

Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries

Yuanwen Yue¹ Theodora Kontogianni^{1,2} Konrad Schindler^{1,3} Francis Engelmann^{1,2}

¹ETH Zurich ²ETH AI Center ³Photogrammetry and Remote Sensing, ETH Zurich

Abstract

We address 2D floorplan reconstruction from 3D scans. Existing approaches typically employ heuristically designed multi-stage pipelines. Instead, we formulate floorplan reconstruction as a single-stage structured prediction task: find a variable-size set of polygons, which in turn are variable-length sequences of ordered vertices. To solve it we develop a novel Transformer architecture that generates polygons of multiple rooms in parallel, in a holistic manner without hand-crafted intermediate stages. The model features two-level queries for polygons and corners, and includes polygon matching to make the network end-to-end trainable. Our method achieves a new state-of-the-art for two challenging datasets, Structured3D and SceneCAD, along with significantly faster inference than previous methods. Moreover, it can readily be extended to predict additional information, i.e., semantic room types and architectural elements like doors and windows. Our code and models will be available at: <https://github.com/ywyue/RoomFormer>.

1. Introduction

The goal of floorplan reconstruction is to turn observations of an (indoor) scene into a 2D vector map in birds-eye view. More specifically, we aim to abstract a 3D point cloud into a set of closed polygons corresponding to rooms, optionally enriched with further structural and semantic elements like doors, windows and room type labels.

Floorplans are an essential representation that enables a wide range of applications in robotics, AR/VR, interior design, etc. Like prior work [2, 3, 8, 9, 30], we start from a 3D point cloud, which can easily be captured with RGB-D cameras, laser scanners or SfM systems. Several works [8, 9, 22, 30] have shown the effectiveness of projecting the raw 3D point data along the gravity axis, to obtain a 2D density map that highlights the building’s structural elements (e.g. walls). We also employ this early transition to 2D image space. The resulting density maps are compact and computationally efficient, but inherit the noise and data gaps of the underlying point clouds, hence floorplan recon-

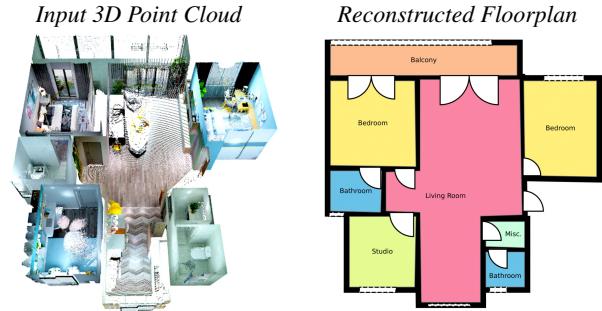


Figure 1. Semantic floorplan reconstruction. Given a point cloud of an indoor environment, *RoomFormer* jointly recovers multiple room polygons along with their associated room types, as well as architectural elements such as doors and windows.

struction remains a challenging task.

Existing methods can be split broadly into two categories that both operate in two stages: *Top-down methods* [8, 30] first extract room masks from the density map using neural networks (e.g., Mask R-CNN [15]), then employ optimization/search techniques (e.g., integer programming [29], Monte-Carlo Tree-Search [4]) to extract a polygonal floorplan. Such techniques are not end-to-end trainable, and their success depends on how well the hand-crafted optimization captures domain knowledge about room shape and layout. Alternatively, *bottom-up methods* [9, 22] first detect corners, then look for edges between corners (i.e. wall segments) and finally assemble them into a planar floorplan graph. Both approaches are strictly sequential and therefore dependent on the quality of the initial corner, respectively room, detector. The second stage starts from the detected entities, therefore missing or spurious detections may significantly impact the reconstruction.

We address those limitations and design a model that directly maps a density image to a set of room polygons. Our model, named *RoomFormer*, leverages the sequence prediction capabilities of Transformers and directly outputs a variable-length, ordered sequence of vertices per room. *RoomFormer* requires neither hand-crafted, domain-specific intermediate products nor explicit corner, wall or room detections. Moreover, it predicts all rooms that make up the floorplan at once, exploiting the parallel nature of the Transformer architecture.

In more detail, we employ a standard CNN backbone to extract features from the birds-eye view density map, followed by a Transformer encoder-decoder setup that consumes image features (supplemented with positional encodings) and outputs multiple ordered corner sequences, in parallel. The floorplan is recovered by simply connecting those corners in the predicted order. Note that the described process relies on the ability to generate hierarchically structured output of variable and a-priori unknown size, where each floorplan has a different number of rooms (with no natural order), and each room polygon has a different number of (ordered) corners. We address this challenge by introducing two-level queries with one level for the room polygons and one level for their corners. The varying numbers of both rooms and corners are accommodated by additionally classifying each query as valid or invalid. The decoder iteratively refines the queries, through self-attention among queries and cross-attention between queries and image features. To enable end-to-end training, we propose a polygon matching strategy that establishes the correspondence between predictions and targets, at both room and corner levels. In this manner, we obtain an integrated model that holistically predicts a set of polygons to best explain the evidence in the density map, without hand-tuned intermediate rules of which corners, walls or rooms to commit to along the way. The model is also fast at inference, since it operates in single-stage feed-forward mode, without optimization or search and without any post-processing steps. Moreover, it is flexible and can, with few straight-forward modifications, predict additional semantic and structural information such as room types, doors and windows (Fig. 1).

We evaluate our model on two challenging datasets, Structured3D [38] and SceneCAD [2]. For both of them, RoomFormer outperforms the state of the art, while at the same time being significantly faster than existing methods. In summary, our contributions are:

- A new formulation of floorplan reconstruction, as the simultaneous generation of multiple ordered sequences of room corners.
- The RoomFormer model, an end-to-end trainable, Transformer-type architecture that implements the proposed formulation via two-level queries that predict a set of polygons each consisting of a sequence of vertex coordinates.
- Improved floorplan reconstruction scores on both Structured3D [38] and SceneCAD [2], with faster inference times.
- Model variants able to additionally predict semantic room type labels, doors and windows.

2. Related Work

Floorplan reconstruction turns raw sensor data (*e.g.* point clouds, density maps, RGB images) into vectorized geometries.

Early methods rely on basic image processing techniques, *e.g.* Hough transform or plane fitting [1, 5, 23, 26–28, 34]. Graph-based methods [6, 14, 16] cast floorplan reconstruction as an energy minimization problem. Recent deep learning methods replace some hand-crafted components with neural networks. Typical top-down methods such as Floor-SP [8] rely on Mask R-CNN [15] to detect room segments and reconstruct polygons of individual room segments by sequentially solving shortest path problems. Similarly, MonteFloor [30] first detects room segments and then relies on Monte-Carlo Tree-Search to select room proposals. Alternative bottom-up methods, such as FloorNet [21] first detect room corners, followed by an integer programming formulation to generate wall segments. This approach, however is limited to Manhattan scenes. Recently, HEAT [9] proposed an end-to-end model following a typical bottom-up pipeline: first detect corners, then classify edge candidates between corners. Although end-to-end trainable, it cannot recover edges from undetected corners. Instead, our approach skips the heuristics-guided processes from both approaches. Without explicit corner, wall or room detection, our approach directly generates rooms as polygons in a holistic fashion.

Transformers for structured reconstruction. Transformers [32], originally proposed for sequence-to-sequence translation tasks, have shown promising performance in many vision tasks such as object detection [7, 39], image/video segmentation [10, 11, 33] and tracking [24]. DETR [7] reformulates object detection as a direct set prediction problem with Transformers which is free from many hand-crafted components, *e.g.* anchor generation and non-maximum suppression. LETR [35] extends DETR by adopting Transformers to predict a set of line segments. PlaneTR [31] follows a similar paradigm for plane detection and reconstruction. These works show the promising potential of Transformers for structured reconstruction without heuristic designs. Our work goes beyond these initial steps and asks the question: *Can we leverage Transformers for structured polygon generation?* Different from predicting primitive shapes that can be represented by a fixed number of parameters (*e.g.*, bounding boxes, lines, planes), polygons are more challenging due to the arbitrary number of (ordered) vertices. While some recent works [18, 20, 37] utilize Transformers for polygon generation in the context of instance segmentation or text spotting, there are two essential differences: (1) They assume a fixed number of polygon vertices, which is not suitable for floorplans. This results in over redundant vertices for simple shapes and insufficient vertices for complex shapes. Instead, our goal is to generate polygons that match the target shape with the correct number of vertices. (2) They rely on bounding box detection as instance initialization, while our single-stage method directly generates multiple polygons in parallel.

