# Floorplan Generation from Noisy Point Cloud

Anselmo Talotta
INTech, Amazon
talotta@amazon.lu

Valentin Radu
Ring Robotics, Amazon
valrad@amazon.com

Lorenzo Sorgi
Ring Robotics, Amazon
sorgil@amazon.com

## ABSTRACT

Floorplans are useful for navigating indoor spaces, for resource allocation and for indoor space management among many others. But in the absence of readily available digital floorplans, these are hard to generate. In this work, we enhance the generation of floorplans from point clouds to be more robust to noisy measurements of sensor data. In our approach, we train an object detector to expose room shapes in a density map produced from a 3D point cloud, as well as the position of relevant landmarks, such as doors and windows. We improve the robustness of the room detector by training in two stages, firstly using point clouds extracted from synthetic 3D graphical representations of plausible indoor spaces; and secondly, extending the training of the model in a new domain by using real-world data collected with Tango devices. This two-tier training nudges the model closer to our target domain, of generating floorplans from easily collected point cloud scans in the real-world. Finally, we showcase the capability of our solution when operating with noisy Lidar scans collected from a drone with pose estimation.

## CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; Robotics.

## KEYWORDS

floor plan generation, neural networks, density map, room detection, indoor spaces

**ACM Reference Format:**
Anselmo Talotta, Valentin Radu, and Lorenzo Sorgi. 2023. Floorplan Generation from Noisy Point Cloud. In *2nd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications (GeoIndustry '23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3615888.3627810

## 1 INTRODUCTION

Floorplan generation, or the process of converting 3D sensor data into a 2D layout, has gained traction in the fields of computer vision and machine learning. This 2D representation delineates room boundaries and other architectural elements. Typically, the source is 3D data in the form of point clouds generated from depth images of LiDAR (Light Detection and Ranging) and depth cameras.

The utility of floorplan generation is evident across various sectors. The real estate, constructions and architecture industries rely heavily on precise floorplans for diverse activities, from marketing campaigns to renovation projects and minute building management. Besides these applications, in robotics floorplans are indispensable for localization, navigation and route planning. Equipped with fine-detail layouts of indoor spaces, robots can navigate and take decisions more efficiently for smooth operation in those environments.

However, the journey from 3D data points to 2D layouts is not straightforward. Some important challenges need to be overcome:

- **Noise in Data**: Devices like LiDAR or depth cameras might generate noisy data, potentially skewing the actual layout or introducing unwanted artifacts.
- **Incomplete Data**: Some indoor spaces may have partial or incomplete scans, resulting in data gaps. Bridging these gaps without compromising on accuracy remains a challenge.
- **Complex Architectures**: Unique building designs or unconventional architectural features can hinder the floorplan reconstruction.
- **Accurate Scale**: The pose and motion of sensing devices may introduce translations in data points. As effect of this, the proportion and scale of walls is affected.
- **Semantic Understanding**: Scans contain more information that just the walls of rooms. Recognising doors and windows from minuscule local deviations in data samples brings tremendous value to many applications.

Recently, the authors of RoomFormer [12] attempted to tackle these challenges. RoomFormer leverages Transformers to reconstruct room polygons from 2D projections. Creatively, they formulate the task of floorplan generation as a variant of object detection. Based on this formulation, a standard process is to convert data representations into the COCO [6] style format for easier used of the model. While RoomFormer demonstrates good performance on the synthetic data in Structured3D [13], its performance on real-world data is drastically reduced. To overcome this limitation, we revisit the training of RoomFormer.

Structured3D synthetic data is generated for 3500 virtual apartments. The cost to create a similar size dataset of real indoor scans is prohibitive. FloorNet [7] is a rare effort of public dataset, although it captures data from only 155 single floor homes. We adopt this dataset and revive its associated tools to produce a data format similar to the well established Structured3D data format.

Our research shows that training with synthetic data is useful, but not sufficient to achieve adequate performance with RoomFormer on real indoor scans. We fine-tune the model with data adapted from FloorNet. We find that this improves the performance on real data. Even more, our training makes the model robust to domain adaptation, observing good performance on data produced by depth measurement from a Lidar, despite the training being done


*GeoIndustry '23, November 13, 2023, Hamburg, Germany*

ACM ISBN 979-8-4007-0350-8/23/11.
https://doi.org/10.1145/3615888.3627810

only on data from a device that uses optical flow to produce the depth information.

Our main contributions are as follows:

(1) Extend the scope of the FloorNet dataset by formatting it to the more widely adopted COCO style.
(2) Improve the robustness of RoomFormer by fine-tuning it on real data from FloorNet, on top of the initial training on the synthetic data from Structured3D.
(3) Demonstrate that our training approach over two domains (synthetic data for pre-training and real data for fine-tuning) makes the model relevant for other domain transfers, in particular observing good performance on Lidar depth measurements despite not having such domain data during training.

## 2 RELATED WORK

The task of generating 2D floorplans from 3D point clouds has been an evolving research area in recent years. The advancements in 3D scanning and data processing technologies have accelerated this line of research. Here, we review some of the more significant works in this space, emphasizing two primary dimensions: the distinction between traditional computer vision (CV) and deep learning approaches, and the transition from 3D sources to 2D density maps as intermediate representations that encode spatial information about the 3D environment.

