Computational Visual Media
https://doi.org/10.1007/s41095-0xx-xxxx-x

Research/Review Article

Fuzzy-based Indoor Scene Modeling with Differentiated Examples

Qiang Fu[1], Shuhan He[1], Hongbo Fu[2], Xueming Li[1], and (✉)Zhigang Deng[3]

Abstract Well-designed indoor scenes contain interior design knowledge, which has been an essential prior for most of indoor scene modeling methods. However, the layout qualities of indoor scene datasets are often uneven, while most of existing data-driven methods do not differentiate indoor scene examples in terms of their qualities. In this work, we aim to explore an approach that leverages datasets with differentiated indoor scene examples for indoor scene modeling. Our solution is to conduct subjective evaluations on lightweight datasets that have various room configurations and furniture layouts, via pairwise comparisons based on the fuzzy set theory. We also develop a system to use such examples to guide indoor scene modeling according to user-specified objects. Specifically, we focus on object groups associated with certain human activities, and define room features to encode the relations between the position/direction of an object group and the room configuration. Given an empty room, our system first assesses it in terms of the user-specified object groups, and then places the associated objects in the room guided by the assessment results thus completing indoor scene modeling. A series of experimental results and comparisons with the state-of-the-art indoor scene synthesis methods are presented to validate the usefulness and effectiveness of our approach.

Keywords Indoor scene modeling, modeling by examples, fuzzy measurement, membership degree

1 Introduction

The problem of indoor scene modeling has been extensively studied in the past decades. From the early guideline-based approaches [1, 2] to example-based approaches [3, 4], as well as the latest activity-centric methods [5–9] and deep learning models [10, 11], their capability advances steadily in generating visually pleasing and functionally-valid 3D indoor scenes that benefit many applications including games and interior design.

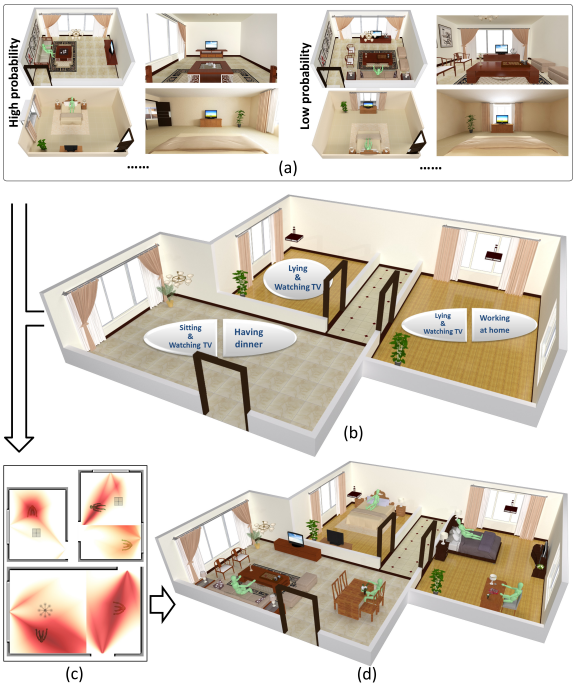


Fig. 1 The workflow of our method, includes the datasets (a), input room and activity labels (b), pre-room assessment results (c), and the synthesized scene based on the assessment (d).

To obtain plausible indoor layouts and object arrangements, most of existing indoor scene modeling approaches rely on either expert-designed guidelines or examples. For data-driven methods, large-scale datasets of indoor scenes could improve the quality of the synthesized scenes as a result of abundant examples. However, the layout qualities of the indoor scene examples in large-scale datasets may not be at the same level. Due to the lacking of metrics to evaluate the quality of indoor scenes based on the associated functionalities, low-quality

1 School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications.
2 School of Creative Media, City University of Hong Kong.
3 Department of Computer Science, University of Houston, TX, USA. E-mail: zdeng4@central.uh.edu.

Manuscript received: 2022-01-01; accepted: 2022-01-01

examples have the same weights as high-quality ones for indoor scene modeling in most existing methods. Intuitively, indoor scenes with different qualities, called *differentiated examples* in this work, should play different roles in indoor scene modeling, i.e., the impacts of high-quality examples should be enhanced while the others should be weakened.

To address the above issues, methods that exploit differentiated examples for indoor scene modeling need to be investigated. We observe that the layouts of high-quality indoor scenes typically well support their assumed functionalities. Even for rooms with specially-designed layouts, their furniture layouts can still have some common relations to the room configurations including room size and shape, positions of windows, positions of doors, etc. For example, since a TV set is rarely placed in front of a window, the layout of an object group with a TV set, a couch, and a tea table could be influenced by the window positions in a room. These observations motivate us to exploit the common layout relations as the metrics to differentiate examples in the datasets. Besides handling a variety of room layouts, the evaluation metrics need to be also associated with object functionalities. Therefore, examples in an ideal interior design dataset should have the following: i) they can be classified into functionality-associated object groups; ii) they include differentiated layout examples and the associated evaluations; and iii) they include sufficient layout variations to support generality and robustness.

In this paper, we propose a new method that uses datasets with differentiated samples as priors for room assessment, and then further use the assessment results to generate indoor scenes. The collected differentiated samples have various room layouts with respect to certain object groups. Since quantitative analysis on indoor scenes is challenging, we leverage fuzzy measures and subjective comparisons to evaluate the layout quality of the differentiated samples. Specifically, we adopt the *membership degree*, a concept borrowed from the fuzzy set theory [12, 13], to evaluate the samples in the dataset after we conduct pair-wise comparisons on the samples. For example, in Figure 1, our method collects differentiated examples of indoor scenes to facilitate indoor scene modeling (a). Given input rooms and the assigned activity labels representing certain object groups (b), our method uses differentiated examples, which have been labeled with evaluation scores during the subjective evaluation, to conduct the per-room assessment (c). Specifically, we first calculate the weighted feature distances of different room features between the input room and dataset scenes. Then the given room can be assessed via transferring the membership degrees of the differentiated examples based on their room feature distances, with respect to a certain object group. Since the assessment is performed for all positions in the given room with four different directions of the object group, we can place the object group into the room based on the assessment results thus synthesizing 3D scenes with plausible indoor layouts (d). Moreover, we provide an ease-of-use tool to assist users to design indoor scenes. It also allows users to merge multiple rooms into a larger and more complex scene.

In sum, our work makes two novel contributions: i) a novel metric to assess indoor scenes through differentiated examples in a dataset, based on the fuzzy set theory, and ii) a framework to model indoor scenes based on the room assessment with respect to certain groups of objects. We demonstrate the advantages of our method for indoor scene synthesis through various experiments, as well as direct comparisons with state-of-the-art, data-driven indoor scene synthesis methods [8, 10, 14].