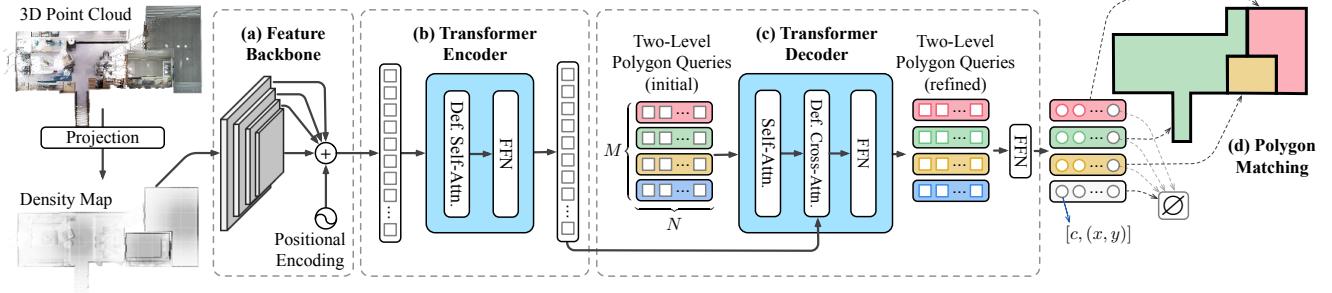


Figure 2. Illustration of the RoomFormer model. Given a top-down-view density map of the input point cloud, (a) the feature backbone extracts multi-scale features, adds positional encodings, and flattens them before passing them into the (b) Transformer encoder. (c) The Transformer decoder takes as input our *two-level* queries, one level for the room polygons (up to M) and one level for their corners (up to N per room polygon). A feed-forward network (FFN) predicts a class c for each query to accommodate for varying numbers of rooms and corners. During training, the polygon matching guarantees optimal assignment between predicted and groundtruth polygons.

3. Method

3.1. Floorplan Representation

A suitable floorplan representation is key to an efficient floorplan reconstruction system. Intuitively, one can decompose floorplan reconstruction as intermediate geometric primitives detection problems (corners, walls, rooms) and tackle them separately, as in prior works [8, 9, 22, 30]. However, such pipelines involve heuristics-driven designs and lack holistic reasoning capabilities.

Our core idea is to cast floorplan reconstruction as a direct set prediction problem of polygons. Each polygon represents a room and is modeled as an ordered sequence of vertices. The edges (*i.e.*, walls) are implicitly encoded by the order of the vertices – two consecutive vertices are connected – thus a separate edge prediction step is not required. Formally, the goal is to predict a *set of sequences* of arbitrary length, defined as $S = \{V_m\}_{m=1}^{M^{\text{gt}}}$, where M^{gt} is the number of sequences per scene, and each sequence $V_m = (v_1^m, v_2^m, \dots, v_{N_m}^m)$ represents a closed polygon (*i.e.* room) defined by N_m ordered vertices.

As each polygon has an arbitrary number of vertices N_m , we model each vertex v_n^m in a polygon V_m by two variables $v_n^m = (c_n^m, p_n^m)$, where $c_n^m \in \{0, 1\}$ indicates whether v_n^m is a valid vertex or not, and $p_n^m \in \mathbb{R}^2$ are the 2D coordinates of the corner in the floorplan. Once the model predicts the ordered corner sequences, we connect all valid corners to obtain the polygonal representation of all rooms.

3.2. Architecture Overview

Fig. 2 shows the model architecture. It consists of (a) a feature backbone that extracts image features, (b) a Transformer encoder to refine the CNN features, (c) a Transformer decoder using *two-level* queries for polygon prediction. (d) During training, the polygon matching module yields optimal assignments between predicted and groundtruth polygons, enabling end-to-end supervision.

CNN backbone. The backbone extracts pixel-level feature maps from the density map $\mathbf{x}_d \in \mathbb{R}^{H \times W}$. Since both local and global context are required for accurately locating corners and capturing their order, we utilize the L multi-scale feature maps $\{I_l\}_{l=1}^L$ from each layer l of the convolutional backbone, where $I_l \in \mathbb{R}^{C \times H_l \times W_l}$. Each feature map is flattened to a feature sequence $I_l \in \mathbb{R}^{C \times H_l W_l}$ and sine/cosine positional encodings $E_l \in \mathbb{R}^{C \times H_l W_l}$ are added to each pixel location. The flattened feature maps are concatenated and serve as multi-scale input to the Transformer encoder.

Multi-scale deformable attention. To avoid the computational and memory complexities of standard Transformers [32], we adopt deformable attention from [39]. Given a feature map, for each query element, the deformable attention only attends to a small set N_s of key sampling points around a reference point, instead of looking over all $H_l W_l$ spatial locations on the feature map, where $N_s \ll H_l W_l$. The multi-scale deformable attention applies deformable attention across multi-scale feature maps and enables encoding richer context. We use multi-scale deformable attention for the self- and cross-attention in the encoder and decoder.

Transformer encoder takes as input the position-encoded multi-scale feature maps and outputs enhanced feature maps of the same resolution. Each encoder layer consists of a multi-scale deformable self-attention module (MS-DSA) and a feed forward network (FFN). In the MS-DSA module, both the query and key elements are pixel features from the multi-scale feature maps. The reference point is the coordinate of each query pixel. A learnable scale-level embedding is added to the feature representation to identify which feature level each query pixel lies in.

Transformer decoder is stacked with multiple layers (Fig. 3a). Each layer consists of a self-attention module (SA), a multi-scale deformable cross-attention module (MS-DCA) and a FFN. Each decoder layer takes in the enhanced image features from the encoder and a set of polygon queries from the previous layer. The polygon queries

first interact with each other in the SA module. In the MS-DCA, the polygon queries attend to different regions of the density map. Finally, the output of the decoder is passed to a shared FFN to predict binary class labels c for each query indicating its validity as a corner.

3.3. Modeling Floorplan As Two-Level Queries

We model floorplan reconstruction as a prediction of a set of sequences. This motivates the two-level polygon queries, one level for room polygons and one level for their vertices. Specifically, we represent polygon queries as $Q \in \mathbb{R}^{M \times N \times 2}$, where M is the maximum number of polygons (*i.e.* room level), N is the maximum number of vertices per polygon (*i.e.* corner level). Using this formulation, we can directly learn *ordered* corner coordinates for each room as queries, which are subsequently refined after each layer in the decoder (Fig. 3b).

We illustrate the structure of one decoder layer in Fig. 3a. The queries in the decoder consist of two parts: *content* queries (*i.e.*, decoder embeddings) and *positional* queries (generated from polygon queries). We denote $Q^i = (x, y)^i$ as the polygon queries in decoder layer i ¹, and $D^i \in \mathbb{R}^{M \times N \times C}$ and $P^i \in \mathbb{R}^{M \times N \times C}$ as the corresponding content and positional queries. Given the polygon query Q^i , its positional query P^i is generated as $P^i = \text{MLP}(\text{PE}(Q^i))$, where PE (Positional Encoding) maps the 2D coordinates to a C -dimensional sinusoidal positional embedding. The decoder performs self-attention on all corner-level queries regardless of the room they belong to. This simple design not only allows the interaction between corners of a single room, but also enables interaction among corners across different rooms (*e.g.* corners on the adjacent walls of two rooms are influencing each other), thus enabling global reasoning. In the multi-scale deformable attention module, we directly use polygon queries as reference points, allowing us to use explicit spatial priors to pool features from the multi-scale feature maps around the polygon vertices. The varying numbers of both rooms and corners are achieved by classifying each query as valid or invalid. In Sec. 3.4, we describe the polygon matching strategy that encourages the queries in corner level to follow a specific order (*i.e.* a sequence) while the queries in room level can be un-ordered (*i.e.* a set).

