Floor plan generation

The interplay among data, machine, and designer

Mostafavi, Fatemeh; van Engelenburg, Casper; Khademi, Seyran; Vrachliotis, Georg

DOI
10.1177/14780771241290649

Publication date
2024

Document Version
Final published version

Published in
International Journal of Architectural Computing

Citation (APA)
Mostafavi, F., van Engelenburg, C., Khademi, S., & Vrachliotis, G. (2024). Floor plan generation: The interplay among data, machine, and designer. *International Journal of Architectural Computing*, *23*(2), 370-386. https://doi.org/10.1177/14780771241290649

IJAC

International Journal of
Architectural Computing
2024, Vol. 0(0) 1–17
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/14780771241290649
journals.sagepub.com/home/jac
Sage

Floor plan generation: The interplay among data, machine, and designer

Fatemeh Mostafavi, Casper van Engelenburg, Seyran Khademi and Georg Vrachliotis

Abstract

Recent advancements in machine learning (ML) in architectural design led to new developments in automated generation of floor plans. However, critical evaluation of ML-based generated floor plans has not progressed proportionally due to the subjectivity and complexity of the assessment, particularly for large and more complex floor plans. Accordingly, a hybrid (quantitative and qualitative) floor plan evaluation scheme is introduced in this study, focusing on multiple architectural aspects. To verify the effectiveness of the proposed framework, the evaluation scheme is applied on the generated floor plans resulting from two baseline computer vision models. The models have been trained on a newly introduced large-scale floor plan dataset called Modified Swiss Dwellings (MSD). The results showed that despite the progression of computer vision models for floor plan generation, they still have difficulty capturing the more complex architectural qualities. In addition, the prospect of floor plan generation and evaluation and possible future developments are discussed.

Keywords

Architectural dataset, computer vision, floor plan generation, floor plan evaluation, human-machine interaction

Introduction

Population expansion and climate change have been the main driving forces to address the challenge of adequate housing aligned with sustainable development goals.[1] However, the building design process is iterative, time-consuming, and costly.[2] Despite specifications concerning climatic and contextual differences, facilitating this process would therefore make a significant contribution. Detailed technical drawings, such as floor plans are to be prepared in the design development stage to illustrate more refined aspects of the design.[3] A floor plan is arguably the most used visual representation by architects, and serves as a compact blueprint of

Delft University of Technology, Architecture and the Built Environment Faculty, Delft, Netherlands

Corresponding author:
Fatemeh Mostafavi, Delft University of Technology, Architecture and the Built Environment Faculty, Julianalaan 134, 2628 BL Delft, Netherlands.
Email: f.mostafavi@tudelft.nl

a building, effectively communicating the arrangement and spaces, structural elements, and openings. In addition, floor plans convey the architectural qualities of a building through the function of the spaces, and connection to the outside.

While computational design methods formalize the input features and some of the objectives through parameterized functions to cast the problem as an optimization problem,[4] the architectural qualities are often linked to complex functions or unknown and high dimensional parameters that cannot be expressed in closed formulations. One approach to formulating the generative process is learning from real-world examples where architectural intelligence is manifested through real design examples. In recent years, data-driven approaches - most prominently deep learning models - have shown to be capable of extracting useful information from complex and high-dimensional data.[5] Hence, such approaches are particularly interesting for the design of floor plans, which is a complex generative task. Nevertheless, formalizing the evaluation metrics to validate the competence of the deep learning model is rather challenging and the focus of this article.

The advancement of artificial intelligence technologies has made it possible for computers to display human-like capabilities for comprehending, interpreting, and generating unique solutions.[6] The human-machine collaboration has been investigated in many areas, from transferring human visual knowledge to robot vision in disassembly tasks[7] to autonomic architectures for smart buildings.[8] The way machines assist architects in designing floor plans harkens back to the onset of digital culture in architecture. The idea of automatically generating and evaluating floor plans was established by the conceptual foundation built upon the “Flatwriter: Choice by Computer” project.[9] The fundamental concept beyond the Flatwriter not only laid the foundation for machine-aided generative design, but also tapped into the concept of human-machine interaction and participatory design.[10] The notion of using intelligent machines for generating and evaluating floor plans has been a persistent theme in data-supported architectural design. Currently, similar questions are being explored, but under different computational conditions and with more sophisticated digital tools.

In automated floor plan generation, the format of the generated output is a crucial factor to ensure further flexibility and editing possibilities. Compared to raster images, vector graphic formats (e.g., dwg, svg, etc.) offer more possibilities of integration into the computer-aided architecture design (CAAD) procedure,[11] better performance at global reasoning,[12] and no post-processing requirement.[13] Moreover, due to the inherent characteristics of architectural technical drawings such as collinearity, orthogonality, and corner sharing between adjacent spaces, the task of generating floor plans is inherently different from the generation of natural images or languages.[11,12] As a result, developing intelligent machines for floor plan generation is challenging, especially when machines’ and designers’ inputs are both considered in the design process.

In this study, the interplay among *machine*, *data*, and *designer* is introduced (Figure 1) and investigated through a floor plan generation task framework. As the three main pillars of this framework, each step containing elements of machine, data, and designer is explained in detail. Accordingly, the Modified Swiss Dwelling (MSD) dataset, as a benchmark for training and testing computer models for the task of floor plan generation is introduced – a floor plan datasets of high-quality buildings. Two baseline models learned from MSD are introduced. Most importantly, the evaluation of AI-generated floorplans, as one of the open problems in the design and development of generative models, is critically addressed by proposing a novel assessment scheme. The significance of human expert evaluation by comparing the results of quantitative and qualitative analysis of the generated floor plans is highlighted, emphasizing the role of the designer in the building evaluation and design process. The objective of the research concerns the evaluation and capabilities of the state-of-the-art models in floor plan generation as well as a thorough examination of the effectiveness of the proposed hybrid evaluation scheme.