2 RELATED WORK

Many systems and approaches for indoor scene modeling have been proposed in the past decades. The first task is to understand and describe contextual scenes and their hierarchical structures. For example, data-driven methods, which encode semantic scene structures from existing indoor scene examples, have been well studied in recent years (e.g., [15–17]). The co-existence and hierarchical relations of indoor objects have often been used to describe indoor scene contexts, e.g., Xu et al. [18] proposed to cluster a set of co-existing object groups, called focal points, in order to organize a collection of heterogeneous indoor scenes. Liu et al. [19] proposed to use probabilistic grammars for hierarchical decomposition of a scene into semantic components. Zhang et al. [20] proposed to learn discrete priors to accurately represent exact layout patterns, by measuring the strengths of spatial relations of indoor objects based on tests for complete spatial randomness (CSR). Moreover, some works also leverage action or even natural language to establish the object relations of indoor scenes (e.g., [21, 22]). In recent years, deep learning techniques have been successfully adopted for contextual scene understanding. For example, Li et al. [14] presented *GRAINS*, which encodes information about objects' spatial properties, semantics, and their relative positioning with respect to other objects in a hierarchy using a variational recursive autoencoder (RvNN-VAE), trained on a dataset of annotated scene hierarchies. The analyzed scene context information benefits indoor scene modeling and can be used as constraints to determine object categories and locations inside

a synthesized scene[23–25]. Such priors of object relations can also be used in interactive indoor scene modeling systems (e.g., [26]). In our work, for simplicity, the relationships within a group of objects for a certain activity are pre-defined, so that we can focus on how to place the objects into the given room in terms of their associated activity.

On the other hand, how to evaluate the quality of indoor scenes is a challenging yet widely-open problem. Analyzing the effect of indoor environmental factors on subjective human perception is a long-standing topic in both architecture and environmental psychology. In general, some major environmental factors, including illumination, air quality, temperature, noise, and space, are utilized to measure the quality of an indoor environment [27]. Researchers have revealed that indoor environment can impact the comfort and cognitive performance of human beings, and there exist acceptable ranges to keep people comfortable [28]. Thus, a task to explore proper environmental factor ranges is then raised for indoor scene design. For example, Konis [29] provided a system to predict the visual comfort of indoor scene core zones, based on high dynamic range images that capture the indoor illumination. Ochoa and Capeluto [30] proposed a similar analysis on indoor illumination with simulated indoor environments to evaluate visual comfort. In our work, we extract expert knowledge from datasets of differentiated indoor scene examples. The evaluations on the differentiated examples are through subjective comparisons, and we use the evaluation results, i.e., the fuzzy membership degrees, as the assessment scores to label the indoor scenes in our datasets. These examples are used to assess input rooms for placing certain object groups.

Based on various scene representations, a large number of indoor scene synthesis methods have been proposed. Most of these works rely on pre-defined guidelines or relations learned from 3D scene datasets (e.g., [1, 2]). To increase the efficiency of indoor scene synthesis, some works adopt example-driven methods to transfer interior design styles from existing indoor scenes [4], or indoor images [3] to a given room. Human-centric approaches provide another way to make indoor scene synthesis more automated. Jiang et al. [5] proposed to use human context for object arrangement by learning how objects relate to human poses. Fisher et al. [6] proposed to generate 3D scenes given noisy and incomplete 3D scans, by arranging objects based on certain activities. Savva et al. [7] proposed to learn a probabilistic model to connect human poses and the arrangement of object geometry, for jointly generating 3D scenes and interaction poses. These works motivate us to gather certain activity-related objects into groups, and place such a group into the given room as a whole. More specifically, our work relies on methods such as [7] to determine the relevant object positions/directions in a group (e.g., a group of the couch, tea table, and TV set). Our method focuses on the next task, i.e., how to place such a group in a given room. Different from the human-centric methods (e.g., [5, 6]) that directly measure the probability of various activities on a certain region in a room, we leverage the subjective experiments and fuzzy metrics to evaluate the dataset indoor scene examples with respect to certain activities. We adopt a data-driven strategy that uses the dataset indoor scenes weighted by the fuzzy metrics to guide the scene synthesis.

Some recent works tackle large-scale interior design by utilizing deep neural networks for indoor scene synthesis. For example, Wang et al. [10] employ a deep convolutional neural network to learn priors from a large-scale indoor scene database for indoor scene synthesis. Zhang et al. [31] proposed a generative model using a feed-forward neural network that maps a prior distribution like normal distribution to the distribution of primary objects in indoor scenes. This work focused on the 3D object arrangement representation within a group of objects. Our work focuses more on the global layout of object groups in a given room, so the local arrangement for objects in a group can be pre-assigned. We also consider the relations between the layout of a certain object group and the room configuration, aiming to create scenes more suitable for performing certain human activities. To describe such relations, we define the room features in terms of the environment-related components like windows and doors. Moreover, comparing to these deep-learning-based methods, our method does not rely on a large-scale indoor scene dataset.

3 Data Preprocessing

In this section, we first introduce how to construct differentiated scene datasets and the room features we adopt, and then we give the details on how to label the scenes in the datasets through fuzzy-based subjective comparisons.

Scene data collection and representation. To verify the usability of differentiated examples as priors for indoor scene modeling, we collect lightweight datasets in which each scene example only has one object group, so that each type of indoor scene dataset is associated with a single group of objects. Namely, scenes in the same dataset have the same kind of object group. Considering the object groups are generally associated with certain activities, we choose the activity name as the object group label in the user interfaces of our system.

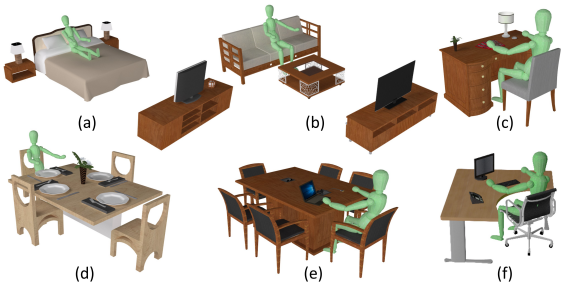


Fig. 2 Object groups and the associated agents with respect to various activities.

Fig. 3 Examples of dataset scenes with different room configurations or furniture layouts.

To establish the relations between room configuration and furniture layout, we first normalize the sizes of the objects, based on a human agent with a fixed body size. Then set the human agent on the object group to represent its front direction and position. Note that, for the given room, the user could specify multiple activity labels to generate an indoor scene with more than one object group. As a preliminary attempt, we only focus on six types of common object groups. As shown in Figure 2, each object group is associated with a certain activity including lying and watching TV (a), sitting and watching TV (b), working at home (c), having dinner (d), conferencing (e), and working in office (f).