### 2.1 Traditional CV Approaches

*2.1.1 Direct Use of 3D Data.* [1] explored the potential of laser scanners in capturing intricate indoor details, representing one of the pioneering efforts in creating digital representations of interior environments. [10] introduced a system where mobile robots, equipped with LiDAR sensors, were deployed to produce 2D floorplans from point cloud data, emphasizing the identification of walkable areas to deduce wall positions.

*2.1.2 2D Density Maps Derived from 3D Sources.* [8] adopted a strategy of transforming 3D data into a 2D density histogram via voxelization. This histogram serves as an intermediate representation that encodes spatial information about the 3D environment, making it easier to delineate architectural elements. Similarly, [5] introduced the concept of mobile crowdsensing for indoor floorplan reconstruction, emphasizing the use of 2D representations derived from 3D data.

### 2.2 Deep Learning Approaches

Deep learning has brought forth innovative methodologies in floorplan reconstruction, often leveraging the Structure3D [13] dataset. Structure3D is a comprehensive synthetic dataset that provides a variety of indoor scenes, facilitating the training of deep learning models for tasks like floorplan reconstruction. While synthetic datasets offer a large volume of data, real datasets, such as the one introduced with FloorNet [7], capture the intricacies and nuances of real-world environments, making them invaluable despite their smaller size.

*2.2.1 Direct Use of 3D Data.* **FloorNet** [7], alongside the above mentioned dataset, presented a unified framework for reconstructing floorplans from 3D scans, abstracting 3D complexities into 2D representations. **MonteFloor** [9] extended the Monte Carlo Tree Search algorithm for large-scale floor plan reconstruction, directly utilizing 3D point cloud data.

**Table 1: The main characteristics of the two chosen datasets.**

| Dataset | Type of source | Number of scenes | COCO style |
|---|---|---|---|
| Structure3D | Synthetic | 3500 | Yes |
| FloorNet | Real | 155 | No |

*2.2.2 2D Density Maps Derived from 3D Sources.* The use of 2D density maps as intermediate representations in deep learning approaches has gained traction due to several inherent advantages. Firstly, 2D representations reduce the complexity of the data, allowing models to focus on essential spatial relationships without the overhead of 3D data processing. This simplification often leads to faster training times and requires less computational resources. Secondly, 2D density maps provide a more intuitive space for architectural elements, making it easier to identify and delineate structures like rooms, corridors, and doors. This spatial clarity often results in more accurate and coherent floorplan reconstructions.

**Floor-SP** [4] introduced an inverse CAD technique, formulating the reconstruction as a sequential room-wise shortest path problem, and leveraged 2D representations for better structure and accuracy. **RoomFormer** [12], implemented a novel approach to 2D floorplan reconstruction from 3D scans using a single-stage structured prediction task. Instead of traditional multi-stage pipelines, the model employs a Transformer architecture to generate polygons of multiple rooms in parallel, eliminating the need for hand-crafted intermediate stages. This holistic approach can directly map a density image to a set of room polygons, leveraging the sequence prediction capabilities of Transformers.

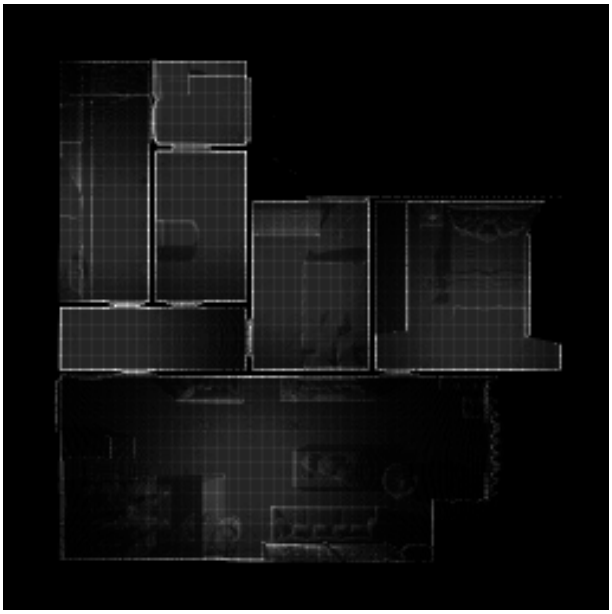
Given these advantages, our research has chosen to adopt the deep learning approach on 2D density maps, anticipating that this methodology will yield superior results in terms of accuracy, efficiency, and architectural coherence. In particular, inspired by its elegant formulation and performances, we elected RoomFormer as our baseline method and conducted our experiments using this model architecture.

## 3 DATASET

This section presents the data processing and dataset correction we performed for the evaluation of our proposed solution.

Based on our preliminary research, we identified two datasets of 3D interior environments containing floorplan annotations: Structure3D [13] and FloorNet [7]. Structure3D contains a considerably bigger corpus of floorplans, with 3500 renderings of synthetic apartments. Floorplan contains only 155 floorplans but they are acquisitions of real apartments from RGB-D video scans (See table 1 for reference) operated by Tango devices [11].