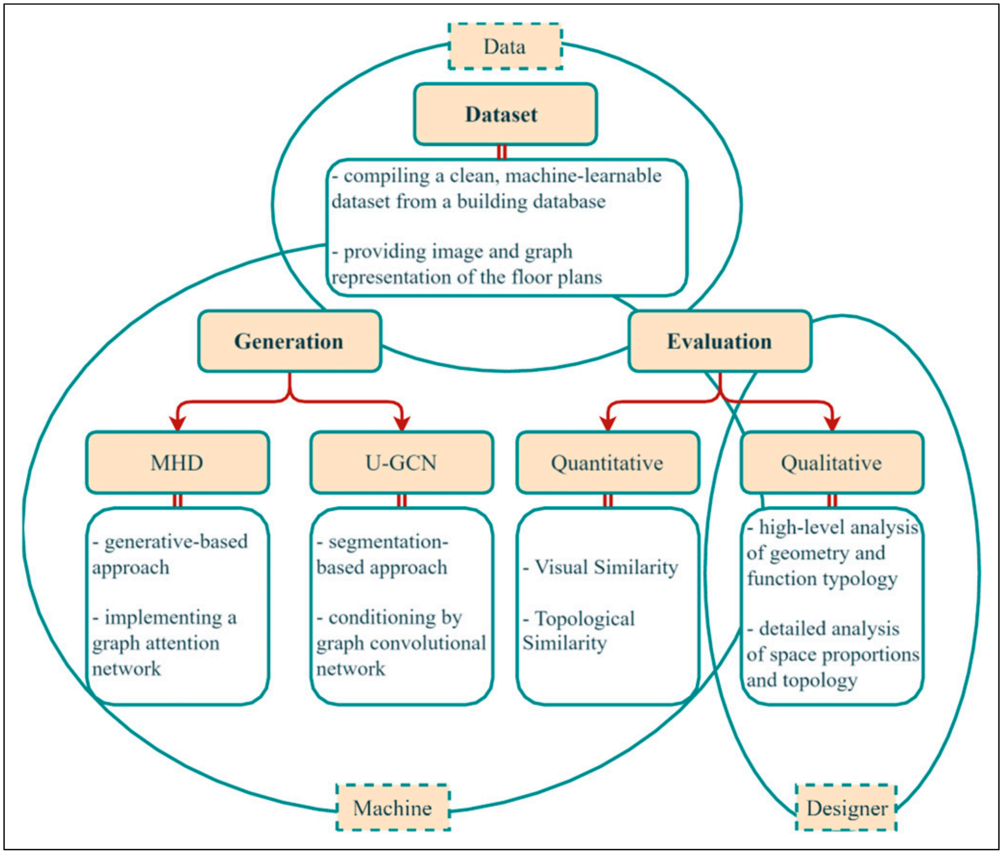
Data
Dataset
- compiling a clean, machine-learnable dataset from a building database
- providing image and graph representation of the floor plans
Generation
Evaluation
MHD
U-GCN
Quantitative
Qualitative
- generative-based approach
- implementing a graph attention network
- segmentation-based approach
- conditioning by graph convolutional network
- Visual Similarity
- Topological Similarity
- high-level analysis of geometry and function typology
- detailed analysis of space proportions and topology
Machine
Designer

Figure 1. The framework of the interplay among data, machine, and human in a floor plan generation task.

Related work

The related works are organized based on the three bubbles: data, machine, and designer (Figure 1). Specifically, the following is discussed: the publicly available floor plan datasets; the involvement of machines in the floor plan generation task; the involvement of the designer in the evaluation process; and their interrelationships.

Floor plan datasets

Data plays a critical role in relation to both machine and designer. Data affects what and to what extent generative models can learn the task and determine, depending on the format, which evaluation strategies can be used. Architectural design datasets come in a range of complexity, annotation style, and quantity, tailored to suit different purposes.[14] For instance, ROBIN[15] was intentionally designed to serve as a plan retrieval study, and HouseExpo[16] was drafted for indoor layout learning. As the affordances that the employed dataset can offer have a significant impact on the problem formulation and complexity of the experiments on floor plans, the choice of a suitable dataset is critical. More specifically on the floor plan generation studies, RPLAN dataset has been relatively widely used.[11,13,17–22] Despite the availability and high quantity, RPLAN still lacks some real-world design-related characteristics, such as the limitation of floor layouts to single apartment units (as opposed to multi-unit apartments) and lack of data about the elevation and furniture.

As technical architectural drawings convey not only geometrical but also topological and semantic information, a comprehensive analysis of floor plans requires sufficient data on these three levels.[14] More recently, studies have been conducted on a dataset presenting comprehensive geometrical, environmental, and infrastructural data of 45k apartments in Switzerland, called Swiss Dwellings (SD).[23] Although previous research investigated the importance of feature selection process in predicting room labels,[24] and microclimate context visualization of floor plan instances in SD,[25] no generation task has been experimented with using this dataset.

Floor plan generation

Machines often play intermediate role between data and designer, linking the curated data and designers' critical insights through floor plan generation and evaluation. The idea of automated floor layout generation has attracted considerable attention in recent decades, even before the data-driven methods started flourishing.[26,27] Since the advent of deep learning (DL),[28] learning-based methods for generating floorplans have become a de facto approach.[11–13,17–22,29] The input to these generative models have taken different formats, addressing either of geometrical constraints and topological requirements or both of them. Accordingly, the learning-based floor plan generation studies can be categorized into three groups depending on the input data modality: building boundary, graph, and combined building boundary and graph.

Building boundary as input. Floor plan generation workflows that use building boundaries as fixed input data are more similar to real-world design circumstances in which the building must be designed on a specific site. In the RPLAN study,[17] an encoder-decoder network was trained to predict the position of internal walls within a fixed building boundary given as input and predicted room types. A post-processing step was employed to turn the pixel-level wall predictions into the vectorized representation. Later in the WallPlan model,[13] the front door location was also considered as the fixed input besides the building boundary. After initializing the building boundary with windows prediction, a graph generation and semantics generation networks were joined to predict the internal walls as well as the room types.

Graph as input. A shift in the input type from the building boundary to the graph, representing the bubble diagram in the architectural design workflow, was made with the purpose of the control over the spatial relationships between spaces. In the HouseGAN model,[12] a generative adversarial network (GAN) was trained, taking the rooms' type, number, and spatial adjacency gathered in a single input graph. With the defined architectural constraints, the model resulted a set of axis-aligned bounding boxes of rooms as possible solutions. Integrating the same network strategy with a conditional GAN, the HouseGAN++ model[19] was trained to convert the input bubble diagrams to the segmentation masks of rooms and doors. With similar layout connectivity representation as input graph, the HouseDiffusion model[20] was later introduced to directly generate a vector floor plan by predicting the coordinates of rooms and doors using denoising a diffusion process. In another study using the connection graph as the input, the GTGAN model[29] used a graph transformer GAN to control the room relations as graph nodes. As a result of the mentioned studies, using the graph as the input to the floor plan generation pipeline would maintain the topological requirements, while overlooking the geometrical (e.g., area or building boundary) constraints.

Building boundary and graph as inputs. Considering both building boundary and graph as the input to the floor plan generation pipeline would benefit both geometrical and topological design constraints. The Graph2Plan model[18] was devised to take the building boundary and user constraints in the form of room numbers, locations, and adjacencies, as the input in the floor plan generation framework. In this pipeline, graph layouts were retrieved from a dataset and then adjusted in the given input layout, leading to a set of bounding boxes for rooms and a floor plan raster (i.e., pixel-based) image. In another study introducing the FLNet model,[22] user inputs were set in the form of building boundary, room types, and spatial relationships as graphs. The space layout resulting from the embedded vectors of the input graph were aligned inside the given boundary, resulting in a raster floor layout output. Although the two mentioned studies differed in the sequence of applying boundary and graph constraints in the floor plan generation workflow, both resulted in floor plans satisfying topological and geometrical requirements.

Floor plan evaluation

As the last element of the presented framework, evaluation of generated floor plans influences the way the performance of computer vision models is measured and the extent to which the designer's insights are incorporated. In general, two types of evaluation methods have been implemented in the related studies. The first category incorporated quantitative analysis of the generated floor plans such as diversity, compatibility, and mean positional error (MPE), whereas the second category corresponded to qualitative analysis (also referred to as user study).