The key advantage of the above approach is that the room polygons can directly be obtained by connecting the valid vertices in the provided order, without the need for an explicit edge detector as in prior bottom-up methods, *e.g.* [9].

Iterative polygon refinement. Inspired by iterative bounding box refinement in [39], we refine the vertices in each polygon in the decoder layer-by-layer. We use a prediction head (MLP) to predict relative offsets ($\Delta x, \Delta y$) from the

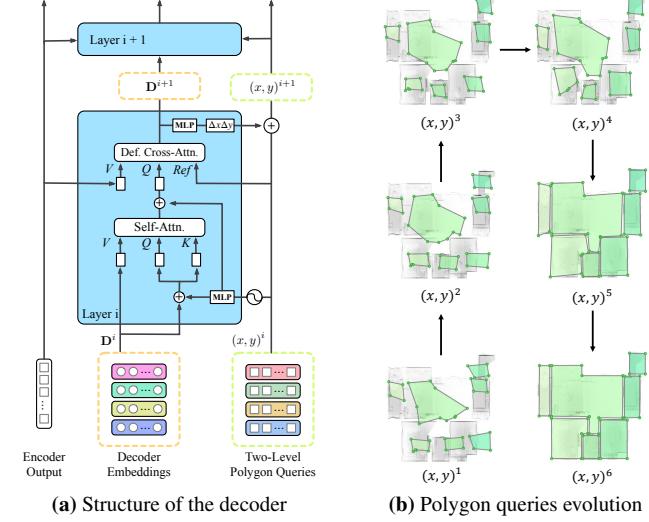


Figure 3. (a) Illustration of one layer of the Transformer decoder (we omit the FFN blocks for clarity). (b) Visualization of the evolution of the polygon queries after each decoder layer.

decoder embeddings and update the polygon queries for the next layer. Both decoder embeddings and polygon queries input to the first layer are randomly initialized. We visualize this iterative refinement process in Fig. 3b. The final predicted labels are used to select valid queries and visualize their position after each layer.

3.4. Polygon Matching

The prediction head of the Transformer decoder outputs a fixed-number M of polygons with a fixed-number N vertices (including non-valid ones, mapped to \emptyset) while the groundtruth contains an arbitrary number of polygons with an arbitrary number of vertices. One of the challenges is to match the fixed-number predictions with the arbitrary-numbered groundtruth to make the network end-to-end trainable. To this end, we introduce a strategy to handle the matching at two levels: set and sequence level.

Let us denote $\hat{S} = \{\hat{V}_m = (\hat{v}_1^m, \hat{v}_2^m, \dots, \hat{v}_N^m)\}_{m=1}^M$ as a set of predicted polygon instances. Each predicted vertex is represented as $\hat{v}_n^m = (\hat{c}_n^m, \hat{p}_n^m)$, where \hat{c}_n^m indicates the probability of a valid vertex and \hat{p}_n^m is the predicted 2D coordinates in normalized space $[0, 1]$. Assume there are M^{gt} polygons in the groundtruth set $S = \{V_m\}_{m=1}^{M^{\text{gt}}}$, where V_m has a length of N_m . For each groundtruth polygon, we first pad it to a length of N vertices so that $V_m = (v_1^m, v_2^m, \dots, v_N^m)$, where $v_n^m = (c_n^m, p_n^m)$ and c_n^m maybe 0, *i.e.* \emptyset . Then we further pad S with additional $M - M^{\text{gt}}$ polygons full of $N \emptyset$ vertices so that $\forall (m \in [M^{\text{gt}} + 1, M] \wedge n \in [1, N]), c_n^m = 0$. At the set level, we find a bipartite matching between the predicted polygons and the groundtruth by searching for a permutation $\hat{\sigma}$ with minimal cost:

¹For simplicity, we drop the polygon and vertex indices.

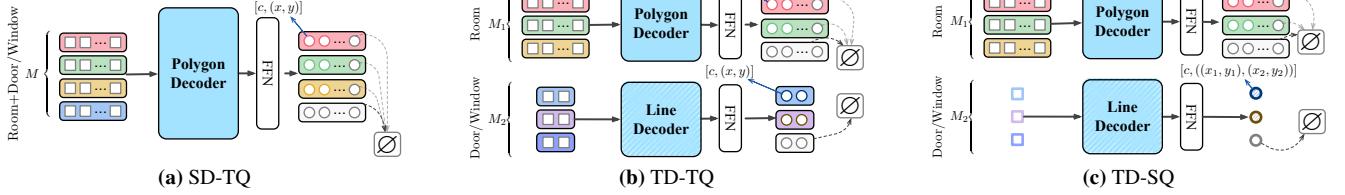


Figure 4. Model variants for semantically-rich floorplans. SD-TQ: Single decoder with two-level queries. TD-TQ: two decoders with two-level queries. TD-SQ: two decoders with single-level queries in the line decoder.

$$\hat{\sigma} = \arg \min_{\sigma} \sum_{m=1}^M \mathcal{D}(V_m, \hat{V}_{\sigma(m)}) \quad (1)$$

where \mathcal{D} is a function that measures the matching cost between groundtruth polygon V_m and a prediction with index $\sigma(m)$. Since we view a polygon as a sequence of vertices, we calculate the matching cost at sequence level and define \mathcal{D} as $\mathbb{1}_{\{m \leq M^{\text{gt}}\}} \lambda_{\text{cls}} \sum_{n=1}^N \|c_n^m - \hat{c}_n^{\sigma(m)}\| + \mathbb{1}_{\{m \leq M^{\text{gt}}\}} \lambda_{\text{coord}} d(P_m, \hat{P}_{\sigma(m)})$, where d measures the sum of pair-wise L_1 distance between groundtruth vertex coordinates *without* padding $P_m = (p_1^m, p_2^m, \dots, p_{N_m}^m)$ and the prediction sliced with the same length $\hat{P}_{\sigma(m)} = (\hat{p}_1^{\sigma(m)}, \hat{p}_2^{\sigma(m)}, \dots, \hat{p}_{N_m}^{\sigma(m)})$. The matching cost \mathcal{D} takes into account both the vertex label and coordinates with balancing coefficients λ_{cls} and λ_{coord} . A closed polygon is a cycle so there exists multiple equivalent parametrizations depending on the starting vertex and the orientation. Here, we fix the groundtruth P_m to always follow counter-clockwise orientation, but can start from any of the vertices. We calculate the distance between $\hat{P}_{\sigma(m)}$ and all possible permutation of P_m and take the minimum as the final d .

Loss functions. After finding the optimal permutation $\hat{\sigma}$ with the Hungarian algorithm, we can compute the loss function which consists of three parts: a vertex label classification loss, a vertex coordinates regression loss and a polygon rasterization loss. The vertex label classification loss is a standard binary cross-entropy:

$$\mathcal{L}_{\text{cls}}^m = -\frac{1}{N} \sum_{n=1}^N c_n^m \cdot \log(\hat{c}_n^{\hat{\sigma}(m)}) - (1 - c_n^m) \cdot \log(1 - \hat{c}_n^{\hat{\sigma}(m)}) \quad (2)$$

Similar to the matching cost function, the L_1 distance serves as loss function for vertex coordinates regression:

$$\mathcal{L}_{\text{coord}}^m = \frac{1}{N_m} \mathbb{1}_{\{m \leq M^{\text{gt}}\}} d(P_m, \hat{P}_{\hat{\sigma}(m)}) \quad (3)$$

We additionally compute the Dice loss [25] between rasterized polygons as auxiliary loss:

$$\mathcal{L}_{\text{ras}}^m = \mathbb{1}_{\{m \leq M^{\text{gt}}\}} \text{Dice}(R(P_m), R(\hat{P}_{\hat{\sigma}(m)})) \quad (4)$$

where R indicates the rasterized mask of a given polygon, using a differentiable rasterizer [18]. We only compute

$\mathcal{L}_{\text{coord}}^m$ and $\mathcal{L}_{\text{ras}}^m$ for predicted polygons with matched non-padded groundtruth while computing $\mathcal{L}_{\text{cls}}^m$ for all predicted polygons (including \emptyset). The total loss \mathcal{L} is then defined as:

$$\mathcal{L} = \sum_m^M (\lambda_{\text{cls}} \mathcal{L}_{\text{cls}}^m + \lambda_{\text{coord}} \mathcal{L}_{\text{coord}}^m + \lambda_{\text{ras}} \mathcal{L}_{\text{ras}}^m) \quad (5)$$

3.5. Towards Semantically-Rich Floorplans

Our method can easily be extended to classify different room types and reconstruct additional architectural details such as doors and windows, while using the same input.