For each type of scene dataset, e.g., Figure 3, we first choose a well-designed indoor scene (Top-Left) and then change its room size, the positions of windows/doors/artificial lights, or the positions/directions of the object groups to generate the other scene examples. This would lead to differentiated examples with various room configurations and layout quality differences. The room configuration variations ensure that the constructed datasets can be used to assess more kinds of indoor scenes, while the layout quality differences ensure that meaningful assessment results can be obtained from subjective comparisons. Note that we use the left-bottom corner of the floor as the origin of its associated coordinate system to encode the room size and the positions of windows, doors, artificial lights, and object groups. The directions of the object groups are limited to four directions (i.e., up, down, left, and right). We also limit the range of room sizes to

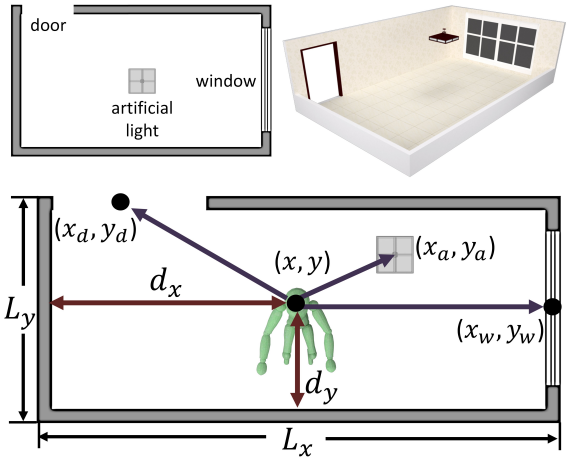


Fig. 4 **Top**: the 2D plan of an empty room (Left) and the corresponding 3D scene (Right). **Bottom**: room features and parameters which are related to the human agent (representing object groups).

avoid too small or too large rooms in the datasets. In our lightweight datasets, we totally have 48 scene examples in all 6 types with different layouts. On this basis, we also duplicate the well-designed scene example two or three times in each type, aiming at balancing the quantities between good- and poor-quality scene examples. Since these datasets are small, we can conduct subjective comparisons on them to label their assessment scores. We generate a snapshot for each scene example with the same view of the human agent in the scene for comparison.

We define the room features that focus on the relations between furniture layout and room components (i.e., wall/window/door), rather than only use the whole shape of the room. Specifically, we consider four types of room features to establish such relations. The features of windows and doors are associated with the relative directions, especially the angles between the direction of the object group and the vector from the object group to the windows and doors. This is mainly because such included angles generally determine the front view of the agent on the object group, and thus impacting the subjective perception of humans on the associated activity. Considering the symmetry, we use the sine-squared functions of the included angles as the features. The features of artificial lights and walls are associated with their relative distances to the object group. We directly use the Euclidean distance to represent the features of artificial lights; while for walls, we use the room size to normalize the distances between the object group and walls.

As illustrated in Figure 4-(Bottom), we define the room features based on the front direction and the position of an object group (where we set the agent). Here we first focus on the case of a room with a single window, a door, and a light (Figure 4-(Top)), and will discuss general cases later.

We assume that $\mathbf{d}$ denotes the front direction and (x, y) denotes the position of the object group, respectively. Let d_x and d_y be the distances from the object group to the front and right-side walls; L_x and L_y are the corresponding side lengths of a rectangular room (or the oriented bounding box of a non-rectangular room); (x_w, y_w), (x_d, y_d), and (x_a, y_a) denote the respective positions of the window, door, and artificial light. To describe the relations between the object group and room components including windows, doors, walls, and lights, the room features consist of the angles between the front direction of the object group and the object-to-window/-door direction that measures the relations to windows and doors, and the relative positions between the object group and walls and lights. Specifically, the sine-squared functions of the included angles (denoted as F_w and F_d, respectively) between $\mathbf{d}$ and the directions from (x, y) to (x_w, y_w) and from (x, y) to (x_d, y_d), the distance F_a from (x, y) to (x_a, y_a), and the relative position F_l of the object group in the room, are described as follows:

$$\begin{aligned} F_w(x, y, \mathbf{d}) &= 1 - \left(\frac{\mathbf{d} \cdot (x - x_w, y - y_w)}{||\mathbf{d}||_2 \cdot ||(x - x_w, y - y_w)||_2}\right)^2, \\ F_d(x, y, \mathbf{d}) &= 1 - \left(\frac{\mathbf{d} \cdot (x - x_d, y - y_d)}{||\mathbf{d}||_2 \cdot ||(x - x_d, y - y_d)||_2}\right)^2, \\ F_a(x, y) &= ||(x - x_a, y - y_a)||_2, \\ F_l(x, y, \mathbf{d}) &= \left(\frac{d_x}{L_x}, \frac{d_y}{L_y}\right). \end{aligned} \tag{1}$$

Fuzzy-based Subjective Comparisons. Due to the lack of quantitative metrics for layout quality evaluation, we aim to leverage subjective evaluations to discriminate the indoor scene examples in the datasets. Inspired by the fuzzy set theory [13], especially the analytic hierarchy process [12], we employ pairwise comparisons to label the assessment scores of the differentiated indoor scene examples by calculating their membership degrees. To reduce the impacts from individual biases, we recruited 32 participants to compare the randomly chosen scene pairs from our datasets. The participants were informed of the related object group for each type of scene, and asked to compare the snapshots of each scene pair by choosing the one they preferred. We collect such intuitive but fuzzy comparisons instead of accurate and professional evaluations due to two reasons: 1) such comparisons do not require professional interior designers thus making it easy to conduct; 2) the non-experienced users can still judge the quality of the indoor scene based on their perspective, even they might not know the hows and whys, that is mainly because the high-quality scenes will always make people feel comfortable in visual. In total, we collected 2,962 comparisons including sitting and watching TV (496), lying and watching TV (713), having dinners (372), working at home (544), conferencing (310), and working in office (527).

Let $\mathcal{S} = \{s_1, s_2, \cdots s_M\}$ be a set of scenes with the same type, we construct the pairwise comparison matrix G as follows:

$$G = \begin{bmatrix} p(s_1|s_1) & p(s_1|s_2) & \cdots & p(s_1|s_M) \\ p(s_2|s_1) & p(s_2|s_2) & \cdots & p(s_2|s_M) \\ p(s_3|s_1) & p(s_3|s_2) & \cdots & p(s_3|s_M) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}. \tag{2}$$

Different from [12], which uses *intensity of importance* (from 1 to 9) to construct the pairwise compassion matrix, the entries of the above matrix in our method are defined below: each entry $p(s_i|s_j)$ represents the degree of preference of s_i over s_j. Since the comparison in our implementation is *either-or*, we simply define the matrix entries using the following equation:

$$p(s_i|s_j) = \frac{p_{s_j}(s_i)}{p_{s_i}(s_j) + p_{s_j}(s_i)}, \forall s_i, s_j \in \mathcal{S}, \tag{3}$$

where $p_{s_j}(s_i)$ represents the count of the votes where the participants felt the scene s_i is better than the scene s_j. Since each participant only compared a portion of the dataset scene pairs to avoid fatigue, we use the weighted average of each row of G as the membership degree function. The weight is the squared root of the comparison frequency $t(i, j) = \sqrt{\frac{n_{i,j}}{N}}$, where N is the total number of the participants and $n_{i,j}$ is the number of comparisons for the pair of the scenes s_i and s_j. Note $t(i, j) = 0$ if scene pair (i, j) has not been compared (e.g., $t(i, i) = 0$). Mathematically, for each dataset scene s_i, the membership degree function on the scene quality is defined as:

$$M_C(s_i) = \frac{1}{T} \sum_{j=1,\cdots,M} G(i, j) \cdot t(i, j), \tag{4}$$

where $T = \sum_{j=1,\cdots,M} t(i, j)$. Based on this definition, we can calculate all membership degrees $M_C(s_i) \in [0, 1]$ for all dataset scenes. The better a scene, the larger its membership degree $M_C(s_i)$.

4 Indoor Scene Modeling

Given an empty room with specified activity labels that indicate the user-expected object groups, our method first assesses the given room by transferring the assessment scores of the dataset scene examples, and then places the object groups into the room guided by the assessment results to synthesize an indoor scene.