Quantitative approach in floor plan evaluation. Different quantifiable metrics have been either employed from other research domains or particularly devised for the floor plan analysis task. One of the earliest methods was calculating the actual distance from the predicted location of generated internal walls in the floor plan to those in the ground truth, expressed in meters.[17] This expression is limited to the applications where the correspondence of each pixel in the floor plan image to the real measurements is known. Other than measuring the distance between corresponding elements, diversity of the generated floor plans were also regarded as a quantifiable metric to measure the creativity of the trained computer vision models. Diversity has been assessed through calculating the Fréchet inception distance (FID) metric,[11,12,19,20,29] comparing the distribution of generated floor plan images with the distribution of a set of real images known as ground truth. Moreover, in the studies in which floor plans have been represented as graphs, a metric called graph edit distance[30] was employed to assess the topological aspects of the generated layouts. This metric quantifies the compatibility between the input bubble diagram and the one reconstructed from the generated floor plan.[12,19,20,29] Based on both image and graph distances, a metric called SSIG was later introduced for evaluating the structural similarity of floor plans.[31] SSIG was coined to address the lack of structural awareness of pixel-based metrics (e.g., Intersection-over-Union), as well as practical difficulties of pairwise graph matching approaches.

Qualitative approach in floor plan evaluation. Besides the abovementioned quantitative analysis, some qualitative assessments were also employed in form of user studies to integrate the expert insights in floor plan generation process. The plausibility of the generated floor plans against the real ones was assessed by a group of participants in the earlier studies,[18] also including a vigilance test to verify the results from experiment.[17] With the same idea, the *Realism* metric was defined in a way that a generated floor layout is subjectively compared against a ground truth by human participants in a user study.[12] The Realism, also referred to as average user rating, was later followed in similar studies alongside the quantitative evaluations.[19,20,29] Formulated in other frameworks, the authenticity check of the generated floorplans in a pool of real ones designed by licensed architects and graduate students of architecture was also performed as a way of qualitatively assessing the layouts.[11]

The combination of domain-specific quantitative metrics and user studies would lead to an effective mixed-methodology assessment framework for the evaluation of floor plans. Despite the importance of incorporating expert knowledge in the floor plan generation workflows, the quantitative evaluation approach has received considerably more attention in previous studies. The reasons could stem from the relative ease of the calculation as well as the lack of access to expert evaluators. Moreover, all the studies in which the qualitative approach was followed, sought to verify the realism of the generated floor plans. However, comprehensive qualitative assessment of the architectural layouts goes beyond this assessment, also addressing topological and geometrical aspects.

Current study scope

Benchmark datasets for the training and evaluation of generative methods are crucial for the development of d deep learning models. In this study, the focus is on the evaluation of the generated floor plans from two computer vision models that have been trained on the MSD dataset. Specifically, attention is given to the human evaluation beyond merely comparing the results with the dataset samples. Accordingly, the novelties of this study are as follows:

- Applying two computer vision models trained on the newly introduced MSD dataset for the floor plan generation task.
- Proposing a hybrid evaluation scheme of AI-generated floor plans
- Highlighting the role of the designer in the process of intelligent floor plan generation and evaluation

The remainder of the papers is divided into four main subjects: The method and materials including the dataset, computer vision models, and the hybrid evaluation scheme of the generated floor plans in Section three, the quantitative and qualitative results of the floor plan evaluation process in Section 4, discussion and conclusion including the interpretation of the results, limitations of the current study, and the further developments in Sections 5 and 6, respectively.

Method and materials

Based on the framework presented in Figure 1, the structure of this section also contains three main aspects, each corresponding to one of the “data”, “machine”, and “designer” bubbles. Accordingly, the development of a novel floor plan dataset, the specifications of two computer vision models trained on the introduced dataset, and the details of the hybrid floor plan evaluation scheme are presented.

Modified Swiss dwelling (MSD) dataset

MSD[32] originates from the SD database[23] – an extensive collection of building layouts in numerical format. The geometrical data in SD is contained in a DataFrame in which each row is an architectural detail that describes a stand-alone space (e.g., living room, corridor) or element (wall segment, window, furniture). The columns define the related data of each architectural entity, for instance, “geometry” as a polygon represented by WKT format, “entity type” categorizing the type of the entity such as “feature” or “area”, and an identifier (ID), which associates the geometrical detail to the belonged site, building, plan, floor, apartment, and unit. The , SD dataset was further processed into MSD by visualizing the plans and curating a well-balanced ML-ready dataset. The dataset focuses on medium- to large-scale residential buildings, excluding mixed-use floor plans, and minimizing similar data instances. The cleaning process includes Removing features such as kitchen sinks, and bathtubs

- Filtering out non-residential floor layouts by detecting certain room types
- Sampling one-floor level per building to avoid duplicate data instances
- Keeping more complex layouts by eliminating layouts with few areas

Ultimately, the number of floor plans in MSD equalled 5.372. Each floor plan is defined as an image in which the pixel value indicates the room type. In addition, the corresponding access graph is attributed to each floor plan. The nodes in the graph are defined by room type, zone type, centroid, and polygon. The edges in the graph indicate access connectivity and are either “door”, “front door”, and “passage”.

Computer vision models

Even though the focus of the paper is the evaluation of the generated floorplans, we provide the technical details of the baseline computer vision models to be able to correlate the results with the approach. As an output of the floor plan generation competition on the MSD dataset,[33] the two top approaches were selected for evaluation in our study. The inputs are the structural elements of the floor plan (i.e., external walls and interior load-bearing walls) and an access graph representing the connection of grouped areas (i.e., zones). The zones are defined based on environmental and architectural requirements of interior residential spaces as follows: (1) zone 1: bedrooms (2) zone 2: living room, kitchen, dining room, etc., (3) zone 3 storerooms, bathrooms, and toilets, and (4) zone 4: balconies. The learning process is tuned to turn the inputs into a complete floor plan. The two baselines are referred to as: (1) segmentation-based, and (2) generative-based models.

U-GCN.

In the segmentation-based approach,[34] a U-Net[35] was used to encode the building structure into a compressed feature vector representation and subsequently decode the representation into the rasterized floor plan. To condition the U-Net on the zoning graph, a graph convolutional network (GCN)[36] was used, encoding the zoning graph into a feature vector. The feature vector was concatenated to the latent representation of the U-Net; together used as input to the decoder part of the U-Net. The U-Net and GCN were updated simultaneously throughout training. In addition, the pre-trained *Segment Anything model*[37] was used in this approach to predict the interior from the exterior. The model architecture and training details are as follows:

- The encoder of the U-Net consisted of four down-sampling convolutional layers, each doubling the channel dimensions (64 → 128 → 256 → 512). The layers comprised 3 × 3 learnable convolutional, batch normalization, ReLU activation, and 2 × 2 Maxpool, respectively. The decoder had the same structure as the encoder; however, used up-sampling convolutional layers.
- The GCN was a stack of two graph convolutional layers, each with a 256-sized hidden node feature dimension. Global mean pooling on the output node features was used to compute the graph-level feature vector.
- The multi-class cross-entropy loss on the pixel predictions, and Adam optimizer with an initial learning rate of 0.001, and batch size of 16 were employed.

Modified HouseDiffusion.

