NadirFloorNet: reconstructing multi-room floorplans from a minimal set of registered panoramic images
Supplementary material

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

NadirFloorNet *a novel deep-learning approach for predicting complex indoor floor plans with ceiling heights from a minimal set of registered gravity-aligned 360° images of cluttered rooms with vertical walls. This document complements the main paper by providing additional material. In particular, we include details on the network architecture and its implementation, additional qualitative and quantitative results, details on the performance of the component that infers depth of individual rooms, and an analysis of failure cases.*

1. Network architecture and implementation details

Our *NadirFloorNet* pipeline, whose structure is summarized in Fig. 1, 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 step.

The pipeline includes a network module, called *Nadir Shape Network*, for processing individual images (Fig. 2). The network infers from a single equirectangular image its Nadir shape, i.e., the estimation of the free floor of the imaged room in the local reference frame of the camera. The training (Fig. 2) is performed by combining and extending depth and layout losses, simultaneously supervising the uncluttered depth (i.e., layout depth) and the room footprint in the horizontal plane (i.e., Nadir shape). At the beginning of the network, an attention mask is estimated with a very lightweight ($4M$ parameters for this paper) autoencoder based on a U-Net architecture (purple network in Fig. 2). Using a binary cross-entropy loss, we pre-train such attention mask network with the Structured3D [5] dataset, which contains the registered representation of empty and non-empty rooms. The Nadir shape network is trained by combining indoor depth and layout losses (see main paper) with $\lambda_d = 1.0$, $\lambda_{dss} = -0.5$, $\lambda_l = 1.0$, $\lambda_h = 0.1$, $\lambda_n = 0.1$, $\lambda_g = 0.1$.

The Floorplan processor, which processes the Nadir floorplan image, is implemented by two elements: the *NadirFloor network* and the *2D/3D floorplan builder* (Fig. 1). The NadirFloor network predicts rooms' logits and corners. Its training is supervised with the same strategy employed by *RoomFormer* [4], and we refer the reader to the original paper for details. It should be noted, however, that while the strategy remains the same, we take as input the registered representations inferred for each room by the pre-trained Nadir shape prediction network module rather than occurrence maps from point clouds.

The 2D/3D floorplan builder finally converts the network output into closed 2D polygons, eliminating redundant corners to have consistent polygons ready for mesh creation. Then, exploiting the heights predicted by the NadirShape network (Fig. 2) and the metric information also provided by it (i.e., metric scaling and original aspect), the polygons are transformed into watertight 3D meshes, representing each room.

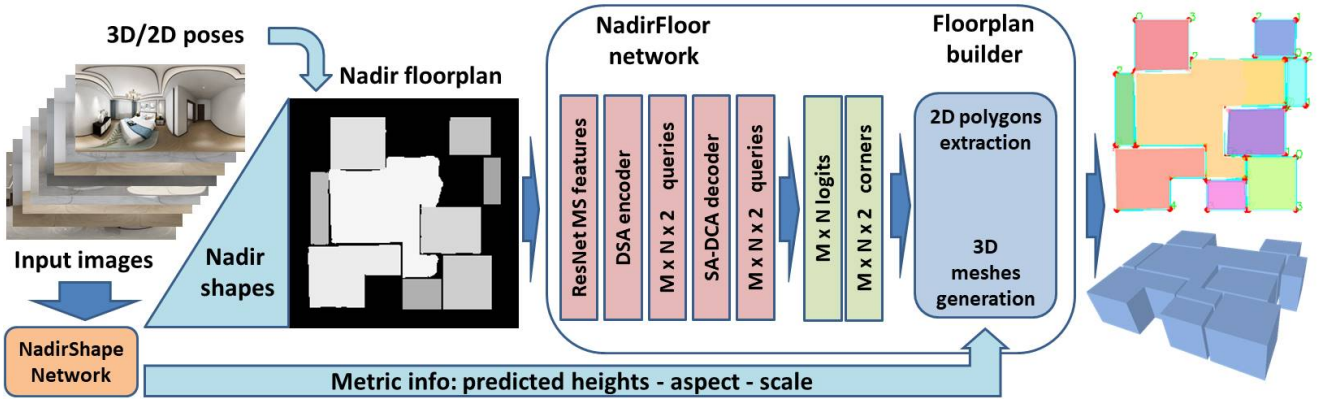


Figure 1. Final floorplan generation pipeline and Floorplan module implementation (i.e., NadirFloor network and 2D/3D floorplan builder).