Room types. The two-level polygon queries make our pipeline very flexible to be extended to identify room types. We denote the output embedding from the last layer of the Transformer decoder as $D^{\text{last}} \in \mathbb{R}^{M \times N \times C}$. We then aggregate room-level features by simply averaging corner-level features and obtain an aggregated embedding $\widehat{D}^{\text{last}} \in \mathbb{R}^{M \times C}$. Finally a simple linear projection layer predicts the room label probabilities using a softmax function. Since M is usually larger than the actual number of rooms in a scene, an additional empty class label is used to represent invalid room. We denote t_m as the type for polygon instance V_m . We use the same matching permutation $\hat{\sigma}$ from Eq. 1 to find the matched prediction $\hat{t}_{\hat{\sigma}(m)}$. The room type is supervised by a cross-entropy loss $\mathcal{L}_{\text{room.cls}}^m$.

Doors and windows. A door or a window can be regarded as a line in a 2D floorplan. Intuitively, we can view them as a special ‘‘polygon’’ with 2 vertices. This way, our pipeline can be directly adapted to predict doors and windows *without* architecture modification. It is only required to increase the room-level queries M since more polygons need to be predicted (Fig. 4a). Alternatively, we could use a separate line decoder to predict doors and windows. Since a line can be parameterized by a fixed number of 2 vertices, we can either represent them as two-level queries but with a fixed number of corner-level queries (Fig. 4b), or single-level queries (Fig. 4c). The two-level queries variant is simply an adaptation of our polygon decoder. For the single-level queries variant, we follow LETR [35] that directly predicts the two end points from the query. The performance of each variant is analyzed in Sec. 4.4.

Method	Venue	Fully-neural	Single-stage	t (s)	Room			Corner			Angle		
					Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Floor-SP [8]	ICCV19	X	X	785	89.	88.	88.	81.	73.	76.	80.	72.	75.
MonteFloor [30]	ICCV21	X	X	71	95.6	94.4	95.0	88.5	77.2	82.5	86.3	75.4	80.5
HAWP [36]	CVPR20	✓	X	0.02	77.7	87.6	82.3	65.8	77.0	70.9	59.9	69.7	64.4
LETR [35]	CVPR21	✓	X	0.04	94.5	90.0	92.2	79.7	78.2	78.9	72.5	71.3	71.9
HEAT [9]	CVPR22	✓	X	0.11	96.9	94.0	95.4	81.7	83.2	82.5	77.6	79.0	78.3
RoomFormer (Ours)	-	✓	✓	0.01	97.9	96.7	97.3	89.1	85.3	87.2	83.0	79.5	81.2

Table 1. Foorplan reconstruction scores on Structured3D test set [38]. Our method offers state-of-the-art results while being significantly faster than existing works. Runtime is averaged over the test set. Scores of prior works are as taken from [9, 30].

4. Experiments

4.1. Datasets and Metrics

Datasets. Structured3D [38] is a large-scale photo-realistic dataset containing 3500 houses with diverse floor plans covering both Manhattan and non-Manhattan layouts. It contains semantically-rich annotations including doors and windows, and 16 room types. We adhere to the pre-defined split of 3000 training samples, 250 validation samples and 250 test samples. As in [9, 30], we convert the registered multi-view RGB-D panoramas to point clouds, and project the point clouds along the vertical axis into density images of size 256×256 pixels. The density value at each pixel is the number of projected points to that pixel divided by the maximum point number so that it is normalized to [0, 1].

SceneCAD [2] contains 3D room layout annotations on real-world RGB-D scans of ScanNet [12]. We convert the layout annotations to 2D floorplan polygons. Annotations are only available for the ScanNet training and validation splits, so we train on the training split and report scores on the validation split. We use the same procedure as in Structured3D to project RGB-D scans to density maps.

Metrics. Following [9, 30], for each groundtruth room, we loop through the predictions and find the best-matching reconstructed room in terms of IoU. For the matched rooms we then report precision, recall and F1 scores at three geometry levels: rooms, corners and angles. We compute precision, recall and F1 scores also for the semantic enrichment predictions. For the room type, the metrics are computed like the room metric described above, with the additional constraint that the predicted semantic label must match the groundtruth. A window or door is considered correct if its L_2 distance to the groundtruth element is <10 pixels.

4.2. Implementation Details

Model settings. We use the same ResNet-50 backbone as HEAT [9]. We generate multi-scale feature maps from the last three backbone stages without FPN. The fourth scale feature map is obtained via a 3×3 stride 2 convolution on the final stage. All feature maps are reduced to 256 channels by

Method	t(s)	Room			Corner			Angle		
		IoU	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
Floor-SP [8]	26	91.6	89.4	85.8	87.6	74.3	71.9	73.1		
HEAT [9]	0.12	84.9	87.8	79.1	83.2	73.2	67.8	70.4		
RoomFormer (Ours)	0.01	91.7	92.5	85.3	88.8	78.0	73.7	75.8		

Table 2. Foorplan reconstruction on the SceneCAD val set [2]. a 1×1 convolution. The Transformer consists of 6 encoder and 6 decoder layers with 256 channels. We use 8 heads and $N_s=4$ sampling points for the deformable attention module. The number of room-level queries and corner-level queries are set to $M = 20$ and $N = 40$.

Training. We use the Adam optimizer [17] with a weight decay factor 1e-4. Depending on the dataset size, we train the model on Structured3D for 500 epochs with an initial learning rate 2e-4 and on SceneCAD for 400 epochs with an initial learning rate 5e-5. The learning rate decays by a factor of 0.1 for the last 20% epochs. We set the coefficients for the matcher and losses to $\lambda_{cls} = 2$, $\lambda_{coord} = 5$, $\lambda_{ras} = 1$ and use a single TITAN RTX GPU with 24GB memory for training.

4.3. Comparison with State-of-the-art Methods

Results are summarized in Tab. 1 for Structured3D and in Tab. 2 for SceneCAD. We compare our method to a range of prior approaches that can be grouped in two broad categories: Floor-SP [8] and MonteFloor [30] both rely on Mask R-CNN to segment rooms, followed by learning-free optimization techniques to recover the floorplan. HAWP [36] and LETR [35] are originally generic methods to detect line segments and have been adapted to floorplan reconstruction in [9]. HEAT [9] is an end-to-end trainable neural model that first detects corners, then links them via edge classification. Our RoomFormer outperforms all previous methods on Structured3D (Tab. 1), increasing the F1 score by +1.9 for rooms, +4.7 for corners and +2.9 for angles from the previous state-of-the-art, HEAT. Our RoomFormer is the fastest one among the tested methods, with more than 10 times faster inference than HEAT. MonteFloor additionally employs a Douglas-Peucker algorithm [13] for post-processing to simplify the topology of the output polygons. By contrast, RoomFormer does not rely on any post-processing steps.

Method	t(s)	Room		Corner		Angle		
		IoU	Prec.	Rec.	F1	Prec.	Rec.	F1
HEAT [9]	0.12	52.5	50.9	51.1	51.0	42.2	42.0	41.6
RoomFormer (Ours)	0.01	74.0	56.2	65.0	60.3	44.2	48.4	46.2

Table 3. Cross-data generalization. Models are trained on Structured3D train set but evaluated on SceneCAD val set. Our method shows significant robustness when the train-test domains differ.

Method	Door/Window		Room*		Room		Corner		Angle	
	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1
SD-TQ (Fig. 4a)	81.1		70.7		94.3		83.9		76.7	
TD-TQ (Fig. 4b)		80.8		71.4		93.4		82.0		73.7
TD-SQ (Fig. 4c)		81.7		74.4		94.9		84.2		75.9

Table 4. Semantically-rich floorplan reconstruction scores on Structured3D test set. The Room* metric is similar to Room, but additionally considers the correct room type classification.