In the generative-based approach[38] the Modified HouseDiffusion (MHD) model was introduced. Instead of denoising pixels, as is done in conventional diffusion models, MHD denoises 2D coordinates of the areas. The MHD’s architecture is based on learning the relations between the corner points of each room polygon. To condition the model on the building structure, an extra attention module between all corner points of the polygons and corner points in the building structure was added. Also, to associate the correct room type given the zoning type, a graph attention network (GAT)[39] was trained separately. The generated floor plans resulting from the two baseline models were further investigated through a hybrid evaluation scheme. Model and training details are as follows:

- The GAT model consisted of five consecutive graph attention layers. Initial node features were one-hot encodings of the zoning type, and edge features depending on the connectivity (i.e., door or a wall). The hidden node feature dimension was set to 64, and a ReLU was used between consecutive layers. The output node feature dimension was equal to the number of room types. The cross-entropy loss between the output node features and the ground truth (a one-hot encoding of the room types) was employed. Moreover, Adam were used with an initial learning rate of 0.001, batch size of 128, using dropout (0.2) for regularization, and applying early stopping.
- The MHD model was built on *HouseDiffusion*,[20] yet altered based on conditioning on the building structure. Specifically, the attention layer in the transformer model of *HouseDiffusion* was modified by adding a cross attention between corners and structural wall segments. The wall segments were extracted from the binary image representing the building structure by subsequently (1) using a morphological thinning technique to reduce the thickness of each wall in the binary image to a bare minimum (1 pixel) and (2) converting the “thinned” binary image into a skeleton network from which the set of wall segments (lines) are to be extracted. In addition, the relational cross attention (RCA) – a module in the original *HouseDiffusion* model to discern between different connectivity types – was modified in which diverse types of connectivity (i.e., door, passage, or front door) were each assigned a unique and learned embedding. The same hyperparameters as in *HouseDiffusion* were used, except for the batch size of 32 and training steps of 300k in current experiment.

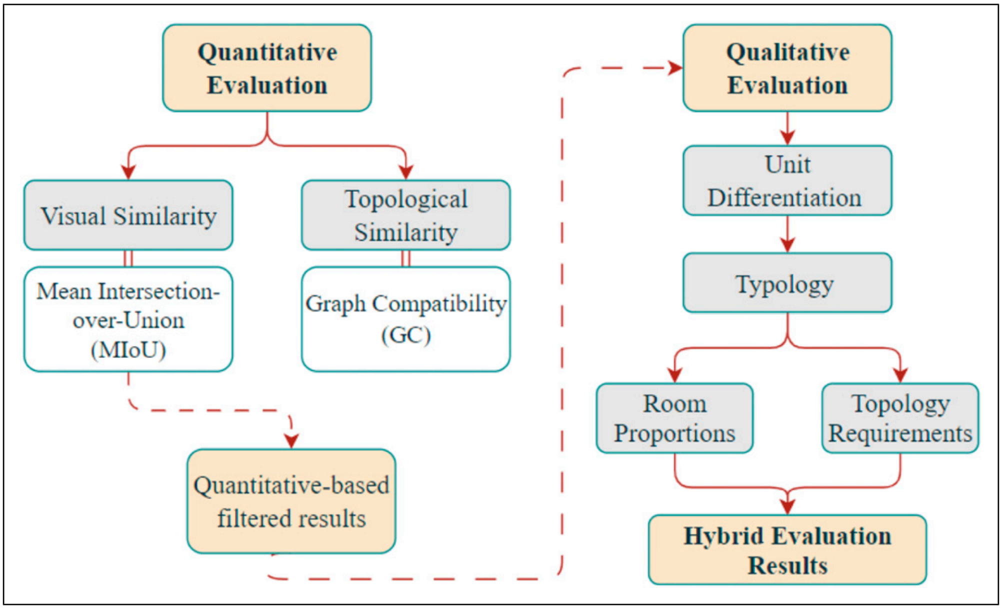

Figure 2. The proposed hybrid evaluation scheme for generated floor plans assessment.

Hybrid evaluation scheme of generated floor plans

Given the necessity of keeping the designer in the loop of the design process, this study proposes a hybrid evaluation scheme for the assessment of generated floor plans (Figure 2). The scheme is divided into two main steps, and the process is done sequentially. To test the trained generative models, a sample of the generated floor plans first undergo a quantitative evaluation, in which the visual and topological similarities between the test data and the ground truth (i.e., samples of training dataset) are measured. Afterward, a fraction of the top results (i.e., the ones passing the quality threshold) are filtered to go through the qualitative evaluation, in which the features that are more efficiently captured by human expert vision are to be assessed. Eventually, the final top results are achieved. The details of the two mentioned steps and their sub-steps are explained in the following sub-sections.

Quantitative evaluation. The generated floor plans of the two models were quantitatively evaluated both at the image pixel level and at the graph level. At the pixel level, the mean Intersection-over-Union (MIoU) is used as a similarity measure between the generated and ground truth floor plans. MIoU stems from a simpler metric called IoU, which measures the pixel-wise intersection divided by the union (i.e., combined) pixel areas of a floor plan pair (the predicted one and the corresponding ground truth). One step further using MIoU, the overlap between two-floor plans is measured as the average per space. This metric would capture the features related to the amount of pixels in different regions of the floorplan, which is architecturally translated to the area of each room. The MIoU values can range from 0 to 1, with higher values showing higher overlapping of the predicted and the ground truth floor plans.

At the graph level, the consistency between the predicted and ground truth graphs is checked. For comparing the predicted and ground truth graphs, the adjacency graphs from the predicted geometries were first extracted. The graph compatibility is then measured by counting the edges in the ground truth room graph that are retained in the predicted adjacency graph, divided by the number of edges in the ground truth room graph. Using the graph compatibility metric, topological qualities of a floor plan would be measured. The range of this metric is similar to MIoU, which values closer to one show higher compatibility of topological aspects between the predicted and the ground truth floor plans.

In the quantitative evaluation phase, 800 floor layout samples were initially selected as test set. To ensure a logical distribution of both simple and complex floor plans, the test set was divided into five subsets based on the spaces counts. Afterwards, 10 samples were randomly selected from the subsets that passed the IoU threshold specific to each subset. The IoU thresholds for the subsets with 15–19, 20–29, 30–39, 40–49, and 50+ area counts were set to 0.35, 0.35, 0.30, 0.25, and 0.22, respectively. As a result of this filtering process, 50 floor plans were qualified for the qualitative evaluation step to be further investigated by a group of knowledgeable users with architecture background.

Qualitative evaluation. In the qualitative evaluation phase, the filtered data instances from the quantitative evaluation step went through a user study by 14 participants with architectural background. A survey was arranged including 102 image data instances, comprising 50 floor plans of each model plus two ground truth for vigilance test purpose. The participants were asked to evaluate the given floor plans based on step-by-step approach (Figure 3). As a result of implementing the vigilance test, the survey results of the participants who did not correctly answer the stated questions for the ground truth floor plans were disregarded. This allowed for more accurate interpretation of survey results of higher quality. Consequently, 10 evaluations were qualified to be further processed.

