FRI-Net: Floorplan Reconstruction via Room-wise Implicit Representation Supplementary Material

Honghao Xu, Juzhan Xu, Zeyu Huang,
Pengfei Xu, Hui Huang, and Ruizhen Hu*

Shenzhen University

In the supplementary material, we provide the technical details of the room-wise encoder in Sec. 1, additional ablation studies in Sec. 2, quantitative and qualitative results of semantically-rich floorplans in Sec. 3, and more visualized comparison results in Sec. 4.

1 Room-wise Encoder Details

This section discusses the details of the room-wise encoder. We utilize a DETR-based [1] transformer architecture to process a single floorplan image input and output several room feature codes, each corresponding to a latent representation of a room. Figure 1 illustrates the architecture of the room-wise encoder, which is divided into three modules: CNN backbone, transformer encoder, and transformer decoder. Given an input image, we first utilize a CNN backbone (ResNet50 [3]) to extract L layers of multi-scale feature maps $\{x_l\}_{l=1}^{L}$, where each layer's feature map has c channels. Subsequently, the multi-scale feature maps are fed into the transformer encoder to generate enhanced feature maps $\{\hat{x}_l\}_{l=1}^{L}$ with the same resolutions as the inputs. The entire transformer encoder is composed of multiple encoder layers, each of which includes a multi-scale deformable self-attention module (borrowed from Deformable DETR [6]) and a feed-forward network. Then, we input m learnable embeddings $F \in \mathbb{R}^{m\times q}$ into the transformer decoder. These embeddings adaptively extract local room features from the global image features output by the transformer encoder, such that each output embedding corresponds to a latent feature representation of a room. For each learnable embedding $f \in \mathbb{R}^{q}$, we discovered that using a single code with channel c to represent each embedding, i.e., $f \in \mathbb{R}^{c}$, is insufficient for accurate reconstruction. Instead, inspired from RoomFormer [4], we employ a stack of d codes to represent each learnable embedding $f \in \mathbb{R}^{d\times c}$. The entire transformer decoder is divided into 7 decoder layers. In the first 6 layers of the decoder layers, we perform self-attention on all local-level codes regardless of the room they belong to. In the cross-attention module, the learnable local-level codes extract different regions of image feature output from the transformer encoder. In the final decoder layer, we additionally introduce room-wise attention, which restricts the local-level codes to attend only to codes within the same room.

* Corresponding author

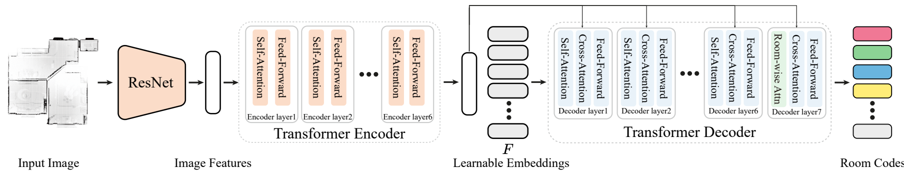

Fig. 1: The room-wise encoder architecture.

Table 1: Additional ablation studies.

Settings	Room		Corner		Angle	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
Default	**99.5**	**98.7**	**90.8**	**84.9**	**89.6**	**84.3**
Using general lines	98.3	97.3	85.8	84.4	78.8	77.5
Using sigmoid function	97.8	93.2	87.8	70.8	86.9	70.2

The local-level codes that belong to the same room are concatenated to produce the respective output room feature codes. The output room feature codes not only capture global image information but also aggregate local information from their corresponding rooms, which is sufficient for the final decoding.

2 Additional Ablation Studies

We provide additional ablation studies on two aspects: (1) Using general lines instead of separate lines, and (2) Using the sigmoid function instead of the loss terms defined in Eq. (7) and (11).

As shown in the 2$^{\text{nd}}$ row of Table 1, directly predicting the diagonal lines is less effective. Our objective is to optimize parameters for all the lines. While diagonal lines offer a more general representation, directly predicting them can lead to an extensive solution space. This complexity can impede the model's ability to accurately identify horizontal and vertical lines, which are predominant in most architectural layouts. To address this, we've adopted a two-phase prediction: first focusing on horizontal and vertical lines to establish a robust initial structure, then introducing diagonal lines for angular features. This methodical approach has yielded substantial performance benefits.

As shown in the 3$^{\text{rd}}$ row of Table 1, using the sigmoid function decreases overall metrics, indicating that the loss terms defined in Eq. (7) and (11) are more effective for optimizing the binary matrix.

3 Semantically-Rich Floorplans

Our method can easily be extended to semantically-rich floorplans, where we input the room-level features into a simple linear layer to predict the room label

Table 2: Semantically-rich floorplan reconstruction scores on Structured3D test set.