Room Assessment. The continuous variations of the room configurations form a space containing all possible scenes for a certain group of objects, denoted as domain $\mathcal{U}$, and the differentiated scene examples in our datasets can be

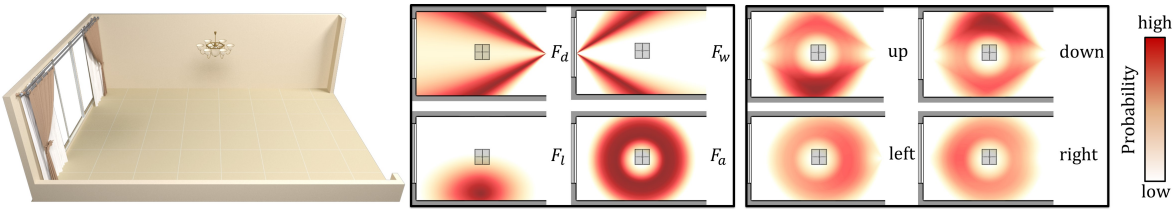


Fig. 5 **Left**: A room with the activity label of "sitting and watching TV". **Middle**: The assessment results with respect to the up direction of the associated object group about different room features (Equation (5)). **Right**: The compound assessment results with four different directions (Equation (6)). Note we conjointly normalize the four maps to reveal that the proper direction (up or down in this case) has the extremums of the probabilities.

considered as some sparsely sampled examples in $\mathcal{U}$. The mapping $A(\mathcal{U}) \rightarrow \mathcal{V}$ is to assess scenes in $\mathcal{U}$ and output $\mathcal{V} \in [0, 1]$. From the aforementioned user study result, we obtain a sparse set of examples, where $A(s_i) = M_C(s_i)$ based on the degree of membership functions for the scenes in our datasets $\{s_i\}$. Hence we propose to use them as bases to establish mapping A for all the scenes in $\mathcal{U}$.

Since the scene examples in our datasets are characterized through the room features, with respect to the position and direction of the object group, we uniformly sample positions in the given room with four directions to calculate a series of room features. Then, we use the basis set $\{A(s_i)\}$ of the assigned activity to get the assessment energy for all sampled positions of the given room to determine the placement of the object group. For the four types of room features $\{F_w, F_d, F_l, F_a\}$ in Equation (1), let $f_n(x, y, \mathbf{d})$ be the value of the n-th type of feature at the position (x, y) in the room with the direction $\mathbf{d}$ of the object group, and $\tilde{f}_n(s_k)$ be the value of the same type of feature calculated from a scene in our datasets s_k. Note that the feature F_a only depends on the positions. To reuse the assessment of our example scenes on the given room, we define the assessment energy at position (x, y) with the direction $\mathbf{d}$ as follows:

$$E_n(x, y, \mathbf{d}) = \sum_{k=1}^{K} (1 - A(s_k)) \cdot \frac{||f_n(x, y, \mathbf{d}) - \tilde{f}_n(s_k)||_2}{D(k, \mathbf{d})},$$
$$D(k, \mathbf{d}) = \sum_{(x,y)} ||f_n(x, y, \mathbf{d}) - \tilde{f}_n(s_k)||_2, \tag{5}$$

where K is the number of the example scenes with respect to the assigned activity, $D(k, \mathbf{d})$ is the sum of the feature distances for all possible positions in the room for the normalization purpose. As a result of Equation 5, areas in the given room with similar features to the example scenes would have low energy due to the second term in E_n. If those similar example scenes have higher assessment scores, the areas will have even lower energies than others, due to the first term in E_n. Note that such a calculation is conducted on the 2D floor plan, which is equivalent to employ a greedy strategy to traverse all sampled positions in the given room for assessment.

Further, the given room might have multiple components (i.e., windows/doors/artificial lights) that are more complex than the scenes in our datasets. Imagine the scenario that installing a new window to a room that already has one, the new window might have either no influence on a certain object group if that's too far away, or stacked influence cooperated with the original window. In the latter case, the stacked influence leads the probability of a certain position for object placement to be a weighted sum of the assessments about the two windows. Approximately, we use the mean assessment score as the stacked result. In this manner, we have two assumptions when using the above energy function: i) for large-size rooms, the room components that are not close to the object group will not impact the assessment; ii) the influence of the nearby room components on the assessment score is linear and stackable. Then, a compound assessment can be performed by using the energies in Equation (5) to find the proper position and direction for placing the object group in a room. Since different room features might have different effects on the assessment, weights are needed to balance the scores, assuming the feature that is more correlated to its assessment score would have a larger weight. For each pair of the dataset scenes, s_i and s_j, we calculate the correlation coefficient between the feature difference $||\tilde{f}_n(s_i) - \tilde{f}_n(s_j)||_2$ and the assessment difference $||A(s_i) - A(s_j)||_2$, denoted as W_n, and use its absolute value $|W_n|$ to describe the effect of the n-th type of room feature. Note that the weights can be adjusted for achieving better effects in practice. To this end, we obtain a compound assessment for the position (x, y) and the direction $\mathbf{d}$ as follows:

$$\arg\min_{x,y,\mathbf{d}} \sum_n |W_n| \cdot E_n(x, y, \mathbf{d}),$$
$$W_n = corr(||f_n(s_i) - f_n(s_j)||_2, ||A(s_i) - A(s_j)||_2). \tag{6}$$

Since we have traversed all sampled positions in the given

room to calculate the assessment energy, we can easily obtain the minimum value of Equation 6 from these sampled positions. As a result of the weight W_n, the four kinds of room features in Equation 3 have different impacts on different activity-related object groups. In our implementation, we observe that the weight of F_l (i.e., light-human distance) is significantly larger than other features for activities of "sitting and watching TV" and "lying and watching TV", while the weight of F_d (i.e., door-human angle) is significantly smaller than other features for activities of "conferencing" and "working in office". For other activities, the differences of room feature weights are not significant.

Based on our assumption that the influences of different types of room components are linear and stackable, the assessment for each type of feature can be done independently. Figure 5-(middle) visualizes the assessment results of a room based on the energies with the up direction of the object group (represented by the agent): a higher probability area (redder) that has lower energy is more likely to place the objects. The assessment with different directions can also ascertain the proper direction of the placed object group (see Figure 5-(Right)). Besides, our method can be used for rooms with partial features, e.g., a room without windows or a room without artificial lights. Similarly, for a given room with multiple windows/doors/artificial lights, we can first decompose the given room into multiple single-component rooms (i.e., with a single window, door, or artificial light), and then combine the independent assessments of these single-component rooms to obtain the compound assessment of the given room. For example, in Figure 6, the room with multiple windows in row (c) can be decomposed into two rooms with a single window ((a) & (b)) for assessment. Since the given room is decomposed into only two rooms, assessment maps of the decomposed rooms are combined with the same weight of 0.5 in terms of each direction, resulting in the final assessment result in (c). Analogously, rooms with multiple windows/doors can be assessed by our method.