More specifically, the qualitative evaluation of the generated floor plans followed a high-level to detailed approach. Each step contained three possible options to be chosen for the related questions. First, it was checked whether different units on a building floor level are distinguishable enough. The definition of the term

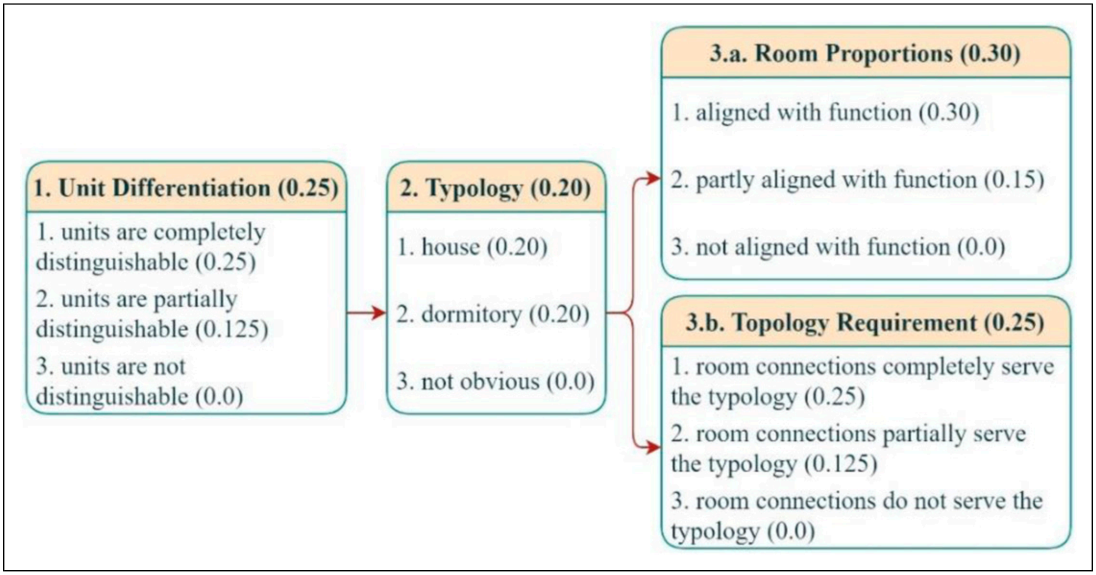
1. Unit Differentiation (0.25)
1. units are completely distinguishable (0.25)
2. units are partially distinguishable (0.125)
3. units are not distinguishable (0.0)
2. Typology (0.20)
1. house (0.20)
2. dormitory (0.20)
3. not obvious (0.0)
3.a. Room Proportions (0.30)
1. aligned with function (0.30)
2. partly aligned with function (0.15)
3. not aligned with function (0.0)
3.b. Topology Requirement (0.25)
1. room connections completely serve the typology (0.25)
2. room connections partially serve the typology (0.125)
3. room connections do not serve the typology (0.0)

Figure 3. The qualitative evaluation steps, scoring system for each step, and answer options.

"unit" in this study is a whole dwelling arrangement containing all the required spaces. Second, the typology of the dwelling was assessed by considering the number, or presence/absence of certain room types. After the first two evaluation steps, participants could choose to continue further with the assessment of either one of the room proportions and topology requirements, or both. The qualitative evaluation process and the scoring system are illustrated in Figure 3. The cumulative scores of all steps equal to 1, each of which is linearly divided among answer options. Between the two parallel steps of room proportions and topology requirement, the former was assigned a higher weight since the computer vision models were initially conditioned by the zoning graph. The results of the proposed hybrid evaluation scheme are reported in the following section.

Results

Following the steps of the evaluation scheme presented in the previous section, the results are reported in two main sub-categories. Firstly, the quantitative results applied on 800 sampled floor plans of two computer vision models are reported. Subsequently, the qualitative results are explained on the 100 filtered floor plans resulting from the previous step.

Quantitative analysis

The quantitative results of the two trained models are presented in Table 1. The MIoU metric is reported based on the floor plan size ranges (i.e., number of rooms), as well as the average among all categories. Regarding the MIoU metric, the U-GCN model outperformed the MHD model both on average and in all floor plan categories. Although the U-GCN model resulted in higher MIoU in the second-floor plan size category (i.e., 20–29 rooms), a general trend of decreasing MIoU in both models can be noticed as the number of rooms increases. The Graph Compatibility metric is only reported for the MHD model since extracting a graph from the rasterized results of the U-GCN model is prone to errors. In the MHD model, the results do not follow a general trend; however, the best results correspond to the more complicated floor plans with more than 50 rooms.

Qualitative analysis

The qualitative results for the 100 filtered floor plans are brought in Table 2. The mean score over 50 samples of each model, as well as the average among all the users, are reported for both models. It can be deduced that the MHD model outperformed the U-GCN model in all categories, with an average final results difference of

Table 1. Quantitative results on floor plan generation for U-GCN and MHD.

Model	Average	15 − 19	20 − 29	30 − 39	40 − 49	50+
	MIoU					
U-GCN	**0.382**	0.398	0.406	0.353	0.328	0.296
MHD	**0.218**	0.235	0.220	0.210	0.201	0.179
	Graph compatibility					
MHD	0.871	0.859	0.873	0.880	0.875	0.886

Table 2. User study results of evaluating 100-floor plans based on different architectural criteria (The values correspond to the mean scores per user in different criteria as well as in total.).

Model	User	Unit differentiation (out of 0.25)	Typology (out of 0.2)	Room proportions (out of 0.3)	Topology (out of 0.25)	Final (out of 1)	Cohen's Kappa
U-GCN	1	0.107	0.096	0.114	0.085	0.402	0.0
	2	0.125	0.140	0.042	0.060	0.367	0.0
	3	0.145	0.180	0.135	0.136	0.657	−0.20
	4	0.150	0.168	0.081	0.115	0.504	0.50
	5	0.072	0.132	0.069	0.072	0.346	0.0
	6	0.115	0.184	0.057	0.082	0.438	−0.14
	7	0.200	0.140	0.168	0.132	0.640	−0.09
	8	0.215	0.188	0.231	0.180	0.814	0.0
	9	0.157	0.124	0.153	0.145	0.583	0.27
	10	0.192	0.164	0.141	0.180	0.677	−0.14
	Average	0.148	0.152	0.119	0.117	**0.537**	-
MHD	1	0.250	0.124	0.195	0.130	0.699	0.63
	2	0.110	0.156	0.108	0.062	0.436	0.0
	3	0.167	0.184	0.209	0.136	0.725	0.0
	4	0.152	0.164	0.183	0.160	0.659	0.20
	5	0.105	0.176	0.114	0.112	0.507	−0.33
	6	0.125	0.192	0.102	0.087	0.506	−0.33
	7	0.140	0.144	0.216	0.110	0.610	0.33
	8	0.200	0.172	0.267	0.215	0.854	0.0
	9	0.172	0.096	0.249	0.197	0.715	0.33
	10	0.162	0.164	0.141	0.127	0.595	0.0
	Average	0.158	0.157	0.178	0.134	**0.627**	-

16.7%. The lowest difference (3.2%) between the models can be seen in the typology criterion, whereas the highest difference corresponds to the room proportion criterion (49.5%), highlighting the strength of the MHD model to represent the geometrical characteristics more accurately. Figure 4 provides example generations of each model. In the case of MHD, by including WCA, the generated building layouts more closely follow the building structure. The generations of U-GCN follow the building structure closely, but the often-noisy pixel maps clutter the readability of the building layouts – which, for MHD, is not the case as MHD, by design, produces shapes.