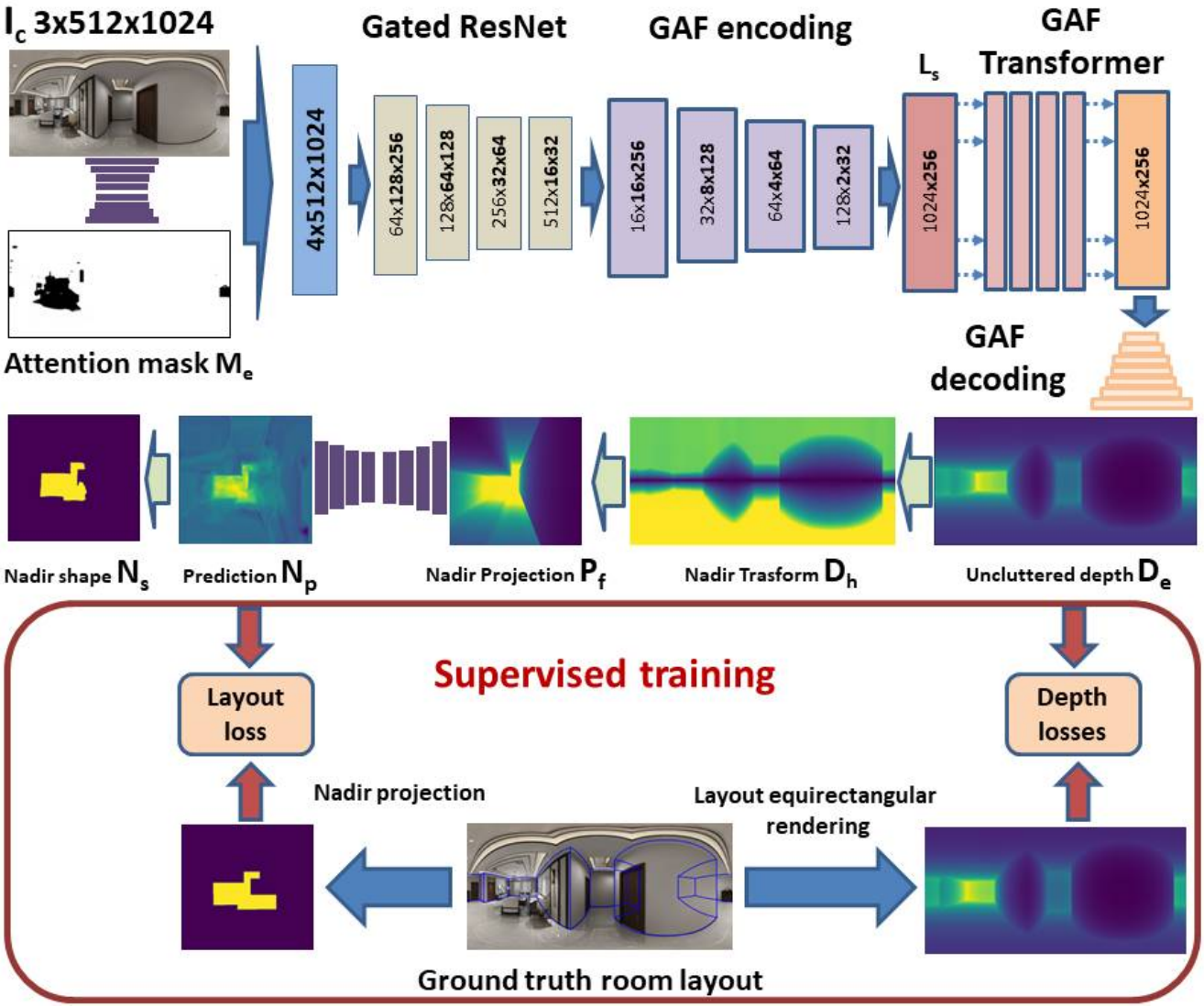


Figure 2. Nadir shape network pipeline and supervised training overview.