On SceneCAD (Tab. 2), we compare with two representative methods (for which code is available) from optimization-based and fully-neural categories Floor-SP and HEAT. RoomFormer achieves notable improvement over other methods especially when it comes to corner/angle precision. For more details on running competing methods, please see the supplementary.

Cross-data generalization. We further evaluate the ability of our model to generalize across datasets. For that, we train on Structured3D training set and evaluate on SceneCAD validation set (without fine-tuning on SceneCAD). We compare with the current state-of-the-art end-to-end model HEAT and report scores in Tab. 3. Our model generalizes better to unseen data characteristics. It outperforms HEAT significantly in almost every metric and particularly on room IoU (**74.0 vs. 52.5**). We attribute the better generalization to the learned global reasoning capacity rather than focusing on separate corner detection and edge classifications as in HEAT.

Qualitative results are presented in Fig. 5 and Fig. 6. The quality of the floorplan generated by two-stage pipelines is strongly affected by errors in the first stage, *e.g.*, missing rooms with MonteFloor (3rd, 4th row) and missing corners and edges with HEAT (5th, 6th row). Instead, our holistic single-stage model produces more accurate predictions while being able to capture geometric details. We observe a similar pattern in Fig. 6. HEAT suffers from missing corners and edges when the input point cloud is sparse (last 4 rows), while our RoomFormer handles these cases more robustly. Floor-SP forces the generated polygon to contain its room segmentation mask completely which, however, results in redundant corners (4th, 5th row). By contrast, our RoomFormer produces more plausible results without imposing any hard constraints.

4.4. Semantically-Rich Floorplans

The quantitative results on semantically-rich floorplan are summarized in Tab. 4. We observe that separating the

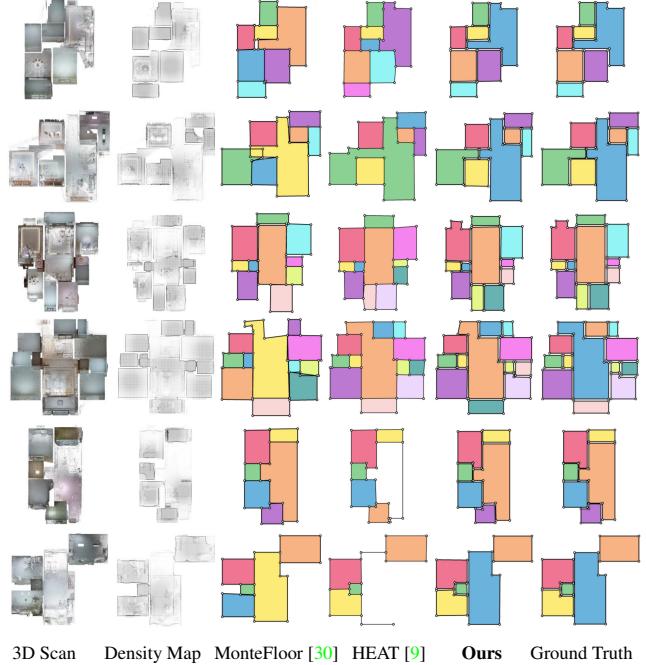


Figure 5. Qualitative evaluations on Structured3D [38]. Best viewed in color on a screen and zoom in. Colors are assigned based on room locations, *without* semantic meaning.

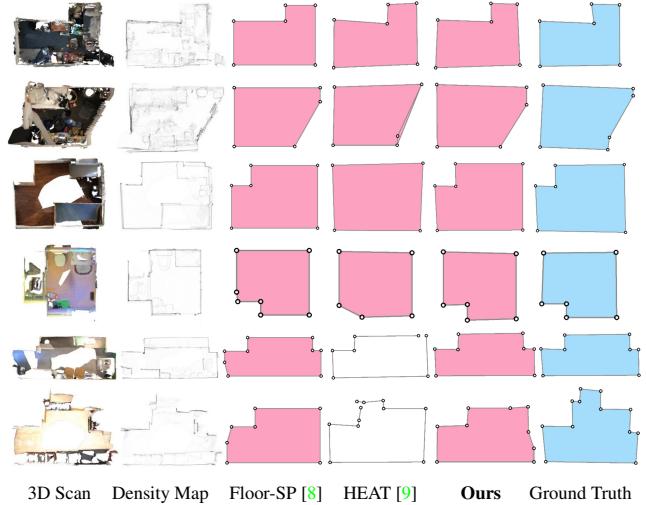


Figure 6. Qualitative evaluations on SceneCAD [2]. HEAT is affected by missing corners and edges (last 4 rows). Floor-SP tends to produce redundant corners due to its containment constraint (4th, 5th row). Our method is more robust in these cases.

room and door/window decoding can help improve room type classification since room and door/window may have significantly different geometry and semantic properties. However, the second-best performing method is SD-TQ where we use our original model with more polygon queries to model the doors and windows. From the two variants with line decoder, the single-level queries (TD-SQ) work better than the two-level queries (TD-TQ), suggesting that for shapes represented by fixed number of parameters (*e.g.*,

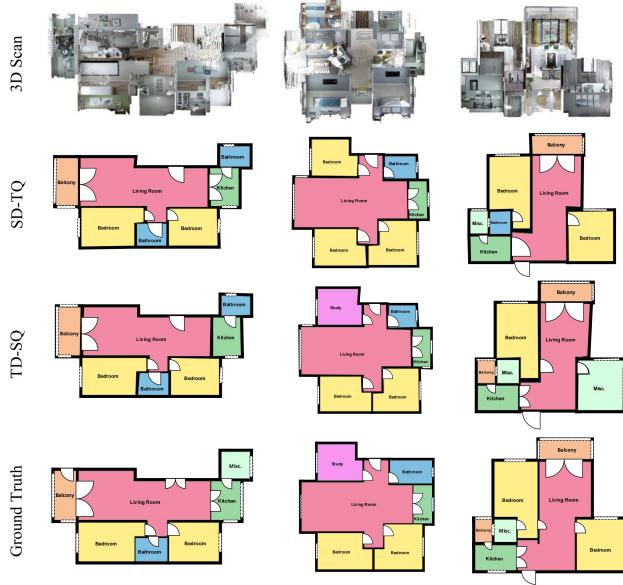


Figure 7. Results of semantic-rich floorplans. Dashed lines represent windows, arcs represent doors. The radius, number and orientation of arcs are determined by the length and direction of the predicted lines. Best viewed in color on a screen.

lines), single-level queries are sufficient. We show our qualitative results on rich floorplan reconstruction in Fig. 7. For visualization purposes, we use arcs to represent doors. Interestingly, our single decoder variant (SD-TQ) incorrectly identifies a room as “bathroom” instead of “misc” (3rd col.). However, this follows house design principles better than the groundtruth, since each house usually has a bathroom.

4.5. Analysis Experiments

Multi-scale feature maps. We leverage multi-scale feature maps to aggregate both local and global context for jointly vertex positioning and order capturing. To validate this design, we conduct ablations by using only a single level feature map obtained from the last stage of ResNet-50. Tab. 5 shows that multi-scales feature maps can significantly improve the metrics for room/corner/angle, which indicates that local and global context are crucial for our Transformers for structured reasoning.

Iterative polygon refinement. We propose to directly learn vertex sequence coordinates as queries and refine them iteratively layer-by-layer, and use the updated positions as new reference points for deformable cross-attention. We ablate this by removing the refinement process and keeping the reference points static in all layers. Tab. 5 suggests that the refinement strategy significantly improves the performance.

Loss functions. We use three loss components to supervise our network (Tab. 6). Since the vertex label classification loss is essential and cannot be removed, we only ablate for $\mathcal{L}_{\text{coord}}$ and \mathcal{L}_{ras} . In the first experiment, we remove $\mathcal{L}_{\text{coord}}$ and replace the matching cost for sequence coordinates with

Multi-Scale Features	Polygon Refinement	Room		Corner		Angle	
		Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
-	✓	93.9	92.9	87.8	79.6	83.3	75.6
✓	-	94.8	93.0	88.7	80.7	84.2	76.7
✓	✓	96.5	95.3	91.2	82.8	88.3	80.3

Table 5. Model analysis. Impact of multi-scale features and polygon refinement. Scores are on Structured3D validation set.