Assessment-guided Synthesis. We have developed a user interface that assists users to easily construct an empty room, by specifying the size of the room, the positions of a window, door, and/or artificial light (e.g., droplight), and then assigning one or multiple activity labels to each room. Based on the assessment of an input room, our system can find the appropriate positions and directions of the object group, as well as its member objects with proper areas from our object database (collected from a well-known 3D Warehouse [32]). Our system can then generate 2D floor plans by applying the 2D projections of these objects. We can easily transfer 2D

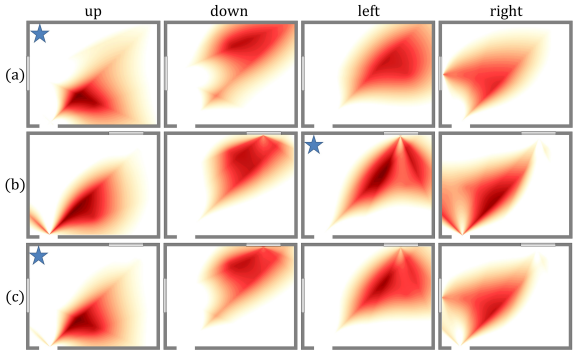


Fig. 6 Assessment results of activity "sitting and watching TV" in terms of the four directions of three given rooms. The suggested directions in three rows are marked with blue stars.

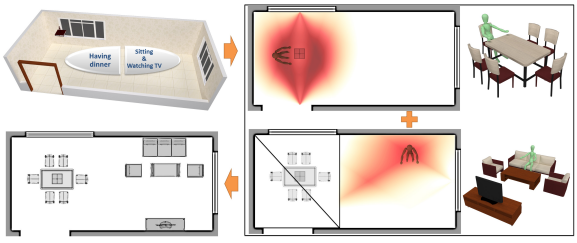


Fig. 7 Scene synthesis with multiple activity label inputs. After the object group for the first activity label is placed in the room (Top-right), the occupied area is then masked for the assessment about the next activity label (Bottom-right).

plans to 3D scenes, benefited from the 3D information of the objects in the database. For some furniture types like beds and TV sets, we snap them to the wall near the suggested position as the constraints to refine the layout.

An input room can have multiple activity labels for the placement of more than one object group. As illustrated in Figure 7, once an object group has been placed based on the first assigned activity label, the areas that have been occupied or too small to place any additional object are masked out. Then, our system assesses the remaining space in the room to place the next object group. Although so far our method focuses on single room modeling, large indoor scenes with multiple rooms can also be tackled by our method room by room. Moreover, aiming at relieving the workload of manually specifying activity labels, we can adopt a similar strategy for adaptive indoor scene modeling as [8], in which the area ratio between objects and room is used to measure whether more objects could be allowed to place in a room. In this way, the user can make a long list of activity labels, but how many labels are available depends on the size of the given room. Namely, a small room would only have few object groups in the list while a large room would have more. We show some application examples in Section 5.

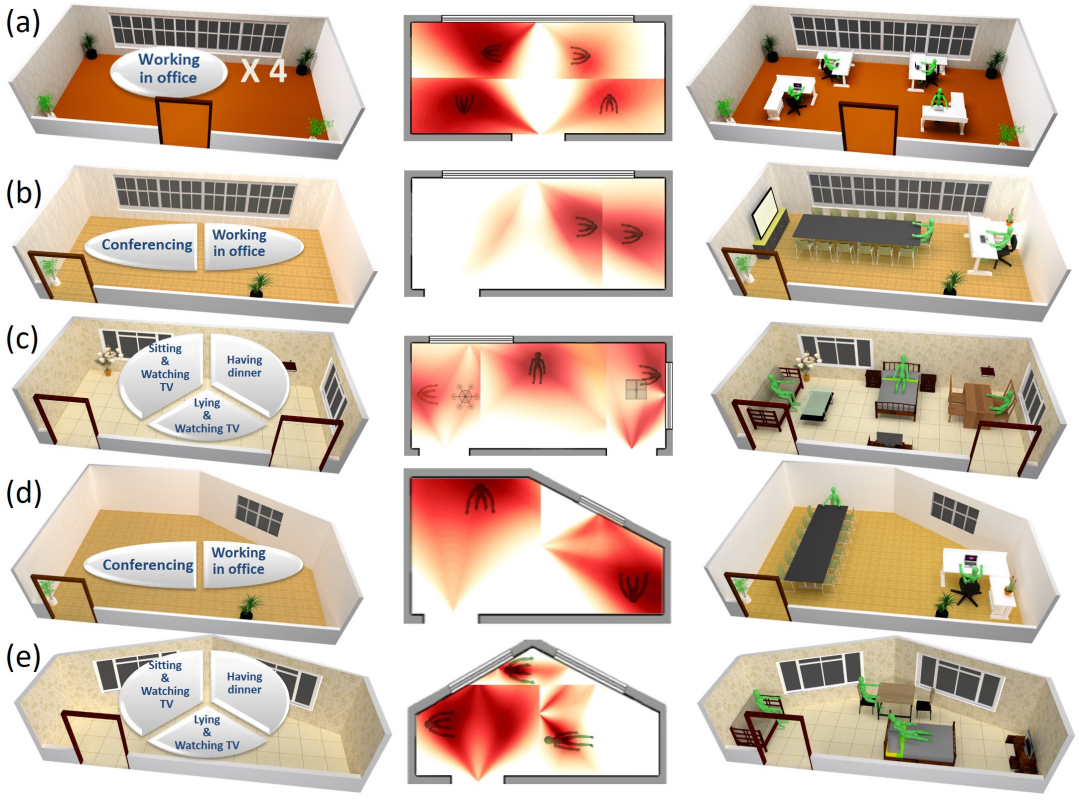


Fig. 8 Galleries of synthesized indoor scenes. In each case we show the input empty room with the assigned activity label(s), the assessment results of the suggested directions in 2D projections, and the 3D scenes generated by our method.

5 Results and Discussion

In this section, we first show various indoor scene modeling results by our approach, then evaluate our method through an ablation experiment, a user study in real-world scenes, and comparisons with an activity-centric method [8] and two deep-learning-based indoor scene modeling methods [10, 14].

Modeling Results. In Figure 8, we show the synthesized indoor scenes along with the assessment results of their corresponding rooms computed by our method, with respect to the user-specified activity labels. In case (a), we choose the assessment results with four different directions to place the same object group in a large room. We can see from the energy map (middle) that the four directions have different probability distributions and suggested positions. Note that we intend to show the relations between the suggested positions of objects and the four directions we focused on in this case. For the synthesis of large rooms with multiple objects of the same kind, symmetry should be considered, e.g., flip half or quarter of the designed scene to obtain symmetrical layouts. The other four cases show more complex scenes with multiple object groups, given different activity labels. In the last two cases ((d) & (e)), we test our method on non-rectangular rooms. Each non-rectangular room is tackled as a whole, with the areas outside the room masked out. It can be seen that, thanks to the defined room features, even though we do not have any non-rectangular scenes in our datasets, our method can still be used for the synthesis of plausible non-rectangular scenes.