To measure the agreement among users, Cohen's Kappa coefficient[40] was calculated for each user compared to the average responses of all the users in each model. The coefficient was calculated pairwise, containing two sequences corresponding to each of the four qualitative evaluation criteria. The negative values of this coefficient demonstrate poor agreement, while the 0–0.20 range shows slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–80 substantial, and 0.81–1.0 almost perfect agreement. Based on the distribution of data, three categories were defined based on 45% and 65% percentile of data at each criterion. Accordingly, the responses of each user were assessed in unit differentiation, typology, room proportions, and topology criteria against the corresponding values for the average responses. As it is shown in Table 2, only 20% of the

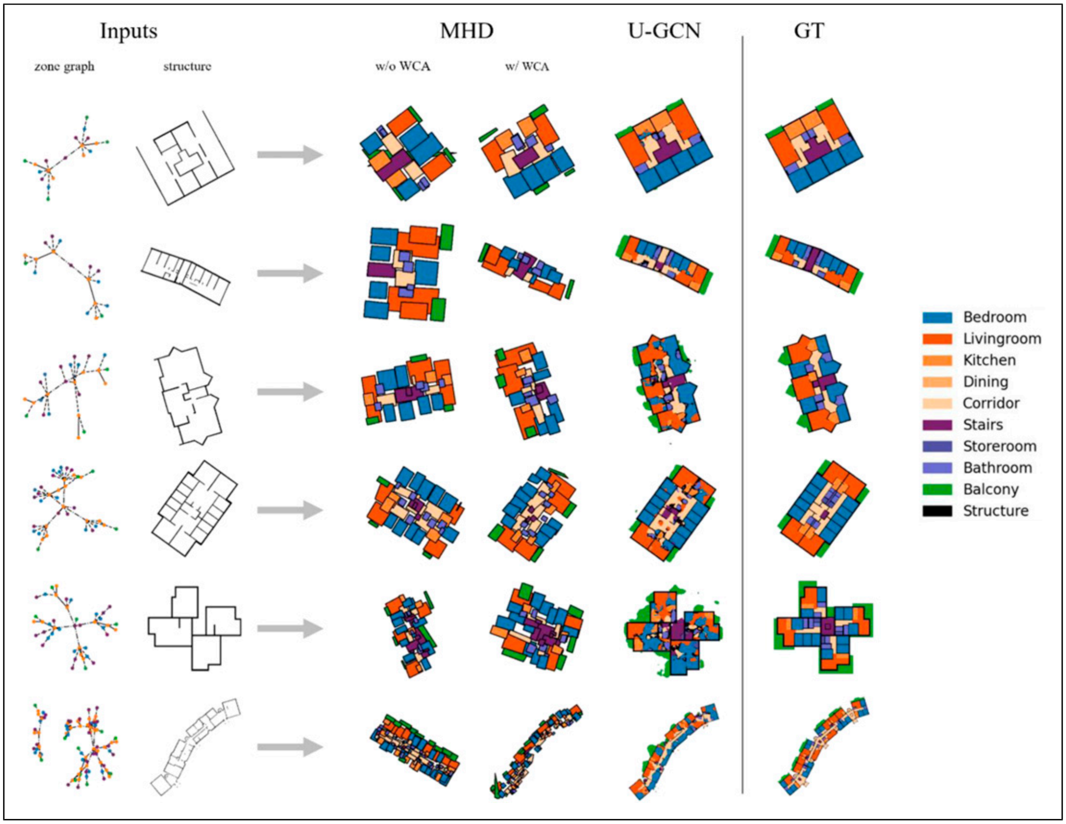

Figure 4. Generations of MHD and U-GCN. The inputs (zone graph and building structure) are provided in the left two columns; the generations of MHD (with and without WCA) are given in the next two columns; those of the U-GCN in the fourth column; and the ground truth pixel maps in the final column. Examples are given from small (top) to large (bottom) building layouts, which was measured by the number of rooms. Colours represents the room types.

users showed fair to moderate agreement for the U-GCN model, whereas 40% agreed on the four defined criteria at the level of fair to substantial for the MHD model.

The values from the two models' results are plotted as boxplots in Figure 5, specifically showing the distribution of the user-study scores, with the size of the box indicating the spread of the distribution. The highest mean score for the U-GCN model corresponds to the typology criterion, whereas MHD gained the higher mean score in the room proportions aspect. Both models performed relatively poorly in representing proper topology, which indicates the drawbacks of the models in satisfying more complicated architectural qualities. The most distinguishable difference between the two models can be seen in the proportion assessment criteria, which demonstrates the superiority of the geometry-based model in showing measurement qualities of spaces in the pixel-based one.

Discussion

Comparing the overall performance of the two computer vision models trained on the MSD dataset for the task of floor plan generation, the results can be interpreted on two different levels. Firstly, the models can be assessed based on the gained scores at different stages of the proposed hybrid evaluation scheme, which would give insights regarding the robustness of different generative models in the floor plan generation task. Secondly, the role of designer in this process can be investigated, focusing on the qualitative analysis study. Accordingly, helpful guidelines can be extracted to further improve the human-machine interaction in the building design process. The details of the two mentioned interpretations are as follows:

The performance of models

For the MHD model, although the MIoU scores are the least for the category of floor plans with 50 rooms, the graph compatibility scores are the highest. Regardless of the model, this shows the importance of evaluating

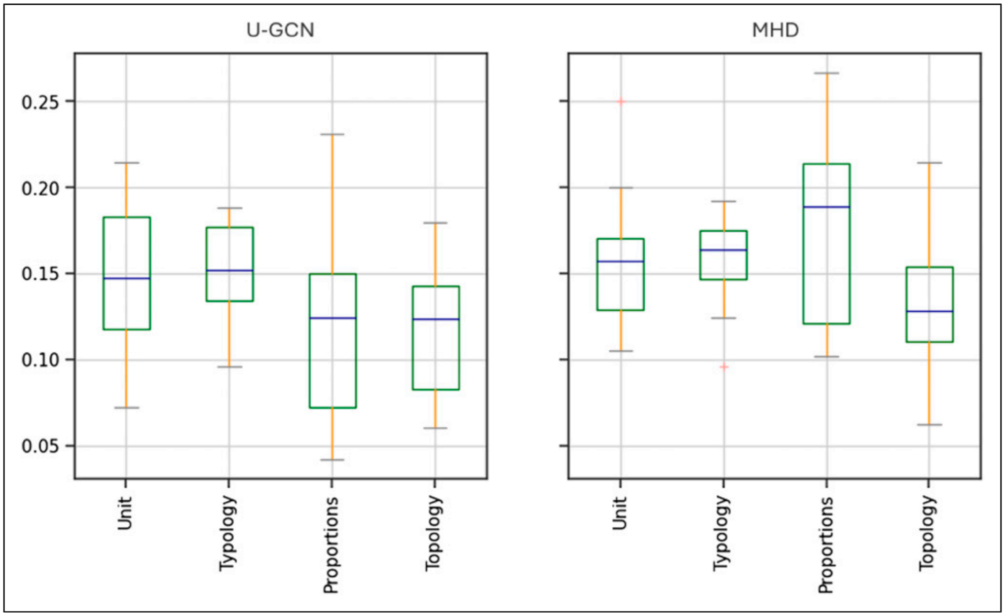

Figure 5. Qualitative results on floor plan generation for the U-GCN and MHD models.