\mathcal{L}_{cls}	$\mathcal{L}_{\text{coord}}$	\mathcal{L}_{ras}	Settings		Room		Corner		Angle	
			Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
✓	-	✓	84.3	83.2	75.4	68.8	70.6	64.5		
✓	✓	-	96.0	94.5	89.9	81.7	87.0	79.2		
✓	✓	✓	96.5	95.3	91.2	82.8	88.3	80.3		

Table 6. Loss analysis. Impact of various losses. \mathcal{L}_{cls} is always required for determining valid corners. The combination of $\mathcal{L}_{\text{coord}}$ and \mathcal{L}_{ras} yields best results. Scores are on Structured3D val set.

M	N	t (ms)	Settings		Room		Corner		Angle	
			Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
15	30	8.4	95.7	94.5	90.2	82.2	86.5	79.0		
20	40	10.8	96.3	95.0	90.8	82.7	87.8	80.0		
20	50	12.8	96.2	94.6	90.5	82.1	87.6	79.5		
30	40	13.9	96.8	95.5	91.1	82.7	88.1	80.1		
30	50	16.8	96.0	94.8	90.6	82.4	87.5	79.7		

Table 7. Analysis on number of queries. The reported scores are on Structured3D validation set averaged over three runs.

a matching cost for rasterized mask. This leads to a large drop in all metrics. Next, we only remove \mathcal{L}_{ras} which incurs a smaller drop in all metrics. We conclude that the sequence coordinates regression loss is essential and the rasterization loss serves as an auxiliary loss.

Number of queries. We study the effect of different number of queries in each level in Tab. 7. The number of queries is based on: 1) the maximum number of rooms and the maximum number of corners in each room in the training dataset. 2) computational efficiency. We chose $M = 20$ and $N = 40$ as a good compromise between model performance, inference time, as well as training time.

5. Conclusion

In this work, we have introduced RoomFormer, a simple and direct model for 2D floorplan reconstruction formulated as a polygon estimation problem. The network learns to predict a varying number of rooms per floorplan, each room represented as a varying length of ordered corner sequence. Our single-stage, end-to-end trainable model shows significant improvements over prior multi-stage and heuristics-driven methods, both in performance and speed metrics. We hope our approach inspires more applications in other polygonal reconstruction tasks.

Acknowledgments We thank the authors of HEAT and MonteFloor for providing results on Structured3D for convenient comparison. Theodora Kontogianni and Francis Engelmann are post-doctoral research fellows at the ETH AI Center.

References

- [1] Antonio Adan and Daniel Huber. 3D Reconstruction of Interior Wall Surfaces Under Occlusion and Clutter. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DPVT)*, 2011. 2
- [2] Armen Avetisyan, Tatiana Khanova, Christopher Choy, Denver Dash, Angela Dai, and Matthias Nießner. SceneCAD: Predicting Object Alignments and Layouts in RGB-D Scans. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 6, 7, 11, 13, 15
- [3] Maarten Bassier and Maarten Vergauwen. Unsupervised Reconstruction of Building Information Modeling Wall Objects From Point Cloud Data. *Automation in Construction*, 120, 2020. 1
- [4] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A Survey of Monte Carlo Tree Search Methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 2012. 1
- [5] Angela Budroni and Jan Boehm. Automated 3D Reconstruction of Interiors From Point Clouds. *International Journal of Architectural Computing*, 8, 2010. 2
- [6] Ricardo Cabral and Yasutaka Furukawa. Piecewise Planar and Compact Floorplan Reconstruction From Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end Object Detection With Transformers. In *European Conference on Computer Vision (ECCV)*, 2020. 2
- [8] Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floor-SP: Inverse CAD for Floorplans by Sequential Room-Wise Shortest Path. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 6, 7, 11, 13, 15
- [9] Jiacheng Chen, Yiming Qian, and Yasutaka Furukawa. HEAT: Holistic Edge Attention Transformer for Structured Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 6, 7, 11, 12, 13, 15
- [10] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for Video Instance Segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 2
- [11] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel Classification Is Not All You Need for Semantic Segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2
- [12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6
- [13] David H Douglas and Thomas K Peucker. Algorithms for the Reduction of the Number of Points Required to Represent A Digitized Line or Its Caricature. *Cartographica: the International Journal for Geographic Information and Geovisualization*, 1973. 6, 15
- [14] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Reconstructing Building Interiors From Images. In *International Conference on Computer Vision (ICCV)*, 2009. 2
- [15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2
- [16] Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. Structured Indoor Modeling. In *International Conference on Computer Vision (ICCV)*, 2015. 2
- [17] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6
- [18] Justin Lazarow, Weijian Xu, and Zhuowen Tu. Instance Segmentation With Mask-Supervised Polygonal Boundary Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 5
- [19] Muxingzi Li, Florent Lafarge, and Renaud Marlet. Approximating Shapes in Images with Low-complexity Polygons. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 15
- [20] Justin Liang, Namdar Homayounfar, Wei-Chiu Ma, Yuwen Xiong, Rui Hu, and Raquel Urtasun. Polytransform: Deep Polygon Transformer for Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [21] Cheng Lin, Changjian Li, and Wenping Wang. Floorplanning-jigsaw: Jointly Estimating Scene Layout and Aligning Partial Scans. In *International Conference on Computer Vision (ICCV)*, 2019. 2
- [22] Chen Liu, Jiaye Wu, and Yasutaka Furukawa. Floornet: A Unified Framework for Floorplan Reconstruction From 3D Scans. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 3
- [23] Josep Lladós, Jaime López-Krahe, and Enric Martí. A System to Understand Hand-drawn Floor Plans Using Subgraph Isomorphism and Hough Transform. *Machine Vision and Applications*, 1997. 2
- [24] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object Tracking With Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [25] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *International Conference on 3D Vision (3DV)*, 2016. 5
- [26] Aron Monszpart, Nicolas Mellado, Gabriel J Brostow, and Niloy J Mitra. RAPter: Rebuilding Man-made Scenes With Regular Arrangements of Planes. *ACM Transactions on Graphics (TOG)*, 34, 2015. 2
- [27] Brian Okorn, Xuehan Xiong, Burcu Akinci, and Daniel Huber. Toward Automated Modeling of Floor Plans. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DPVT)*, 2010. 2
- [28] Victor Sanchez and Avideh Zakhor. Planar 3D Modeling of Building Interiors From Point Cloud Data. In *International Conference on Image Processing (ICIP)*, 2012. 2

- [29] Alexander Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, 1998. 1
- [30] Sinisa Stekovic, Mahdi Rad, Friedrich Fraundorfer, and Vincent Lepetit. MonteFloor: Extending MCTS for Reconstructing Accurate Large-Scale Floor Plans. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 6, 7, 11, 12
- [31] Bin Tan, Nan Xue, Song Bai, Tianfu Wu, and Gui-Song Xia. Planetr: Structure-guided Transformers for 3D Plane Recovery. In *International Conference on Computer Vision (ICCV)*, 2021. 2
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3
- [33] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation With Transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2
- [34] Xuehan Xiong, Antonio Adan, Burcu Akinci, and Daniel Huber. Automatic Creation of Semantically Rich 3D Building Models From Laser Scanner Data. *Automation in Construction*, 31, 2013. 2
- [35] Yifan Xu, Weijian Xu, David Cheung, and Zhuowen Tu. Line Segment Detection Using Transformers Without Edges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5, 6, 11, 13
- [36] Nan Xue, Tianfu Wu, Song Bai, Fudong Wang, Gui-Song Xia, Liangpei Zhang, and Philip HS Torr. Holistically-attracted Wireframe Parsing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 11
- [37] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text Spotting Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [38] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3D: A Large Photo-Realistic Dataset for Structured 3D Modeling. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 6, 7, 11, 12
- [39] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-end Object Detection. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3, 4

Appendix

In the appendix, we provide additional ablation studies on the self-attention design of the two-level queries (Sec. A), we describe an alternative ‘tight’ room layout used by some baselines and the scores of our model trained with this layout (Sec. B), further implementation details and more results on semantically-rich floorplans (Sec. C). We also provide details for running competing methods Floor-SP [8] and HEAT [9], as well as a learning-free baseline on the SceneCAD dataset [2] (Sec. D).