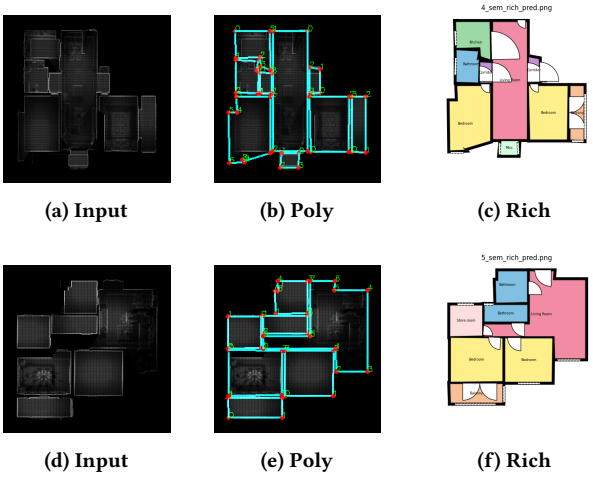
FloorNet has been utilized only by the homonym model that has been released by the same authors and had no further wide adoption in the community. The reasons why this dataset has such a low adoption may be due to its reduced size of floorplans, which made it unsuitable for training deep neural network models. Another reason for its reduced adoption might be the complexity of data



(a) Structure3D sample

(b) FloorNet sample

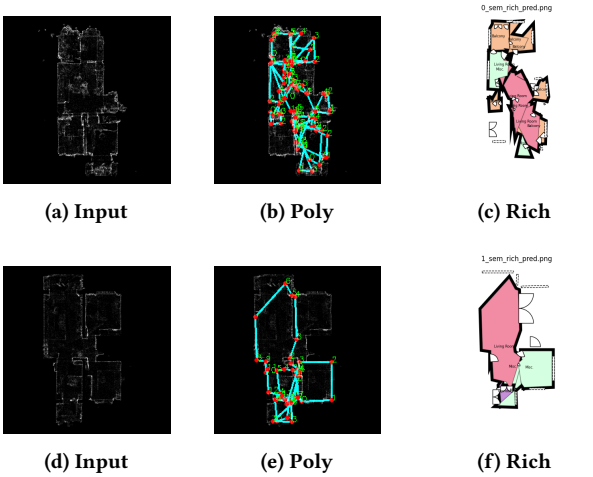
**Figure 1: Density maps produced from the 3D data samples from the two datasets.**




(a) Input (b) Poly (c) Rich

(d) Input (e) Poly (f) Rich

**Figure 2: Predictions generated by the original RoomFormer using samples from Structure3D.**




(a) Input (b) Poly (c) Rich

(d) Input (e) Poly (f) Rich

**Figure 3: Predictions generated by the original RoomFormer using samples from FloorNet.**

annotation format, which is made of a mix of vector information and rasterized semantic segments.

Structure3D, with its expansive corpus of data and rich, well-formatted annotations, has established itself as the standard dataset in literature for the task of Floorplan Generation. Since its release, it has been employed by numerous models, as summarized in Table 2. Notably, in the most recent works, the annotations of this dataset have been adapted to the COCO format for Object Detection [6], albeit with minor customizations. This adaptation to the COCO format, which is particularly popular for the Object Detection task, has enhanced the dataset's compatibility with prevalent libraries and toolkits, further cementing its significance in the research community.

Despite the different adoption state of the two datasets, we believe that data collected from real world is vital to train successfully a machine learning model with desired performances on real data. We verified this assumption empirically, comparing predictions of RoomFormer pretrained with Structure3D on both Structure3D and FloorNet scenes, and we verified that high performances on synthetic data don't guarantee acceptable performances on real data (see Figure 2 and Figure 3). We then elected FloorNet as our dataset of preference and proceeded with an operation of "refurbishment"

**Table 2: Benchmarking different floorplan generation methods on the Structure3D dataset.**

| Method | Stages | Steps | Prec. (Room) | Rec. (Room) | F1 (Room) | Prec. (Corner) | Rec. (Corner) | F1 (Corner) | Prec. (Angle) | Rec. (Angle) | F1 (Angle) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Floor-SP [5] | 2 | - | 89 | 88 | 88 | 81 | 73 | 76 | 80 | 72 | 75 |
| MonteFloor [42] | 2 | 500 | 95.6 | 94.4 | 95 | 88.5 | 77.2 | 82.5 | 86.3 | 75.4 | 80.5 |
| LETR [45] | 1 | 1 | 94.5 | 90 | 92.2 | 79.7 | 78.2 | 78.9 | 72.5 | 71.3 | 71.9 |
| HEAT [7] | 2 | 3 | 96.9 | 94 | 95.4 | 81.7 | 83.2 | 82.5 | 77.6 | 79 | 78.3 |
| RoomFormer [46] | 1 | 1 | 97.9 | 96.7 | 97.3 | 89.1 | 85.3 | 87.2 | 83 | 79.5 | 81.2 |
| RoomFormer (modified) | 1 | 1 | 96.3 | 96.2 | 96.2 | 89.7 | 86.7 | 88.2 | 85.4 | 82.5 | 83.9 |
| RoomFormer (modified) + PolyDiffuse | 2 | 10 | 98.7 | 98.1 | 98.4 | 92.8 | 89.3 | 91 | 90.8 | 87.4 | 89.1 |
| Rough annotations | 1 | - | 17.9 | 18.2 | 18 | 1.3 | 1.4 | 1.3 | 0.1 | 0.1 | 0.1 |
| Rough annotations + PolyDiffuse | 2 | 10 | 97.4 | 98.2 | 97.8 | 91.7 | 92.2 | 91.9 | 89.2 | 89.7 | 89.4 |