2. Additional qualitative Results

We present additional qualitative results of our method on the benchmarks adopted in the paper. Figure 3 illustrates results on several interesting synthetic multi-room layouts from the Structured3D dataset [5], while Figure 4 provides results on selected real-world layouts contained in the ZInD dataset [1]. For each image, we show the predicted Nadir map, the final reconstructed polygons (Predicted FP), the ground truth floorplan (GT FP), and a 3D view of our prediction, once rescaled to metric aspect (2D/3D floorplan builder).

In both the real-world and synthetic cases, our method reconstructs floorplans with many rooms and of non-trivial complexity, even in the presence of wall shapes that are not aligned with canonical directions. A recurring issue in real-world cases, which adds difficulty to the reconstruction, is the inherent ambiguity of the original annotations. A prominent example is in the third row of Figure 4. Such annotations are derived from manual approximations and often do not tightly correspond to the real shape of the structure. Even in these cases, our method still manages to resolve the ambiguity.

3. Nadir shape network depth inference performance

The Nadir shape generation is performed, for each input image, by the Nadir shape network (Figure 1). The network takes as input an equirectangular image I_c of a cluttered room, and outputs in the forward pass a regularized probability map of the free floor area N_p with the floor-ceiling planes distances h_f and h_c (Figure 2). The same network also outputs, as an intermediate result used for training, the equirectangular depth of the emptied scene. This depth is used in loss computation and weight update during the backward pass of training (Figure 2). In Tab. 1 investigate the performance of our network by benchmarking such an intermediate depth map.

Method	**MAE**	**RMSE**	δ_1
SliceNet [2]	0.402	0.194	0.932
PanoFormer [3]	0.064	0.156	0.970
Our NS cluttered	0.060	0.043	0.964
Our NS uncluttered	**0.034**	**0.027**	**0.981**

Table 1. **Nadir shape - pure depth estimation performance.** As an orthogonal experiment, we show our quantitative performance (in bold) compared to other representative state-of-the-art works in terms of depth estimation. We adopt mean absolute error (MAE), root mean square error of linear measures (RMSE) and relative accuracy δ_1, defined as the fraction of pixels where the relative error is within a threshold of 1.25, with training and testing on Structured3D [5].

In particular, we show the single view depth performance of the Nadir shape network as the *Our NS uncluttered* case in Tab. 1, comparing to baselines for which training support and results on Structured3D [5] were available, such as SliceNet [2] and PanoFormer [3]. For clarity, the *Our NS uncluttered* results correspond to the canonical configuration of our network, i.e., related to the prediction of an empty panoramic scene. This explains the significant margin in performance compared to the other methods that predict complete clutter scenes instead. To provide a fairer comparison just in terms of depth estimation, we retrained our Nadir shape network to predict a full, cluttered, scene. The comparison using the output of this retrained network, which is the same expected by the compared methods [2, 3], is reported in Tab. 1 as *Our NS cluttered*. Even in this case, our approach is in line with the state of the art. This shows the benefit of having a training pass additionally supervised by the annotated layout.

Given the good performance of this solution, as illustrated in Tab. 1, we used this configuration (i.e., Our NS cluttered) to estimate full depth maps from the input images to generate the data required by other methods that generate floor plans for panoramic images [4]. This is done by applying the *Our NS cluttered* to predict the depth of the room using monoscopic panoramic depth inference followed by vertical projection and accumulation to compute the occupancy maps. See results sections of the main paper for benchmark discussion.