In Figure 9-Top, using three cases we show how to combine multiple rooms into larger indoor scenes. For each case, the indoor scene modeling is conducted per room. Figure 9-Bottom shows three cases of adaptive indoor scene modeling. For each row, the sizes of the input rooms determine how many activity labels from the user-specified list (Left) can be adopted for indoor scene modeling. Note that the TV set is manually removed in the large room of the middle case for a better passageway. In this manner, our method can be used for both interactive and automated indoor scene modeling. In addition, our metrics for the room assessment with respect to certain object groups can also be used to evaluate indoor scenes. For example, in Figure 10, we normalize the compound assessment result of the object group in each scene in the range of 0 and 1. The scenes in the left column have better paths according to the door positions (e.g., the first and third cases), do not block the windows (e.g., the first and second cases), and better sense

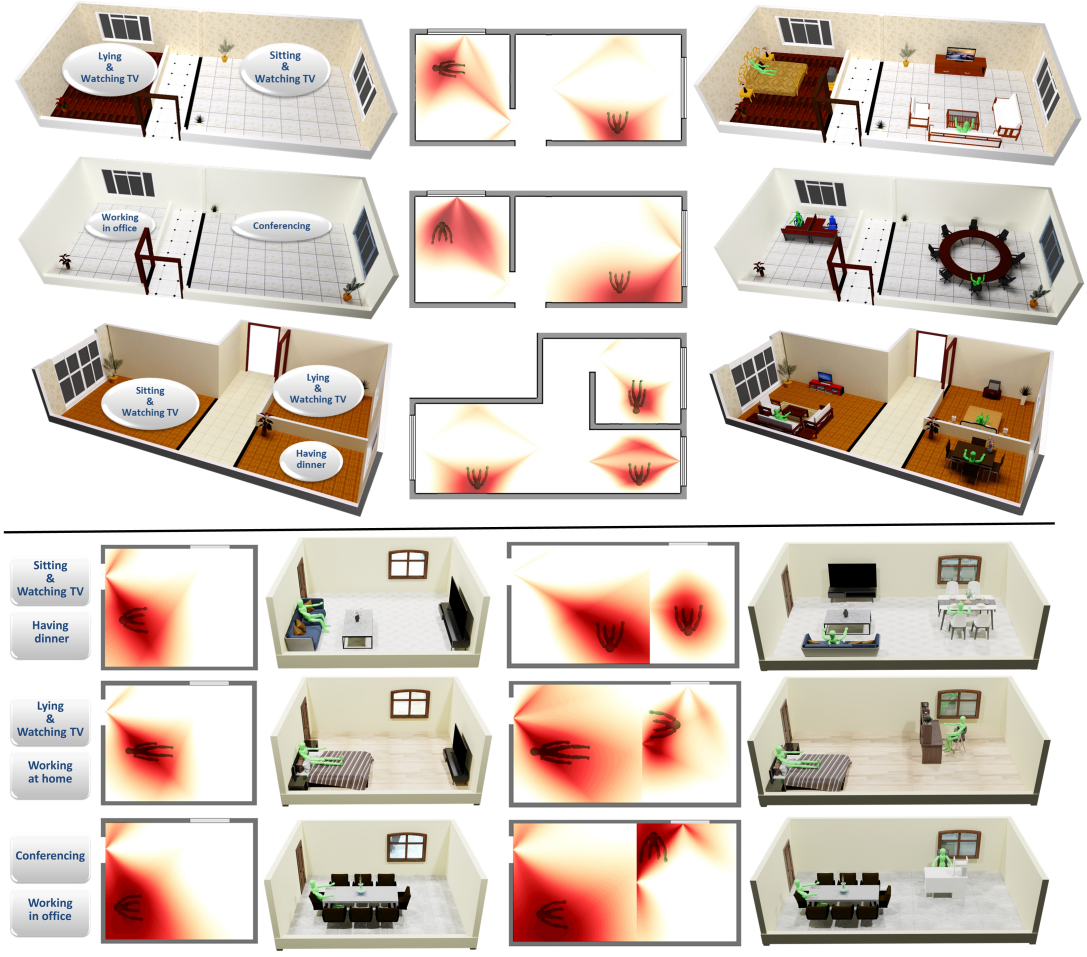


Fig. 9 **Top:** Large indoor scenes that are combined by the per-room modeling results. **Bottom:** Indoor scene modeling with adaptive groups of objects. In each row, we show the candidate activity labels and the synthesized scenes with small and large room inputs.

of privacy (e.g., the last office). We can see that scenes with good layouts would have high scores. This demonstrates that our method can refine large-scale indoor scene datasets via filtering out the examples with low assessment scores. On average, the generation of one indoor scene took less than 5 seconds per activity label for assessment and 10 seconds for object placement and system I/O, on an off-the-shelf computer with Intel Core i7-8550U 1.80GHz CPU and 8GB RAM.

Evaluations. In Equation 5, we set $A(s_i) = M_C(s_i)$ to encourage the high-quality indoor scene examples to play more important roles than the low-quality ones in indoor scene synthesis. We conducted an ablation experiment to evaluate the usability of the weighted indoor scene examples in our datasets. In Figure 11, the first row of energy maps are directly calculated by Equation 5, while the energy maps in the second row are calculated by setting $A(s_i) = 0$ in Equation 5. Four energy maps in each row are corresponding to the up, down, left, and right directions for placing the object groups, respectively. At the bottom of Figure 11, we show two 3D scenes corresponding to energy maps (a) and (b). Since the position of the window might lead backlight problem to scene (b) for watching TV, scene (a) has a relatively better layout than scene (b). If we do not weight the dataset indoor scenes (i.e., set $A(s_i) = 0$), Equation 5 can hardly generate scenes like (a) when the quantity of the scene examples similar to (a) is smaller than the examples similar to (b). Therefore, our method that weights the dataset is effective for the small dataset with differentiated examples.

On the other hand, we conducted a user study to demonstrate that the dataset weights (i.e., the fuzzy measurement $M_C(s_i)$ in 1) are consistent with the subjective perceptions. Since our method can provide common-seen layouts in the real world for most residential scenes, we only conducted the user study on office scenes that always have various layouts. We recruited 10 volunteers (postgraduate students) as two groups (5 participants in each group) to evaluate two office scenes

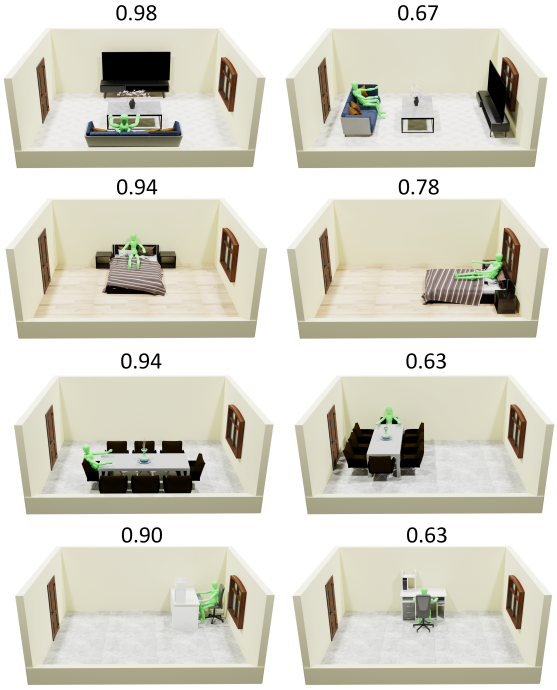


Fig. 10 Indoor scenes evaluated by our metrics and the assessment scores, with respect to the layouts of the groups of objects.