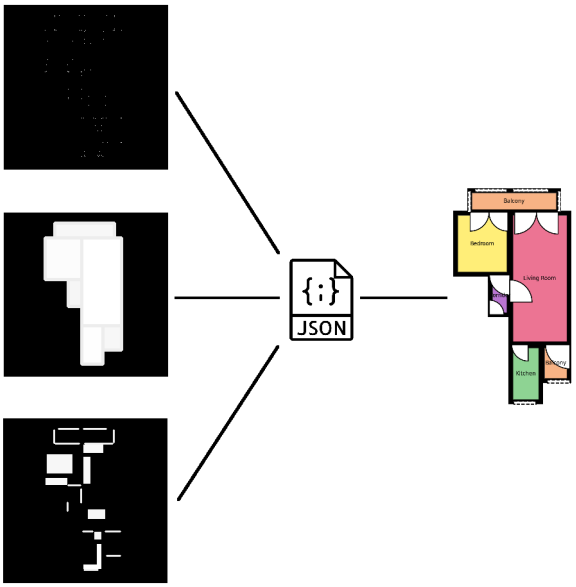



**Figure 4: Representation of the annotation conversion from the FloorNet multi-source format to Structure3D format. Raster images representing rooms and "Icons" and coordinates representing vertexes had to be elaborated and converted in an unique vector description in JSON format. Manual corrections have also been necessary.**

and converted the annotations of FloorNet to the COCO format so that we could use it to train modern models.

## 3.1 Details on data transformation

Beside the plain conversion of annotation format from the original FloorNet format to COCO, as described in Figure 4, in order to facilitate this domain transfer we operated manipulations on both the density maps and the annotation values.

On the input density maps we applied random adjustments of brightness and saturation to training samples. We considered this data processing due to observed chromatic variation in density maps cause by the random selection of points from the point cloud. This filtering of points is required to simplify the creation of density maps, but it creates variations due to non-uniform distribution.

We find that real scans, such as those from FloorNet, are rougher than the smooth surfaces of Structure3D. In those cases we try to smooth hard contrasting variations in the continuity of walls. Also for walls we aim to impose a reasonable wall thickness. Since the scan points are determined at the surface of the wall, thicker walls may be detected as additional rooms between two wall surface scans. To compensate for this, we adjust the FloorNet annotations to reflect thinner walls, which should cause the object detector to be less sensitive to such thicker walls.

Finally we divided the resulting dataset in a training set containing 135 floorplans and a validation set containing 20 images. Given the limited number of data points, we preferred avoiding a further split into a test set. The final version of FloorNet dataset with density maps as input data and annotations in JSON COCO format will be made available as open dataset for other to use.

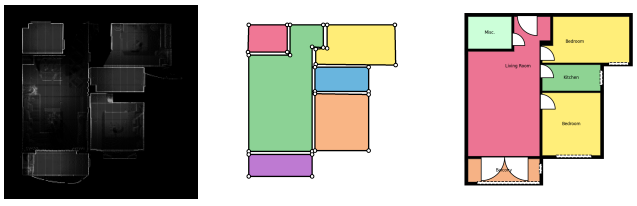
# 4 METHOD

This section presents our technical approach for robust floorplan generation from noisy sensor data.

We extend an object detection based solution, RoomFormer [12], to operate on real-world sensor scans. While we acknowledge the innovative approach proposed by RoomFormer, it has been designed for and evaluated on synthetic data only. We construct a pipeline to transfer the strength of RoomFormer in the synthetic data domain and fine-tune it for real-world sensor data.

## 4.1 Model Baseline identification

In order to define a baseline, it is important to specify the goals of the floorplan extraction. In literature we could find two sub-tasks related to floorplan extraction, defined as: polygonal reconstruction and semantically rich reconstruction. Both tasks are supported by the annotations of Structure3D and FloorNet. While the task of polygonal reconstruction is limited to the detection of the room polygons, the semantically rich reconstruction aims to detect also doors and windows and to classify the room types (see Figure 5).



