MSD: A Dataset for Floor Plan Generation of Building Complexes

Thanks to all reviewers, R1, R2, R3.

We are excited they find our work valuable: **R1**: *unique dataset; comprehensive representation formats; real-world applicability.* **R2**: *dataset is carefully cleaned; dataset is significantly more complex and diverse than prior work.* **R3**: *authors provide great details in data collections; proposed dataset build floor plans with higher complexity.*

Benchmark comparisons

R1: *[paper] lacks a broader comparison against a wider range of existing methodologies or more diverse baseline models ... restricts the understanding of MSD's comprehensive applicability.*

Agreed. We also ran and evaluated FLNet [2] and House-GAN++ [1]. Both required re-purposing to make them applicable to the tasks we set. We will add the results in Table. 1 below to the paper. These additional methods do not perform well, demonstrating the need for our proposed dataset with more realistic building complexes. This confirms the value of our work.

	MIoU (↑)	Compatibility (↑)	Topology (↑)	Proportions (↑)
FLNet (**new**)	19.3	n.a.	n.a.	n.a.
H-GAN++ (**new**)	11.6	64.2	n.a.	n.a.
MHD	21.8	76.2	0.461 ± 0.138	0.514 ± 0.143
U-Net	42.4	n.a.	0.439 ± 0.148	0.371 ± 0.171

Table 1. **HouseGAN++**: 128 x 128 masks, 388k steps, learn. rates: 1e-5 generator, 4e-5 discriminator, structural masks as input. **FLNet**: 128 x 128 masks, 50 epochs. User studies are done for MHD and U-Net: 7 architects, each 50 random IDs. **Topology**: whether the organization of the spaces makes sense. **Proportions**: whether the room proportions makes sense. Scoring: {"yes": 1, "unsure": 0.5, "no": 0}.

Cross-regional generalization

R1: *Paper does not address the potential challenges in generalizing the findings across different regions with varying architectural. norms and styles*

Definitely. This is exactly what our paper is about. We are actively extending the current dataset, to dwellings from other regions in Europe. We completely agree that we do not address the full diversity, because this process is slow due to copyright and privacy issues. In our work, we take an active step towards more diversity, where we will increase variations one step at a time.

Model performance

R2: *MIoU ... low for ... MHD model compared to UN. Not clear about the reason.*

This is an interesting point. We would like to argue that the discrepancy in the performance (in MIoU) might stem from the different losses. Clearly, the loss of UN (cross-entropy at pixel level) is closely aligned with evaluating on MIoU. However, the loss and evaluation are not necessarily as closely aligned in the case of MHD. MHD is a diffusion model, which through a series of T time steps denoises corner points $C^t_{i,j}$ (t: time step, i: room, j: j-th corner in room i): from a randomly initialized set of corner points $C^T_{i,j}$ into a reasonable composition of corner points $C^0_{i,j}$, which (when discretized) is taken as the composition of the final floor layout. The objective is similar to other diffusion models: each iteration, you randomly select t and learn a mapping for the reverse noise for that time step, which is parameterized by a neural network as $e_\theta\left(C_{i,j}, t\right)$. The corner points that come out $C^t_{i,j}$ are compared to the ground truth using the L2-norm (regression). Hence, the neural network $e_\theta\left(\bullet, \bullet\right)$ learns to effectively denoise corner points *for a given time step*. This is not necessarily the same as learning a mapping from input (structure and graph) to a *fixed* output (floor plan layout), which could for a part explain the discrepancy in performance. Whether indeed the different objectives explain the differences in performance in MIoU should, however, be more rigorously investigated. A nice direction for future work. We added these thoughts in the paper.

Evaluating complexity

R2: *How to evaluate the complexity performance?*

Yes, this is an unsolved research question. To better evaluate the complexity, qualitative evaluation (besides the important instrumental measures) will play an essential role. We are actively researching the evaluation methods for topologically more complex floorplans, and some preliminary results of our study are shown in Table. 1 (right) which we will include in the paper. Nonetheless, we believe that *both* quantitative as well as qualitative measures play an important role.

Complex vs. simple

R3: *Unknown if methods trained with more complex multi-apartment layout generation will help ... on simpler single-apartment layout generation tasks, and [vice versa].*

Indeed, both simple and complex floor plan designs are relevant. In consultation with architectural firms in northern Europe we strongly believe that our proposed dataset more closely aligns with the realistic scale of residential architectural projects.

References

[1] Nauata, N., Hosseini, S., Chang, K.H., Chu, H., Cheng, C.Y., Furukawa, Y.: Housegan++: Generative adversarial layout refinement network towards intelligent computational agent for professional architects. CVPR (2021) 1

[2] Upadhyay, A., Dubey, A., Arora, V., Kuriakose, S.M., Agarawal, S.: Flnet: Graph constrained floor layout generation. ICMEW (2022) 1