(Figure 12-Left) in terms of six different positions for the activity label "working in office", by giving scores from 0 (worst) to 1 (best). We also used our method to assess the same two scenes, based on the participants' positions and directions. To normalize the assessment results of our method and make them comparable with the participants' scores. We proportionally mapped our results to the range between the maximum and minimum participants' scores. We post the assessment results of ours and participants' in Figure 12-Right. Even though the perceptions are too subjective to be precisely measured, the comparison results still show that the priors extracted from our dataset are similar to what we can obtain from real-world experience.

Moreover, we compared our modeling results with the state-of-the-art indoor scene modeling methods [8, 10, 14] to demonstrate the effectiveness of our method. In Figure 13-Left-Top, we compare our method with an activity-centric method [8]. For two given rooms (a small one and a large one), we use "lying and watching TV" + "working at home" and "bed" + "bookshelf" as the input labels of our method and the method of [8], respectively. The indoor scene synthesis results show that even employing different strategies of object exploration, both methods can explore proper activity-related indoor objects that suit the size of the input room. More specifically, the method of [8] leverages the activity-associated object relation graphs to determine the proper object categories in the room, while our

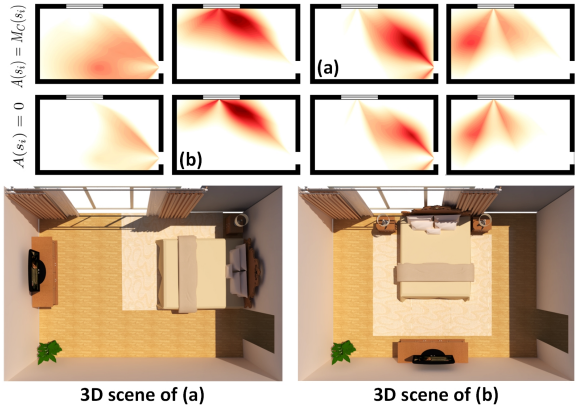


Fig. 11 **Top:** Two rows of the calculated energy maps in terms of four directions, given the same room in an ablation experiment. **Bottom:** 3D scenes corresponding to energy maps (a) and (b).

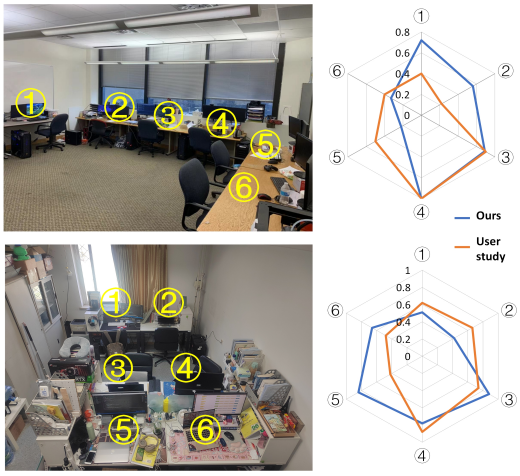


Fig. 12 Comparison of the subjective assessments of humans on two offices with six desks in the real-world and the corresponding assessment results of our method.

method simply uses the order of the specified activity labels as their priorities for choosing the associated pre-defined object groups. Limited by the pre-defined object groups, some indirect interactive objects (e.g., bookshelf) are not in any object groups considered by our method, so the explored objects in our results are fewer than [8]. However, the method of [8] only has a single synthesized layout for the suggested objects, while our method can make more variations on the indoor layout, just by changing the order of the input activity labels or using the sub-optimal energy maps.

In Figure 13-Right, we show the comparison between our results and results of [10]. The model of [10] is based on a convolutional neural network and trained by a large-scale indoor scene database [33]. The comparison results show that both of the methods can generate plausible indoor layouts for the given room. Note the given room may not be similar to any scene examples in the datasets used in the two methods. In Figure 13-Left-Bottom, we compare

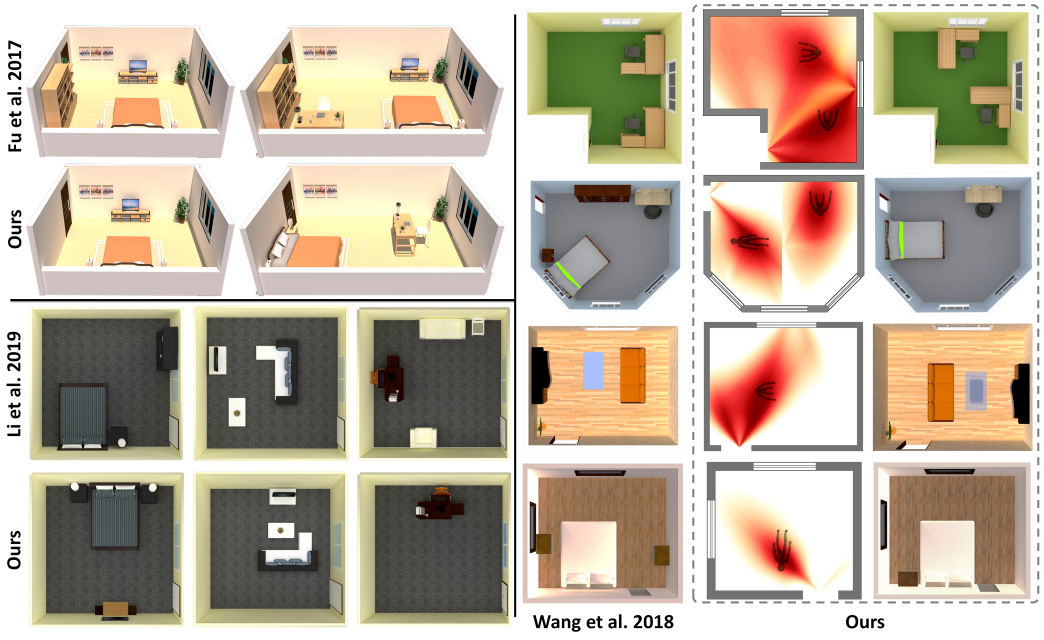


Fig. 13 Comparisons of our results with the synthesized indoor scenes of Fu et al. 2017 [8] (Left-Top), Wang et al. 2018 [10] (Right), and Li et al. 2019 [14] (Left-Bottom)

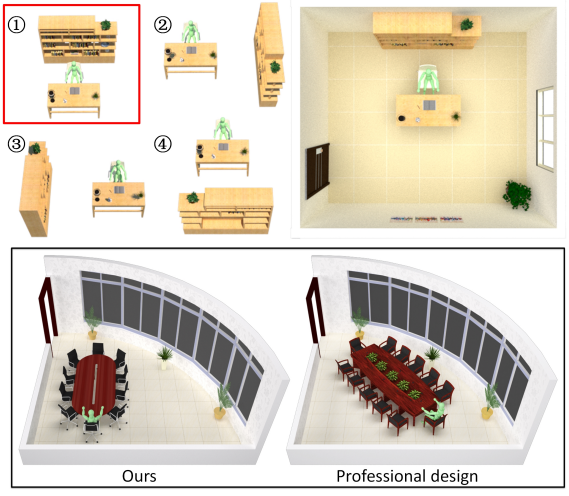


Fig. 14 **Top:** An example that adds a bookshelf into the "working at home" group with four pre-set relations (Left). Users can select one of them (e.g., in the red box) to synthesize the indoor scene (Right). **Bottom:** Our result and the professional design for a room with a cambered edge.