**(a) Density Map from Structure3D.** **(b) Polygonal reconstruction** **(c) Semantically rich reconstruction**

**Figure 5: Struct3D data sample in three representations.**

The task of floorplan generation from point clouds is relatively young in the field of ML but a notable number of models have been already proposed in literature. The most recent and effective models have been trained and benchmarked with Structure3D, as we saw in the previous section (Table 2). Anyway, almost all models are engaging in the polygonal reconstruction task only. The only exception, to the best of our knowledge, is RoomFormer [12], presented at CVPR 2023. Roomformer has been derived from the Deformable DETR [14] model, and proposes an elegant formulation of the task of Floorplan Generation as a variant of Object Detection. It performs remarkably well on the task of polygonal reconstruction alone and, unlike other models, it has been trained and evaluated also on the task of semantically rich floorplan reconstruction on the Structure3D dataset. We elected RoomFormer, pretrained on Structure3D, as our baseline model for our project. We rerun the evaluation scripts provided by the authors to capture the actual loss values computed on the model predictions. Table 3 represents the final loss values of RoomFormer, trained on Structure3D training set, on Structure3D and FloorNet validation sets, for the two sub tasks we are considering.

It is possible to note that losses on FloorNet are of orders of magnitude higher, providing a quantitative evidence that models trained on Synthetic data don't perform well on real data, as we already observed in our preliminary visual inspections (Figure 2 and Figure 3).

## 4.2 The Basic RoomFormer System

The **RoomFormer** model, as presented by Yue et al. in their work [12], offers a novel approach to 2D floorplan reconstruction from 3D scans in synthetic environments. At its core, RoomFormer is heavily influenced by the **Deformable DETR** [14] model, which in turn is an extension of the original **DETR** (End-to-End Object Detection with Transformers) [2]. While DETR introduced the concept of using transformers for object detection, Deformable DETR further enhanced this by introducing deformable attention modules to capture long-range spatial relationships in images. Adapting from this foundation, RoomFormer modifies the input format from RGB images to black and white density representations and shifts the output format from bounding box detection to polygon segmentation. The model can be trained to solve two distinct tasks: predicting room polygons only, or generating semantically-rich floorplans that include doors, windows, and room types.

The loss function introduced in the RoomFormer model is a pivotal innovation, tailored to address the unique challenges of floorplan reconstruction. The total loss, denoted as $\mathcal{L}$, is defined as a summation over multiple components, each catering to different aspects of the task. Specifically, the formula is given by:

$$\mathcal{L} = \sum_{m}^{M} (\lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^{m} + \lambda_{\text{coord}} \mathcal{L}_{\text{coord}}^{m} + \lambda_{\text{ras}} \mathcal{L}_{\text{ras}}^{m})$$

Where:

- $\mathcal{L}_{\text{cls}}^{m}$ represents the classification loss, which is a standard binary cross-entropy loss.
- $\mathcal{L}_{\text{coord}}^{m}$ is the loss function for vertex coordinates regression, ensuring accurate prediction of polygon vertices.
- $\mathcal{L}_{\text{ras}}^{m}$ is an auxiliary loss related to rasterized polygons, further refining the model's predictions.
- $\lambda_{\text{cls}}, \lambda_{\text{coord}}$, and$\lambda_{\text{ras}}$ are the respective weights for each component of the loss.

For the semantically-rich floorplan task, the loss function is extended to account for the additional semantic components, such as doors, windows, and room types. This ensures that RoomFormer not only predicts the room polygons accurately but also captures the detailed semantic information associated with each polygon.

This adaptation also involved tweaking various hyperparameters to better suit the task at hand. The result is a model that, while sharing a significant portion of its codebase with Deformable DETR, is finely tuned for the specific challenges of floorplan generation. RoomFormer's holistic approach allows for a more integrated and comprehensive representation of indoor spaces, and it has demonstrated superior performance on challenging datasets like Structured3D and SceneCAD, achieving faster inference times compared to previous methods.

## 4.3 Model Training Configurations

RoomFormer had been trained with Structure3D using annotations in COCO format. As we converted the FloorNet dataset to the same format, we were able to train RoomFormer against it using the same script provided by the authors, with just minor changes. We trained Roomformer for both sub tasks: polygonal reconstruction and semantically rich reconstruction. Since FloorNet contains only 155 floorplans, we preferred using it to fine-tune a model version already pretrained with Structure3D rather than train the model from scratch. We validated this choice with an ablation experiment, training RoomFormer on the Polygonal Reconstruction task for 800 epochs using the same regime proposed by the authors on Structure3D and we observed that the model started overfitting the training set while reaching on the validation set performances that were worse compared to the model trained with Structure3D only (see Table 4). We then proceeded on the fine-tune training task. After a few calibration attempts, we set the number of epochs to be 50 for the polygonal reconstruction task and 200 for the semantically rich reconstruction task. The learning rate has been dropped by a factor of 10, as recommended by the authors for retraining. We also experimented with additional hyper-parameter tuning and different optimizers, but we found that benefits are minimal compared to what is achieved by the default training on the Structure3D dataset.

In Table 4 and in Table 5 we report the improvements measured on the final looses on the Polygonal Reconstruction and Semantically Rich Reconstruction tasks respenctively.

In Figure 6 and Figure 7 and visual comparisons of the predictions of the models before and after fine-tuning. The first column presents the scenario of running the model without our fine-tuning; the second column is the output of our fine-tuned model; and the third column is the ground-truth. We can see the performance of our model is superior across multiple scenarios compared to the initial model in both polygon and rich-map floorplan construction.