A. Self-Attention of the Two-Level Queries

In the main paper, the Transformer decoder performs self-attention on all vertex-level queries regardless of the polygon they belong to (Tab. 8, ②). Alternatively, in this experiment, we restrict the vertex-level queries to attend only to vertices within the same polygon. Implementation-wise, we add an attention mask to prevent attention from vertex-level queries in one polygon to vertex-level queries of another polygon. We find that this restricted form of attention leads to an overall reduced performance (Tab. 8, ①). We conclude that the self-attention between vertices across all polygons plays an important role for structured reasoning. In particular, the attention mechanism across multiple polygons seems to help fine-tune the vertex positions of one polygon by attending to the vertex positions of its neighboring polygons.

Settings	Room			Corner			Angle		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
① Intra-Poly. Attn.	95.5	94.2	94.8	90.3	81.9	85.9	87.2	79.1	83.0
② Inter-Poly. Attn.	96.5	95.3	95.9	91.2	82.8	86.8	88.3	80.3	84.1

Table 8. Impact of self attention design on Structured3D val. Using self-attention only between the vertices of a single polygon leads to a drop in the F1 score, showing the usefulness of extending the effect of vertices across all polygons.

Method	Room			Corner			Angle		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Floor-SP [8]	89.	88.	88.	81.	73.	76.	80.	72.	75.
MonteFloor [30]	95.6	94.4	95.0	88.5	77.2	82.5	86.3	75.4	80.5
HAWP [36]	77.7	87.6	82.3	65.8	77.0	70.9	59.9	69.7	64.4
LETR [35]	94.5	90.0	92.2	79.7	78.2	78.9	72.5	71.3	71.9
HEAT [9]	96.9	94.0	95.4	81.7	83.2	82.5	77.6	79.0	78.3
RoomFormer (Ours)	97.9	96.7	97.3	89.1	85.3	87.2	83.0	79.5	81.2
RoomFormer* (Ours)	97.2	96.2	96.7	91.6	83.4	87.3	88.3	80.5	84.2

Table 9. Foorplan reconstruction scores on Structured3D test. For a complete and fair comparison, we complement Tab. 1 from the main paper with the result of our model trained on tight room layouts, indicated by *. Cyan and orange mark the two top scores.

B. Tight Room Layouts

We represent floorplans as a set of closed polygons, which is consistent with the groundtruth annotations in

Structured3D [38]. The advantage of this representation is that the thickness of inner walls is implicitly provided by the distance between neighboring room polygons (see Fig. 8, right column). Alternatively, HEAT [9] predict floorplans as planar graphs, which means that the walls of adjacent rooms are represented by a single shared edge in the graph. In particular, this simpler graph representation can approximate the true floorplan only up to the thickness of the walls. Furthermore, to train HEAT, an additional pre-processing step is required which merges the groundtruth edges of neighboring polygons into a single shared edge. It is important to note that the evaluation still runs on the unmodified groundtruth floorplans. Therefore, during evaluation, HEAT performs a post-processing step to obtain a set of closed room polygons from the estimated planar graph.

To show that the improved performance of our model is independent of the layout representation, we train a model on the same ‘tight’ room layout representation as HEAT and report the scores in Tab. 9 marked with a star (*). Note again that all models are evaluated on the same non-processed groundtruth annotations of Structured3D [38]. We observe that the room metrics of the model trained on tight room layouts drop a bit compared with our original model, while still outperforming all other methods. The drop in scores is not surprising since the modified training data is only an approximation of the original groundtruth used for evaluation. Furthermore, the room metrics penalize overlap between rooms [30]. Results from tight room layouts are more likely to overlap, which can potentially degrade the room metrics. Interestingly, the angle metrics improve, especially when it comes to angle precision, which already outperforms MonteFloor. In the tight room layout, we encourage corners in adjacent walls to have exactly complementary angles. This implicit constraint might help the model to reason about angle relationships, thus benefiting the angle metrics. More qualitative results can be found in Fig. 8.

C. Semantically-Rich Floorplan Models

We proposed three model variants for reconstructing semantically-rich floorplans: (1) SD-TQ: single decoder with two-level queries. (2) TD-TQ: two decoders with two-level queries. (3) TD-SQ: two decoders with single-level queries in the line decoder. Here, we explain additional implementation details along with more results. We also describe our heuristic for drawing the door arcs based on the length of the door segments.

C.1. Implementation details

We describe the implementation details of the three model variants to extend our RoomFormer architecture to predict room types, doors and windows (Fig. 4, main paper).

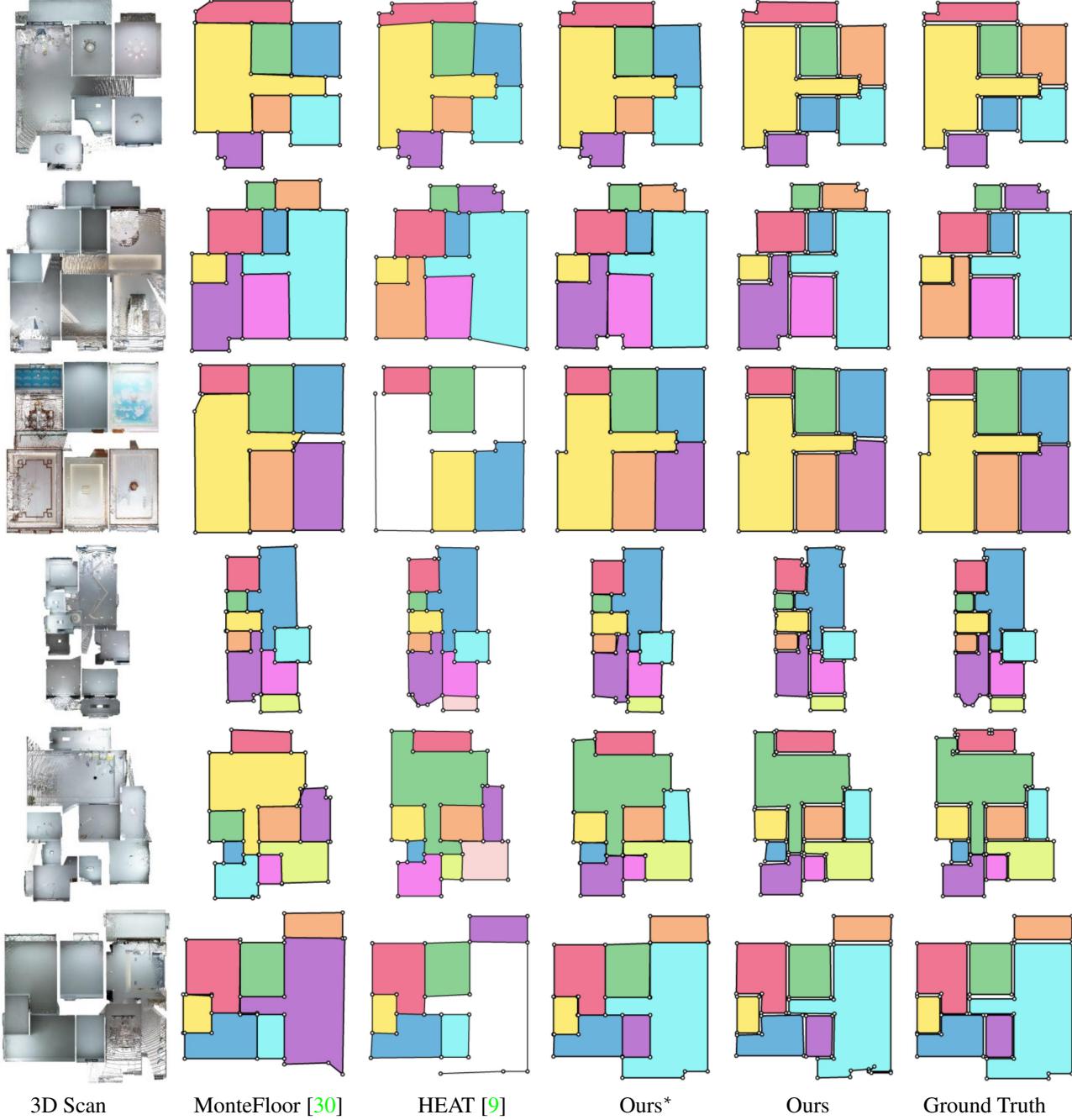


Figure 8. More qualitative evaluations on Structured3D [38]. Colors are assigned based on room locations, *without* semantic meaning. Ours* denotes the result of our model trained on tight room layout. (Best viewed in color on a screen and zoomed in.)