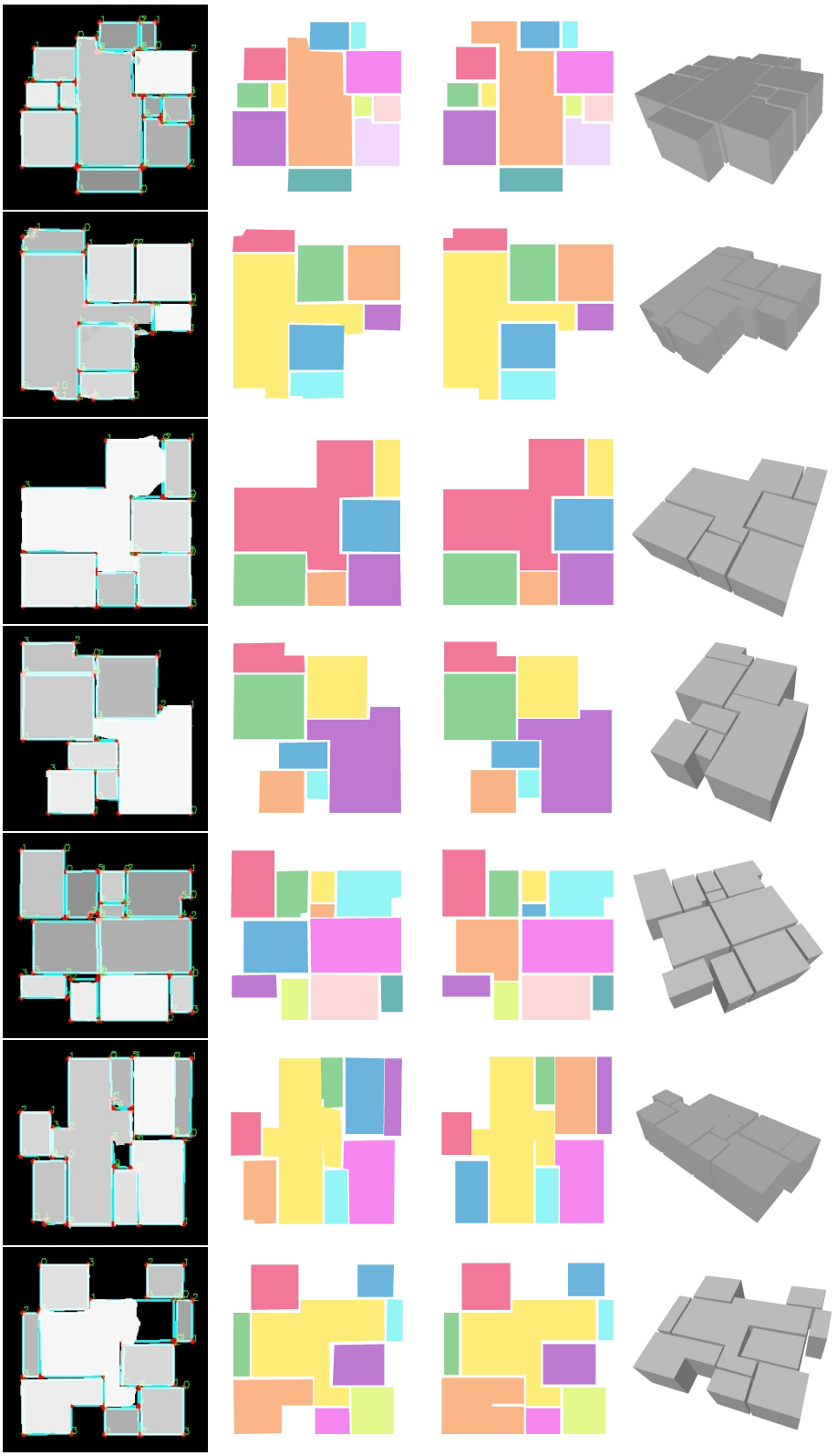
4. Failure cases

Our method exploits indoor-specific priors to permit the reconstruction of plausible structures when minimal information is available. As for all environment- or object-specific methods, this capability also leads to failure cases when the imaged model does not meet our prior assumptions.