# 5 EVALUATION ON CUSTOM DATA

This section presents the evaluation of our fine-tune model for floorplan generation on custom data. We utilized a Lidar device

**Table 3: RoomFormer losses for Different Structures**

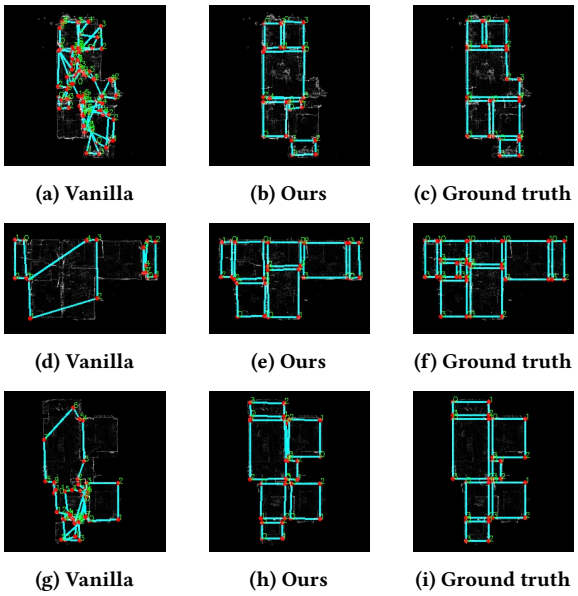
| Metric | Structure3D Polygonal | Structure3D Semantically rich | FloorNet Polygonal | FloorNet Semantically rich |
|---|---|---|---|---|
| $\mathcal{L}$ | 0.2136 | 0.1676 | 1.6661 | 0.7795 |
| $\mathcal{L}_{\text{cls}}$ | 0.0336 | 0.0198 | 0.4792 | 0.1644 |
| $\mathcal{L}_{\text{coord}}$ | 0.0807 | 0.0668 | 0.4109 | 0.3027 |
| $\mathcal{L}_{\text{ras}}$ | 0.0589 | N/A | 0.7897 | N/A |
| $\mathcal{L}_{\text{cls\_room}}$ | N/A | 0.0429 | N/A | 0.3304 |

**Table 4: The loss values used for the polygonal reconstruction task in three conditions: not fine-tuned model, trained from scratch, fine-tuned model.**

| Metric | Not fine-tuned | Trained from Scratch | Fine-tuned |
|---|---|---|---|
| $\mathcal{L}$ | 1.6661 | 1.466 | 0.2136 |
| $\mathcal{L}_{\text{cls}}$ | 0.4792 | 0.1373 | 0.0336 |
| $\mathcal{L}_{\text{coord}}$ | 0.4109 | 0.51097 | 0.0807 |
| $\mathcal{L}_{\text{ras}}$ | 0.7897 | 0.81772 | 0.0589 |

**Table 5: The loss values used for the semantically rich reconstruction task in two conditions: not fine-tuned model and fine-tuned model.**

| Metric | Not fine-tuned | Fine-tuned |
|---|---|---|
| $\mathcal{L}$ | 0.7795 | 0.3377 |
| $\mathcal{L}_{\text{cls}}$ | 0.1644 | 0.0252 |
| $\mathcal{L}_{\text{cls\_room}}$ | 0.3304 | 0.077 |
| $\mathcal{L}_{\text{coord}}$ | 0.3027 | 0.2386 |



**(a) Vanilla** **(b) Ours** **(c) Ground truth**

**(d) Vanilla** **(e) Ours** **(f) Ground truth**

**(g) Vanilla** **(h) Ours** **(i) Ground truth**

**Figure 6: Polygonal Reconstruction grid of a few samples using the same density map from FloorNet as input per row. The first column presents predictions by RoomFormer trained with Structure3D only (Vanilla), the second column presents predictions by RoomFormer retrained with FloorNet (Ours) and the third is the ground truth annotations of polygons.**

to collect the scan of a real environment, processed the data and created a density map representing a floorplan. We then run our model to produce predictions and performed a visual evaluation of the results.

## 5.1 Data cleaning

Shiny surfaces, such a mirrors and windows tend to affect the quality of light-based sensors used for scanning the 3D space. This is cause by the light beam sent by the sensor scattering in different directions, which take a longer path to reach the receptor of the sensor. These distorted measurements create artefacts which are often neglected in synthetic scans.

These artefacts create conditions that are not found in the data used for training the model. To simplify the task, we first need to identify and eliminate those regions in the point cloud that represent obvious anomalies. We do this by clustering points with k-means. Any region that is far from other continuous set of points is marked as anomalies and discarded. Figure 8(a) presents such anomaly points to the top image.