the generated floor plans based on different geometrical and topological criteria, such that diverse architectural qualities could be covered in the assessment. Moreover, implementing the hybrid evaluation scheme in this study allowed different steps of the evaluation framework to be compared. Although the U-GCN model outperformed in quantitative evaluation step, the MHD model showed higher scores in qualitative assessment, in addition to the capability of being assessed by graph-based metrics. Conducting the qualitative evaluation also revealed the extent to which the generated floor plans could be assessed. In the sense that participants were given the option to further continue the assessment based on either room proportions, typology requirement, or both. Nonetheless, all the participants chose to assess all the instances based on both of the parallel steps. Furthermore, based on the results from the qualitative study, the AI-generated plans perform on average half as the ground truth (Figure 6). More specifically, the ML-based models showed more precise performance in capturing high-level attributes of the floor plans, namely unit differentiation and typology. Therefore the progress should be made in the future models to also be robust at extracting more detailed features such as room proportions and topology.

The effectiveness of the evaluation scheme

Besides the comparison of two presented computer vision models by the proposed quantitative and qualitative metrics, this study also shed light on the performance of the evaluation scheme itself. Compared to the evaluation approaches followed in previous works, in which the impact of expert evaluators were limited to assess the realism (i.e., whether the generated floor plans look like real case) of the outputs,[11,12,19,20,29] the current study benefited from step-by-step high-level to detailed evaluation approach, addressing multiple architectural aspects. Moreover, the agreement assessment results among the users in this study proved that the task of floor plan evaluation is highly subjective (Figure 6, right). Although the number of participants in the qualitative assessment part was limited, the main objective of performing such assessment in this study was to explore the role of the designer in the process of generative floor layout design. It was shown that expert knowledge is a determining factor in assessing AI-generated floor plans. The role of designer can be interpreted and further improved by two approaches. In the sense that the designer can either enter the evaluation process of the floor plans later in more detailed steps, or the models can be fine-tuned based on the designer's input in an interactive setting such as the Reinforcement Learning approach.[41,42]

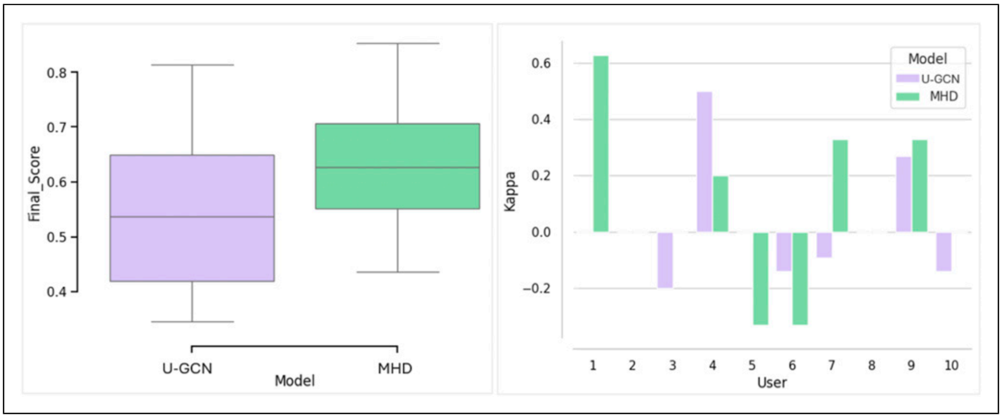

Figure 6. The overall comparison of U-GCN and MHD models (left final qualitative evaluation score, right Kappa coefficient values for each user).

Conclusion

Due to population expansion and climate change, demand for providing adequate housing has raised. Given that the housing design process is inherently iterative, time-consuming, and costly, facilitating the design procedure would of high impact. In this regard, automated floor plan generation task has benefited from recent technologies in two different methodological ways. While computational methods seek to parametrize the design space, the parametrization is learning-based approaches using computer vision models has been shifted to neural networks' parameters. In other words, implementing AI-based techniques into the building design process can solve problems that algorithmic approaches show limitations to tackle. Implementing either of these approaches, attention must be paid to the climatic and contextual specifications of the intended building to be designed or assessed.

In this study, a task of floor plan generation using a novel dataset was defined, and accordingly, two computer vision models were trained and tested. A hybrid evaluation scheme, composed of quantitative and qualitative analysis was proposed and later applied to the trained models. Based on the results, both the competence and the gap to develop more robust models in future iterations were concluded. It was shown that despite the advancements in computer vision models in task of floor plan generation, they still struggle capturing the architectural qualities which can be assessed by expert knowledge. Moreover, the role of the designer in the evaluation of AI-generated floor plans highlighted the need for reconsidering the evaluation and design pipelines within this realm to adapt to the new technological advancements. Consequently, the task of floor plan evaluation is concluded to be non-trivial, calling for unified evaluation systems, metrics, and scoring. Furthermore, employing a coherent evaluation scheme makes the comparison of different studies on architectural datasets more feasible.

The limitations of the current study and therefore the potential for further expanding this line of research fall under the number of expert users and the amount of floor plan being assessed by them. The evaluation scheme could benefit from more collective inputs of the evaluators and hence reducing the possible sources of bias. Moreover, this study focused on assessing geometrical and topological qualities of generated floor plans; however, broader analysis can be conducted based on other architectural qualities such as the environmental demands, particularly context-specific architectural characteristics and regional guidelines. Further improvements of the current study can be envisioned in the following directions:

- Devising new evaluative metrics tailored to floor plan architectural, technical, and performative assessment
- Fine tuning the current generative models towards enhancing the more detailed architectural qualities such as room proportions and topology requirements
- Integrating the designer's input in the AI-driven building design loop side by side of data and the machine

The research data including the test floor plans for the qualitative evaluation phase and their corresponding ground truths can be accessed upon request.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Fatemeh Mostafavi https://orcid.org/0000-0002-8047-2168

