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, and Yasutaka Furukawa. Extreme structure from motion for indoor panoramas without visual overlaps. In *Proc. ICCV*, pages 5683–5691, 2021. 1, 3, 6

[36] Mohammad Amin Shabani, Sepidehsadat Hosseini, and Yasutaka Furukawa. HouseDiffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising. *arXiv preprint arXiv:2211.13287*, 2022. 3

[37] Zhijie Shen, Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Shuai Zheng, and Yao Zhao. Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In *Proc. CVPR*, pages 17337–17345, 2023. 2

[38] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. ECCV*, pages 746–760. Springer, 2012. 3

[39] Bolivar Solarte, Yueh-Cheng Liu, Chin-Hsuan Wu, Yi-Hsuan Tsai, and Min Sun. 360-dfpe: Leveraging monocular 360-layouts for direct floor plan estimation. *IEEE Robotics and Automation Letters*, 7(3):6503–6510, 2022. 3

[40] Bolivar Solarte, Chin-Hsuan Wu, Yueh-Cheng Liu, Yi-Hsuan Tsai, and Min Sun. 360-MLC: Multi-view layout consistency for self-training and hyper-parameter tuning. *NeurIPS*, 35:6133–6146, 2022. 3

[41] Sinisa Stekovic, Mahdi Rad, Friedrich Fraundorfer, and Vincent Lepetit. Montefloor: Extending mcts for reconstructing accurate large-scale floor plans. In *Proc. CVPR*, pages 16034–16043, 2021. 1, 2, 3, 5, 6, 8

[42] Jheng-Wei Su, Kuei-Yu Tung, Chi-Han Peng, Peter Wonka, and Hung-Kuo Chu. SLIBO-Net: Floorplan reconstruction via slicing box representation with local geometry regularization. In *Proc. NIPS*, 2023. 3, 7

[43] Y. Su and K. Grauman. Kernel transformer networks for compact spherical convolution. In *Proc. CVPR*, pages 9434–9443, 2019. 2

[44] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360° imagery. In *NeurIPS*, pages 529–539, 2017. 2

[45] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In *Proc. CVPR*, pages 1047–1056, 2019. 1, 2

[46] Cheng Sun, Min Sun, and Hwann-Tzong Chen. HoHoNet: 360° indoor holistic understanding with latent horizontal features. In *Proc. CVPR*, pages 2573–2582, 2021. 1, 4

[47] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proc. ECCV*, pages 732–750, 2018. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 4

[49] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. BiFuse: Monocular 360° depth estimation via bi-projection fusion. In *Proc. CVPR*, pages 462–471, 2020. 2

[50] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. LED2-Net: Monocular 360 layout estimation via differentiable depth rendering. In *Proc. CVPR*, pages 12956–12965, 2021. 2, 5

[51] Haiyan Wang, Will Hutchcroft, Yuguang Li, Zhiqiang Wan, Ivaylo Boyadzhiev, Yingli Tian, and Sing Bing Kang. PSMNet: Position-aware stereo merging network for room layout estimation. In *Proc. CVPR*, pages 8616–8625, 2022. 3

[52] 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 *Proc. CVPR*, 2019. 2, 5

[53] Yuanwen Yue, Theodora Kontogianni, Konrad Schindler, and Francis Engelmann. Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries. In *Proc. CVPR*, 2023. 1, 2, 3, 5, 6, 7, 8

[54] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. PanoContext: A whole-room 3D context model for panoramic scene understanding. In *Proc. ECCV*, pages 668–686, 2014. 2

[55] Yining Zhao, Chao Wen, Zhou Xue, and Yue Gao. 3D room layout estimation from a cubemap of panorama image via deep Manhattan Hough transform. In *Proc. ECCV*, pages 637–654. Springer, 2022. 2

[56] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *Proc. ECCV*, pages 519–535, 2020. 1, 2, 6, 7

[57] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3, 5, 6

[58] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. OmniDepth: Dense depth estimation for indoors spherical panoramas. In *Proc. ECCV*, pages 453–471, 2018. 2

[59] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. LayoutNet: Reconstructing the 3D room layout from a single RGB image. In *Proc. CVPR*, pages 2051–2059, 2018. 2

[60] Chuhang Zou, Jheng Wei Su, Chi Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung Kuo Chu, and Derek Hoiem. Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods. *International Journal of Computer Vision*, 129:1410–1431, 2021. 1, 2