Figure 5 illustrates some clear failure cases that may occur when using our method.

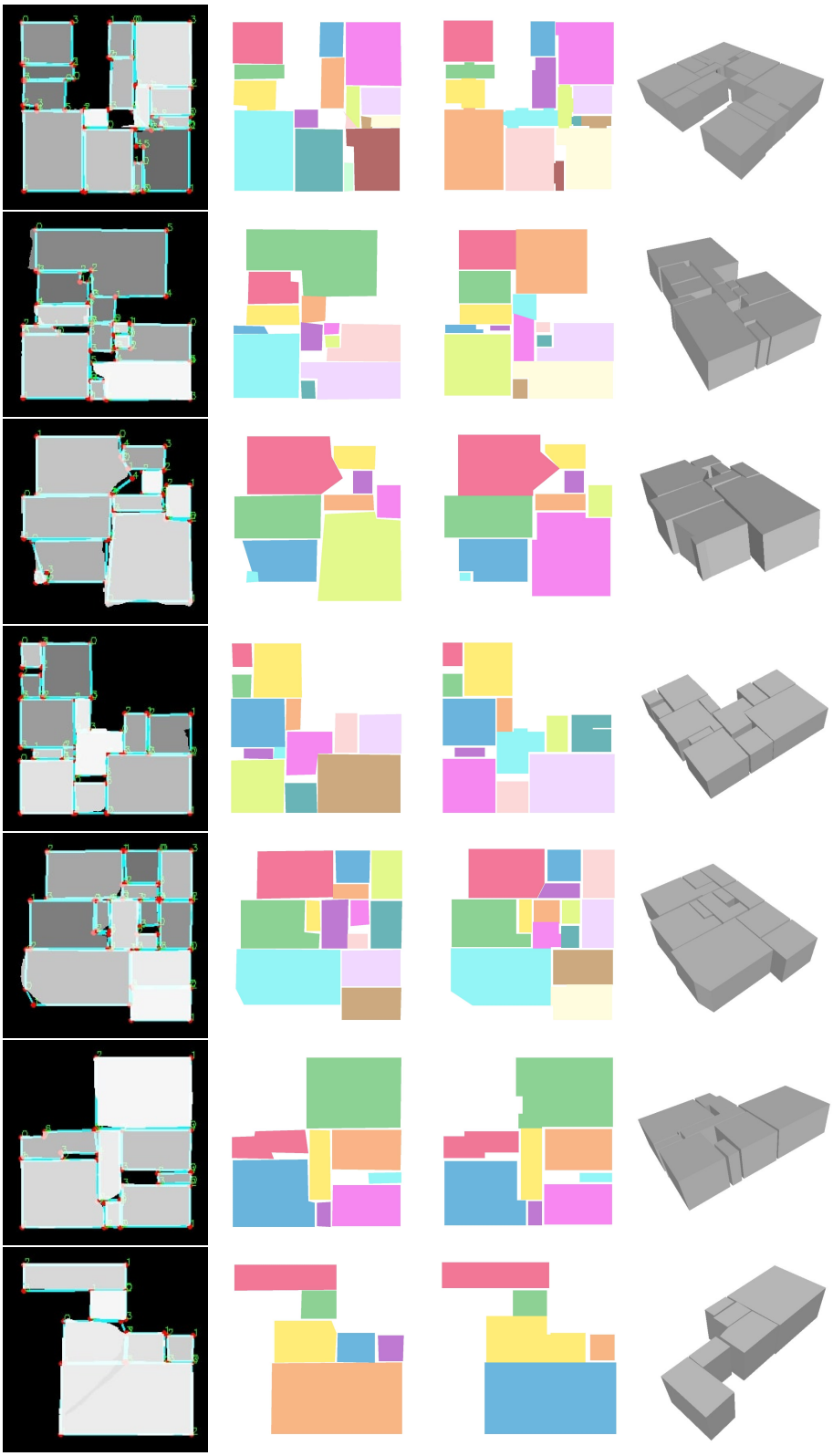
The first row of Figure 5, created from examples present in the ZInD dataset [1], shows the error that arises when some structural parts, such as stairs, become dominant in the scene. Such structures generate ambiguity in both the uncluttering process and the geometric reconstruction. The second example in the same row illustrates, instead, the case of an environment with partially outdoor parts, where the assumption of an indoor environment fully bounded by vertical walls is not met.