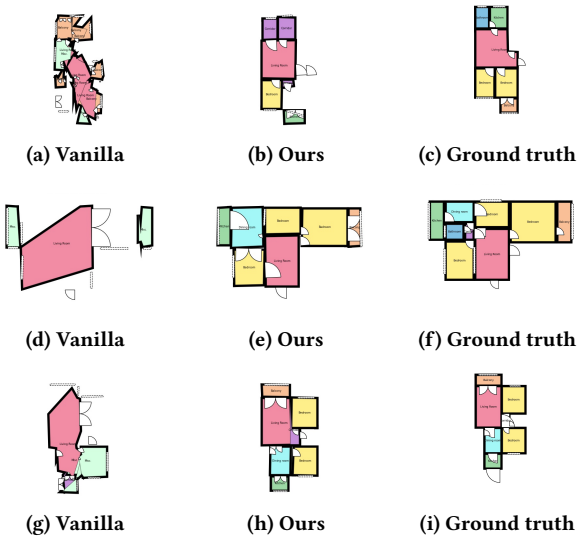
The anomaly of distant points caused by shiny objects is more common to scans collected with Lidar. For scans in the FloorNet we found this data cleaning to be less relevant since their dual-camera depth estimation produces less artefacts.

*Data completion.* Some scans may be dominated by a lot of points localised in a constrained space of the whole scene. We noticed this to be the case for scans collected with a Lidar of a moving agent that spends more time in a room/space that connect multiple rooms, thus accumulating more samples in that space compared to the rest of the spaces. To avoid intensive colors in those locations, the map is split into grids, aiming at having a similar ration of points across grid cells.

## 5.2 Evaluation of the model on Lidar data

So far, all the density maps have been created from either synthetic data (Structure3D dataset) and from point clouds extracted from optical depth measurements collected with Tango devices (FloorNet dataset). We now look at real-world Lidar measurements collected with our device. Data was collected in an experiment site, capturing a run through the environment with the device held in hand.

We align the point cloud to have the majority of walls parallel to the orthogonal coordinates and generate the density map by



(a) Vanilla (b) Ours (c) Ground truth

(d) Vanilla (e) Ours (f) Ground truth

(g) Vanilla (h) Ours (i) Ground truth

**Figure 7: Semantically Rich Reconstruction grid of a few samples using the same density map from FloorNet as input per row. The first column presents predictions by RoomFormer trained with Structure3D only (Vanilla), the second column presents predictions by RoomFormer retrained with FloorNet (Ours) and the third is the ground truth annotations of polygons.**

choosing 50,000 random points, as done by RoomFormer. From the density map we notice that most of the time in scanning the environment was spent in the central room, with fewer points for the adjacent rooms and corridors. Results are visible in Figure 8.

Similar as before, we run the before mode (trained only on Structure3D) and our after model (fine-tuned on the FloorNet dataset). The first observation is that the experiment site is more difficult than the standard apartment samples in both Structure3D and FloorNet datasets used for training. Despite this limitation, we see a substantial improvement in the model after fine-tuning, as shown in Figure 9.

Although the fine-tuned model does not replicate exactly the map geometry, it captures a lot more rooms than the baseline model (trained on Structure3D only).

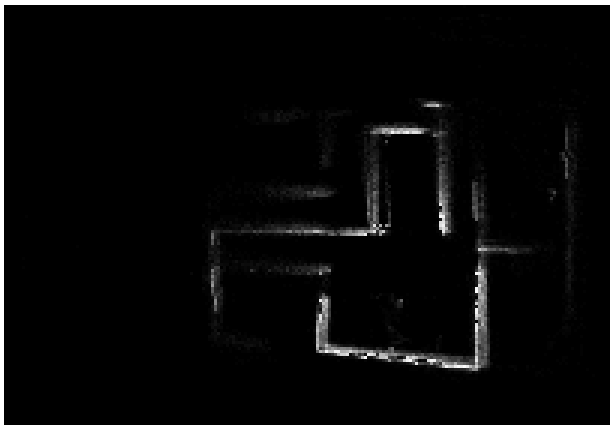
This experiment shows that our fine-tuned model is robust to domain transfer, since we did not train the model with any Lidar generated data, still being able to detect close-enough room shapes. We expect an improved performance once a substantial dataset of Lidar scans has been collected to use for further fine-tuning the model (domain adaptation).

## 6 CONCLUSIONS

This paper shows an improved solution for generating robust floorplans from noisy point clouds of indoor spaces captured using cheap sensors.

Even though the predictions with the fine-tuned version of the model are not perfect, they look extremely encouraging, also considering the limited size of the dataset used for re-training and

(a) Point cloud



(b) Density map

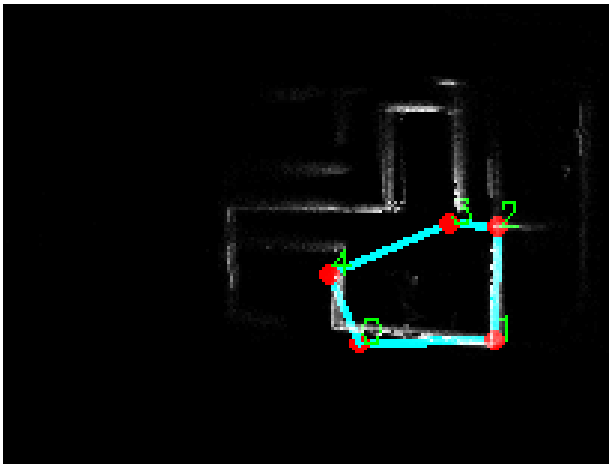
**Figure 8: Lidar data: from Point Cloud to Density Map**

limited effort involved in hyper-parameter tuning. We believe that the model trained with synthetic and real data will produce stronger baselines for future work on this task. We also demonstrated the importance of going beyond synthetic data to real that for model robustness across domains, in particular we show that our model can operate well in noisy sensor data from Lidar scans, a sensing modality that was not present in training data.