Method	Door/Window			Room*			Room			Corner			Angle		
	Prec.	Rec.	F1												
SD-TQ	83.4	79.0	81.1	71.5	70.0	70.7	95.3	93.3	94.3	86.0	81.8	83.9	78.6	74.9	76.7
TD-TQ	82.6	79.1	80.8	71.9	70.9	71.4	94.0	92.8	93.4	84.2	80.0	82.0	75.6	71.9	73.7
TD-SQ	85.6	78.2	81.7	74.8	74.0	74.4	95.4	94.4	94.9	85.8	82.6	84.2	77.3	74.5	75.9

Table 10. Detailed scores of semantically-rich floorplan reconstruction on Structured3D test set [38].

Algorithm 1 Algorithm for plotting doors as arcs

Input: a set of predicted lines $L = \{l_i\}_{i=1}^{N^l}$, where $l_i = ((x_1^i, y_1^i), (x_2^i, y_2^i))$

Output: plotting of a set of arcs

- 1: calculate the median length m of all the lines L
- 2: **for** l_i in L **do**
- 3: **if** $\|y_2^i - y_1^i\| > \|x_2^i - x_1^i\|$ **then**
- 4: **if** $y_2^i > y_1^i$ **then** $e_1^i = (x_1^i, y_1^i)$ and $e_2^i = (x_2^i, y_2^i)$
- 5: **else** $e_1^i = (x_2^i, y_2^i)$ and $e_2^i = (x_1^i, y_1^i)$
- 6: **end if**
- 7: **else**
- 8: **if** $x_2^i > x_1^i$ **then** $e_1^i = (x_2^i, y_2^i)$ and $e_2^i = (x_1^i, y_1^i)$
- 9: **else** $e_1^i = (x_1^i, y_1^i)$ and $e_2^i = (x_2^i, y_2^i)$
- 10: **end if**
- 11: **end if**
- 12: **if** $\text{len}(l_i) < 1.5 \times m$ **then**
- 13: draw a quadrant centered at e_1^i with a radius of $\text{len}(l_i)$ from e_2^i clockwise
- 14: **else**
- 15: **if** $\|y_2^i - y_1^i\| > \|x_2^i - x_1^i\|$ **then**
- 16: draw two opposite quadrants centered at e_1^i and e_2^i with a radius of $\text{len}(l_i)/2$ at the right side of $\overrightarrow{e_1^i e_2^i}$
- 17: **else**
- 18: draw two opposite quadrants centered at e_1^i and e_2^i with a radius of $\text{len}(l_i)/2$ at the left side of $\overrightarrow{e_1^i e_2^i}$
- 19: **end if**
- 20: **end if**
- 21: **end for**

SD-TQ. Single decoder with two-level queries. We take our original architecture and increase the number of room-level queries M from 20 to 70 while keeping the number of corner-level queries N unchanged. In addition, since there is no need to rasterize lines, we remove the rasterization loss \mathcal{L}_{ras} . We predict all the semantic types (room types, door or window) from the aggregated room-level features of the output embedding of the polygon decoder, as described in Sec. 3.5 of the main paper.

TD-TQ. Two decoders with two-level queries. We add a separate line decoder with the same architecture of the polygon decoder. In the line decoder, we set the number of line-level queries to 50 and the number of corner-level queries to 2. We remove the rasterization loss \mathcal{L}_{ras} term for the line decoder (still used in the polygon decoder). Room types are predicted from the aggregated room-level features of the output embedding of the polygon decoder, while door/window types are predicted from the aggregated line-level features of the output embedding of the line decoder.

TD-SQ. Two decoders with single-level queries. We add a separate line decoder which takes single-level queries,

and predicts the coordinates of the two end-points of each line directly, similar to [35]. We set the number of single-level queries to 50. Room types are predicted as in TD-TQ. Door/window types are predicted based on the output embedding of the line decoder.

C.2. Plotting doors

To obtain floorplan illustrations that are closer to actual floorplans used by architects, we follow the typical notation and represent doors as arcs. Note that this step is only for visualization purposes and is not part of the annotations in the training datasets. In particular, we cannot predict towards which side a door opens and if it is a double door or a single door. The visualization used in this paper is based on a heuristic which creates a double door when the predicted length of the door exceeds a certain threshold. The exact heuristic is shown in Algorithm 1.

C.3. Additional results

Tab. 10 complements Tab. 4 in the main paper by providing more detailed scores of the three model variants on semantically-rich floorplan reconstruction. By comparing the single-decoder variant (SD-TQ) with the two-decoder variants (TD-TQ and TD-SQ), we find that separate decoders can help improve room type classification. Our polygon queries are designed for geometries with varying number of vertices. Since a line has a fixed number of 2 vertices, the single-level query variant (TD-SQ) works better than the two-level query variant (TD-TQ). We provide more qualitative results in Fig. 9.

D. Running Competing Approaches

For comparison on the SceneCAD dataset [2] in the main paper, we select two representative methods from both optimization-based and fully-neural categories Floor-SP [8] and HEAT [9] that offer-state-of-the-art results in floorplan reconstruction and have a public codebase. For a more complete comparison, here we also report the performance of a heuristic-guided pipeline which is free of any deep learning components. Here we describe in detail how we adapted those methods for the SceneCAD dataset [2].

Floor-SP: We used the official implementation with a change in the sequential room-wise shortest path module. We project 3D scans into density images of size 256×256 pixels. Unlike Structured3D, the SceneCAD dataset usually contains only one room per scene, which will result in a larger occupancy pixel areas for a single room. While this does not cause problems for HEAT and our approach, it can lead to a large search space for the sequential room-wise shortest path module in Floor-SP since each room mask contains a large number of pixels. Using Floor-SP default settings, we observe the solver cannot find a solution for



Figure 9. Additional qualitative results on semantically-rich floorplans. (Best viewed in color on a screen and zoomed in.)

Method	t(s)	Room		Corner			Angle		
		IoU	Prec.	Rec.	F1	Prec.	Rec.	F1	
Non-learned	0.22	86.0	66.0	80.8	72.7	45.0	55.2	49.6	
Floor-SP [8]	26	91.6	89.4	85.8	87.6	74.3	71.9	73.1	
HEAT [9]	0.12	84.9	87.8	79.1	83.2	73.2	67.8	70.4	
RoomFormer (Ours)	0.01	91.7	92.5	85.3	88.8	78.0	73.7	75.8	

Table 11. Floorplan reconstruction on the SceneCAD val set [2].

many scenes due to the computational complexity from the larger number of pixels per room. Therefore, we down-sample the density map to 64×64 pixels (a similar size with a single room in the density map of Structured3D) to help reduce the search space. Please note, we train other modules (room mask extraction and corner/edge detection) on the original density map without down-sampling.

HEAT: We used the official implementation with a batch size of 10. We train the model for 400 epochs and found that longer training times would not help to improve the performance further. For the experiments on cross-data generalization, we directly load the released official checkpoints

trained on Structured3D and evaluate them on SceneCAD.

Non-learned baseline: We first project the 3D scan along the vertical axis into an occupancy map of size 256×256 pixels. A pixel is occupied if at least one point is projected to this pixel. To mitigate the impact of missing scans, we apply dilation and erosion to fill holes in the occupancy map. Then we employ a learning-free polygon vectorization algorithm [19] to extract closed polygons from the occupancy map, finally followed by a Douglas-Peucker algorithm [13] to remove redundant corners.

We complement Tab. 2 of the main paper with the result of the non-learned baseline (the 1st row in Tab. 11). We report the running time of the polygon vectorization process. Surprisingly, the pipeline achieves a room IoU of 86.0 and a corner recall of 80.8. However, the angle metrics indicate that those polygons usually fail to accurately describe the geometry of the actual room shapes. This suggests that the non-learned baseline can only capture the rough shape of the floorplan compared with methods which utilize deep learning.