The second row of Figure 5, created from examples present in the Structured3D dataset [5], shows instead another specific failure case. In this scene, formally compatible with a standard indoor, the method fails to separate the clutter from the structure, particularly because of the type of shelves present, that partially let see the structure through them. Also to be noted is the error, although minor, in the



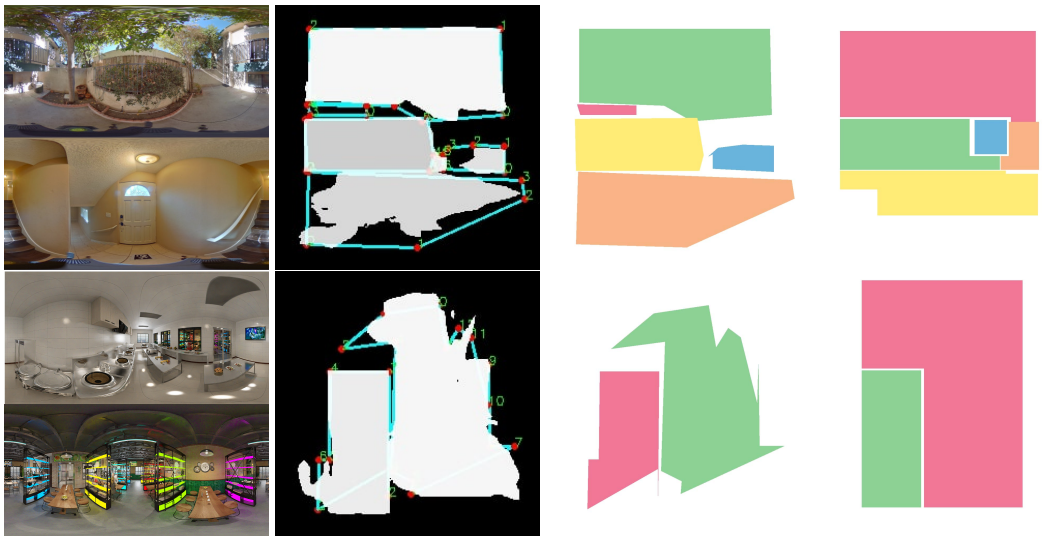
(a) Nadir map (b) Predicted FP (c) GT FP (d) Predicted 3D

Figure 3. **Qualitative examples**. We present additional scenes from the Structured3D [5] dataset reconstructed by our method.



(a) Nadir map (b) Predicted FP (c) GT FP (d) Predicted 3D

Figure 4. **Qualitative examples**. We present additional scenes from the ZInD [1] dataset reconstructed by our method.



(a) Input samples (b) Nadir map (c) Prediction (d) GT

Figure 5. **Failure case examples**. We present some examples where our method fails to reconstruct the floorplan. In the first row, we show a scene from the ZInD [1] dataset, and in the second row a scene from the Structured3D [5] dataset.

other room, mainly due to the presence of a textured outdoor visible through the window and reflective surfaces that confound the reconstruction.

References

[1] 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. 3, 5, 6

[2] 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. 3

[3] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. PanoFormer: Panorama transformer for indoor 360° depth estimation. In *Proc. ECCV*, pages 195–211. Springer, 2022. 3

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

[5] 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, 3, 4, 6