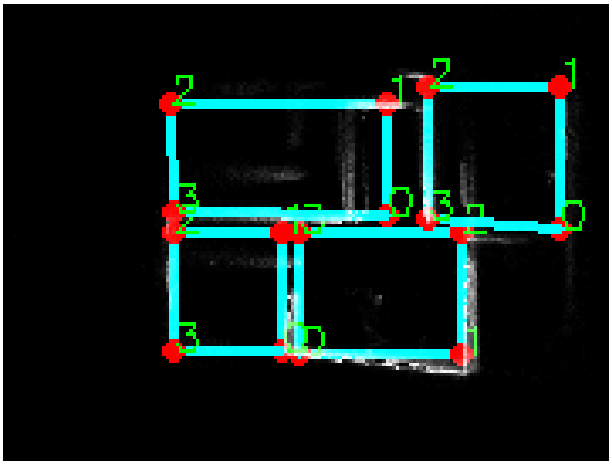
## 7 FUTURE WORK

As next steps, based on the lessons learned from this research, we plan to produce a novel dataset that will be collected from many real-world sites, which we will use to extend the training of our model further. In place measurements will provide the necessary ground truths to complement the Lidar scans, which we will convert into the required COCO format. We believe that investing in the creation and curation of a high quality dataset will prove beneficial in the long-term. Our aim is to make this upcoming dataset publicly available to enable new research in the community.

Nevertheless, we also see the opportunity to design new models that could achieve better floorplan reconstruction. Literature shows that progress has been made in the task of polygonal reconstruction of floorplan, which could be potentially exploited in future iterations of this research stream. One possible candidate is PolyDIffuse [3], a model that makes use of the same diffusion



**(a) Original RoomFormer model**



**(b) RoomFormer model fine-tuned with Floornet**

**Figure 9: Lidar data.**

technique to improve the output of RoomFormer, and is also the current SOTA for the polygonal reconstruction task.

Another stream of research is related to the use of novel techniques for Object Detection. RoomFormer demonstrated that floorplan reconstruction could be seen as a variant of object detection. Porting SOTA techniques from 2020 achieved notable performances on Structure3D. The task of Object Detection has seen major progress over the last years and many advanced techniques have been invented, which brought impressive improvements in the benchmark of the COCO dataset and others. Extensions to our work might also focus on increasing the size, newer backbone models, changes in the architecture or in the training process.

## REFERENCES


[1] W. Boehler and A. Marbs. 2002. 3D scanning instruments. In *Proceedings of the CIPA WG 6 International Workshop on Scanning for Cultural Heritage Recording.*

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. arXiv:2005.12872 [cs.CV]

[3] Jiacheng Chen, Ruizhi Deng, and Yasutaka Furukawa. 2023. PolyDiffuse: Polygonal Shape Reconstruction via Guided Set Diffusion Models. arXiv:2306.01461 [cs.CV]

[4] Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2019. FloorSP: Inverse CAD for Floorplans by Sequential Room-wise Shortest Path. arXiv:1908.06702 [cs.CV]

[5] Ruipeng Gao, Mingmin Zhao, Tao Ye, Fan Ye, Yizhou Wang, Kaigui Bian, Tao Wang, and Xiaoming Li. 2014. Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. 249–260.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV]

[7] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2018. FloorNet: A Unified Framework for Floorplan Reconstruction from 3D Scans. arXiv:1804.00090 [cs.CV]

[8] B. Okorn, X. Xiong, B. Akinci, and D. Huber. 2010. Toward automated modeling of floor plans. In *Proceedings of the Symposium on 3D Data Processing, Visualization and Transmission.*

[9] Sinisa Stekovic, Mahdi Rad, Friedrich Fraundorfer, and Vincent Lepetit. 2021. MonteFloor: Extending MCTS for Reconstructing Accurate Large-Scale Floor Plans. arXiv:2103.11161 [cs.CV]

[10] E. Turner and A. Zakhor. 2014. Floor plan generation and room labeling of indoor environments from laser range data. *Graphical Models* 76, 5 (2014), 338–350.

[11] Wikipedia contributors. 2023. Tango (platform) — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Tango_(platform)&oldid=1159441918 [Online; accessed 12-August-2023].

[12] Yuanwen Yue, Theodora Kontogianni, Konrad Schindler, and Francis Engelmann. 2023. Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries. arXiv:2211.15658 [cs.CV]

[13] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. 2020. Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling. arXiv:1908.00222 [cs.CV]

[14] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv:2010.04159 [cs.CV]


Received 20 August 2023