References

1. United Nations Human Settlements Programme. *Priorities 2022-2023: adequate housing, cities and climate change and localising the sustainable development goals*. UN-Habitat. https://unhabitat.org/priorities-2022-2023-adequate-housing-cities-and-climate-change-and-localising-the-sustainable (2024, accessed 21 June 2024).
2. Bahrehmand A, Batard T, Marques R, et al. Optimizing layout using spatial quality metrics and user preferences. *Graph Models* 2017; 93: 25–38.
3. AIA ETN. Design to construction. https://www.aiaetn.org/find-an-architect/design-to-construction/ (2022).
4. Arora JS. Computational design optimization: a review and future directions. *Struct Saf* 1990; 7: 131–148.
5. Goodfellow I, Bengio Y and Courville A. *Deep learning*. MIT Press, 2016.
6. Topuz B and Çakici Alp N. Machine learning in architecture. *Autom ConStruct* 2023; 154: 105012.
7. Deng W, Liu Q, Zhao F, et al. Learning by doing: a dual-loop implementation architecture of deep active learning and human-machine collaboration for smart robot vision. *Robot Comput Integrated Manuf* 2024; 86: 102673.
8. Genkin M and McArthur JJ. B-SMART: a reference architecture for artificially intelligent autonomic smart buildings. *Eng Appl Artif Intell* 2023; 121: 106063.
9. Vrachliotis G. *Architecture and design in the age of cybernetics*. Berlin, Boston, MA: Birkhäuser, 2022. DOI: 10.1515/9783035624816.
10. Friedman Y. *Toward a scientific architecture*. Cambridge, MA: MIT Press, 1980.
11. Luo Z and Huang W. FloorplanGAN: vector residential floorplan adversarial generation. *Autom ConStruct* 2022; 142: 104470.
12. Nauata N, Chang K, Cheng C, et al. *House-GAN: Relational Generative Adversarial Networks for Graph-constrained House Layout Generation*. CVPR, 2020.
13. Sun J, Wu W, Zhang G, et al. WallPlan: synthesizing floorplans by learning to generate wall graphs. *ACM Trans Graph*; 41: 1–14. DOI: 10.1145/3528223.3530135.
14. Pizarro PN, Hitschfeld N, Sipiran I, et al. Automatic floor plan analysis and recognition. *Autom ConStruct* 2022; 140: 104348.
15. Sharma D, Gupta N, Chattopadhyay C, et al. DANIEL: a deep architecture for automatic analysis and retrieval of building floor plans. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* 2017; 1: 420–425.
16. Li T, Ho D, Li C, et al. HouseExpo: a large-scale 2D indoor layout dataset for learning-based algorithms on mobile robots. In: IEEE International Conference on Intelligent Robots and Systems, Las Vegas, NV, 24 October 2020–24 January 2021: 5839–5846.
17. Wu W, Fu XM, Tang R, et al. Data-driven interior plan generation for residential buildings. *ACM Trans Graph*; 38: 1–12. DOI: 10.1145/3355089.3356556.
18. Hu R, Huang Z, Tang Y, et al. Graph2Plan: Learning Floorplan Generation from Layout Graphs. *ACM Trans Graph*; 39. Epub ahead of print 27 April 2020. DOI: 10.1145/3386569.3392391.
19. Nauata N, Hosseini S, Chang K-H, et al. House-GAN ++: generative adversarial layout refinement networks. *CVPR*. 2021. DOI: 10.48550/arXiv.2103.02574.
20. Shabani MA, Hosseini S and Furukawa Y. HouseDiffusion: vector floorplan generation via a diffusion model with discrete and continuous denoising. https://aminshabani.github.io/housediffusion (2022, accessed 16 October 2023).
21. Hosseini S, Shabani MA, Irandoust S, et al. PuzzleFusion: unleashing the power of diffusion models for spatial puzzle solving. https://www.magicplan.app/ (2023, accessed 17 October 2023).

22. Upadhyay A, Dubey A, Arora V, et al. FLNet: graph constrained floor layout generation. In: ICMEW 2022 - IEEE international conference on multimedia and expo workshops 2022, proceedings, Taipei City, Taiwan, 18–22 July 2022. DOI: 10.1109/ICMEW56448.2022.9859350.
23. Standfest M, Franzen M, Schröder Y, et al. Swiss Dwellings: a large dataset of apartment models including aggregated geolocation-based simulation results covering viewshed, natural light traffic noise, centrality and geometric analysis. *zenodo*. 2023. DOI: 10.5281/zenodo.7788422.
24. Bielik M, Zhang L and Schneider S. Big data, good data, and residential floor plans: feature selection for maximizing the information value and minimizing redundancy in residential floor plan data sets. *Computer-Aided architectural design. INTERCONNECTIONS: Co-computing beyond boundaries*. Cham: Springer Nature Switzerland, 2023, Vol. 1819, pp. 607–622.
25. Mostafavi F and Khademi S. Micro-Climate building context visualization a pipeline for generating buildings' environmental context maps using numerical simulation data. In: 41st Conference on education and research in computer aided architectural design in Europe, eCAADe, 2023, Graz, Austria, 20–22 September 2023, pp. 9–18.
26. Merrell P, Schkufza E and Koltun V. Computer-generated residential building layouts. *ACM Trans Graph* 2010; 29: 181. DOI: 10.1145/1866158.1866203.
27. Peng C-H, Yang Y-L and Wonka P. Computing layouts with deformable templates. *ACM Trans Graph* 2014; 33: 1–11. DOI: 10.1145/2601097.2601164.
28. Szeliski R. *Computer vision, algorithms and applications*. London: Springer London, 2011. DOI: 10.1007/978-1-84882-935-0.
29. Tang H, Zhang Z, Shi H, et al. Graph transformer GANs for graph-constrained house generation 2023. https://arxiv.org/abs/2303.08225
30. Abu-Aisheh Z, Raveaux R, Ramel J-Y, et al. An exact graph edit distance algorithm for solving pattern recognition problems. *4th International Conference on Pattern Recognition Applications and Methods*. 2015. DOI: 10.5220/0005209202710278ï.
31. Van Engelenburg CCJ, Khademi S and Van Gemert JC. *SSIG: a visually-guided graph edit distance for floor plan similarity*. IEEE Xplore, 2023.
32. Van Engelenburg Casper, Mostafavi Fatemeh, Kuhn Emanuel, et al. MSD: A Benchmark Dataset for Floor Plan Generation of Building Complexes. *arxiv*. 2024. DOI :10.48550/arXiv.2407.10121.
33. CVAAD. cvaad-workshop/iccv23-challenge. https://github.com/cvaad-workshop/iccv23-challenge (2023).
34. Jeon Y, Tran DQ and Park S. Skip-connected neural networks with layout graphs for floor plan auto-generation. https://arxiv.org/abs/2309.13881v2 (2023, accessed 6 December 2023).
35. Weng W and Zhu X. U-Net: convolutional networks for biomedical image segmentation. *IEEE Access* 2015; 9: 16591–16603.
36. Kipf TN and Welling M. Semi-supervised classification with graph convolutional networks. In: 5th international conference on learning representations, ICLR 2017 - conference track proceedings. https://arxiv.org/abs/1609.02907v4 (2016, accessed 13 December 2023).
37. Kirillov A, Mintun E, Ravi N, et al. Segment anything. https://arxiv.org/abs/2304.02643v1 (2023, accessed 5 December 2023).
38. Kuhn E. Adapting housediffusion for conditional floor plan generation on modified Swiss dwellings dataset. https://arxiv.org/abs/2312.03938v1 (2023, accessed 11 December 2023).
39. Veličković P, Casanova A, Liò P, et al. Graph Attention Networks. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. Epub ahead of print 30 October 2017. DOI: 10.1007/978-3-031-01587-8_7.
40. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
41. Black K, Janner M, Du Y, et al. Training diffusion models with reinforcement learning. https://rl-diffusion.github.io (accessed 12 January 2024).
42. Brown N, Garland A, Fadel G, et al. Deep reinforcement learning for engineering design through topology optimization of elementally discretized design domains. https://www.aaai.org (2022, accessed 3 January 2024).