Methods	Room*		Room		Corner		Angle	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
RoomFormer [4]	71.9	70.9	94.0	92.8	84.2	80.0	75.6	71.9
Ours	**75.1**	**74.4**	**98.5**	**97.6**	**88.3**	**84.2**	**87.1**	**83.1**

probabilities. The quantitative result on the semantically-rich floorplan is shown in Table 2. Our model still outperforms RoomFormer after considering the room categories. The qualitative results are shown in Figure 2.

4 More Visualization Results

We provide more comparison results with HEAT [2] and RoomFormer [4] on Structured3D [5] in Figure 3 and 4.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 1
2. Chen, J., Qian, Y., Furukawa, Y.: Heat: Holistic edge attention transformer for structured reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3866–3875 (2022) 3
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1
4. Yue, Y., Kontogianni, T., Schindler, K., Engelmann, F.: Connecting the dots: Floorplan reconstruction using two-level queries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 845–854 (2023) 1, 3
5. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 519–535. Springer (2020) 3, 5, 6
6. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020) 1

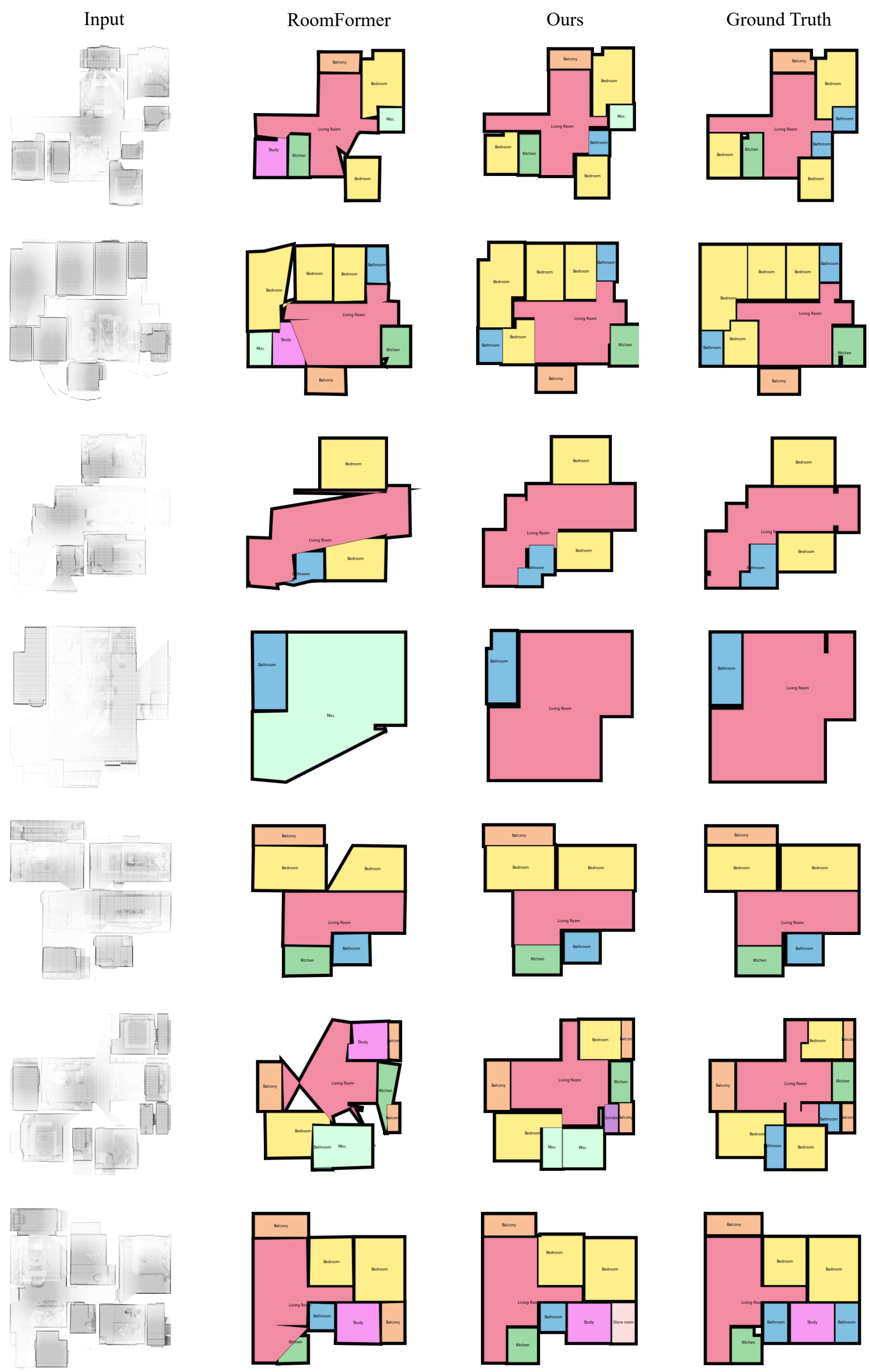

Fig. 2: Qualitative results on semantically-rich floorplans.

Fig. 3: More qualitative results on Structured3D [5].

Fig. 4: More qualitative results on Structured3D [5].