the synthesized indoor scenes given rectangular rooms by our method and by the method in [14]. Since the method in [14] does not consider the impact of windows or doors, even though their generated scenes could have richer content, our results look more appropriate for the environment of the given room. Comparing the layouts between ours and methods [10, 14] in terms of the given environment, furniture in their methods might block the window while ours do not, e.g., the second row in Figure 13-Right and the first column in Figure 13-Left-Bottom. Our results can have better door paths that would not impact the activity of the object groups, e.g., the third row in Figure 13-Right and the second column in Figure 13-Left-Bottom. Note that since the influences of the environmental factors and their mixed effects on the quality of the synthesized scenes are subjective, using fuzzy measurement to collect such priors could be more flexible than directly using hard constraints such as setting the relevant positions/directions between object groups and windows/doors. However, methods of [10, 14] can make more layout variations in their synthesized indoor scenes, benefiting from the large-scale training datasets.

We also conducted a user study to compare the quality of the generated layouts between our method and the above two methods [10, 14]. 17 participants (undergraduate students majored in digital media technology) were recruited to compare 16 pairs of scenes (8 pairs for our method and [10], and 8 for our method and [14]), and choose the better layout from each pair (the order of the two scenes in each pair was randomized). Note such two comparisons were separately conducted. The result of the first one shows that, in 136 comparisons (17×8) our results won 67 times (49.3%), while [10] won 69. The result of the second one shows that our results won 70 times (51.5%) in 136 comparisons, while [14] won 66. These results demonstrate that our method can generate indoor layouts with comparable quality to the two

indoor scene synthesis methods [10, 14]. It is noteworthy that, taking a sharp turn from the state-of-the-art methods that generally train deep-learning-based models with large-scale indoor scene datasets, our method only utilizes much smaller datasets of differentiated examples for indoor scene synthesis and can produce comparable results.

Limitations. Our current approach has several limitations. First, our method is based on several simplified assumptions: utilizing doors and windows as spots in the room features; ignoring the cross effects between object groups when collecting the indoor scene datasets; and assuming the impacts of windows/doors/artificial lights on room assessment are linear and stackable. Second, new subjective comparisons need to be conducted if we add new data, especially with respect to new object groups, to the datasets. This might somewhat limit the scalability of our current method. As future work, we plan to study the transferability of the subjective evaluations between different object groups. Besides, our method needs pre-defined relationships within an object group. For some decorations like potting and mural, it still demands user assistance for the manual placement of objects into a scene. Some indirect interactive objects are also not considered in the currently pre-defined object groups, since we cannot always use a fixed arrangement (i.e., relative positions/directions) for these objects in a group. Aiming at addressing the placement of such objects and making more variations of the in-group object arrangement, we can use flexible object relations that contain multiple examples of in-group object arrangement in the further. For example, in Figure 14-Top, we pre-set four object relations for adding the bookshelf into the object group associated with "working at home", while the user decides which one is better for the synthesized indoor scene. Third, since the activity labels are assigned by users, improper user-specified object groups might lead to unnatural results (e.g., dining room with a bed). This can be relieved by employing priors such as relation graphs which indicate the co-existing possibility between different object groups. Lastly, since we only consider four different directions, our current implementation can only generate axis-aligned layouts even given non-rectangular rooms (e.g., Figure 8(d & e)). However, for non-axis-aligned rooms or rooms with cambered edges, there may always exist better object arrangement solutions rather than axis-aligned layouts. For example, in Figure 14-Bottom, given a room with a cambered bay window and the assigned activity of conferencing, we show a scene synthesized by our method (Left) and one by an interior designer (Right). Our method suggests a small conference table and places it to cover a small part of the room, while the interior designer chooses a larger table with an oblique direction, making better use of both the space and light. To alleviate this problem, our approach assists the user to slightly adjust the position and direction of the objects to refine the indoor layout.

6 Conclusion

In this paper, we present a new approach to using datasets of differentiated examples to support indoor scene modeling. To construct such datasets, we conduct subjective comparisons on special-designed indoor scenes with different room features in terms of certain object groups, and then compute the membership degrees of scene quality for the example scenes in our datasets as their assessment scores. Given a new room and user-specified activity label(s), our approach uses the labeled dataset scenes as priors to assess the given room, and suggests the appropriate positions and directions for the placement of the object groups that are pre-associated with the activity labels. In this way, our approach is able to differentiate the qualities of the indoor scene examples when using them to guide scene modeling. It can open up new research opportunities towards example-driven indoor scene modeling.

In the future we plan to further extend our room features to tackle more types of factors including colors, decorations, and furniture styles, to handle rooms with more complex shapes (e.g., with cambered edges), and to enable the automation of designing more comfortable indoor scenes for target activities. We are also interested in exploring the cross-activity influences on indoor layouts to increase the practicality of our method, especially the scenarios that an object group is associated with multiple activities. This involves the priorities of different activities, which could also be collected from the subjective comparison experiments. Such priorities would enable the weighted superposition of the energy maps for all specified activities on a single object group, thus jointly impacting the output layout. To improve the scalability of our approach, we also plan to study the similarities of different kinds of object groups, to extend a limited number of labeled data for more types of indoor scenes in the future.

Acknowledgements

We thank the anonymous reviewers for the constructive comments. This work was partially supported by grants from the the NSFC (No.61902032), Research Grants Council of the Hong Kong Special Administrative Region, China (No. CityU 11237116), City University of Hong Kong (No. 7004915).

Declaration of Competing Interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Yu LF, Yeung SK, Tang CK, Terzopoulos D, Chan TF, Osher S. Make it home: automatic optimization of furniture arrangement. *ACM Transactions on Graphics*, 2011, 30(4): 86:1–86:12.

[2] Merrell P, Schkufza E, Li Z, Agrawala M, Koltun V. Interactive Furniture Layout Using Interior Design Guidelines. *ACM Transactions on Graphics*, 2011, 30(4): 87:1–87:10.

[3] Chen X, Li J, Li Q, Gao B, Zou D, Zhao Q. Image2Scene: Transforming Style of 3D Room. In *ACM International Conference on Multimedia*, 2015, 321–330.

[4] Fisher M, Ritchie D, Savva M, Funkhouser T, Hanrahan P. Example-based Synthesis of 3D Object Arrangements. *ACM Transactions on Graphics*, 2012, 31(6): 135:1–135:11.

[5] Jiang Y, Lim M, Saxena A. Learning Object Arrangements in 3D Scenes using Human Context. In *International Conference on Machine Learning*, 2012, 907–914.

[6] Fisher M, Savva M, Li Y, Hanrahan P, Nießner M. Activity-centric Scene Synthesis for Functional 3D Scene Modeling. *ACM Transactions on Graphics*, 2015, 34(6): 179:1–179:13.

[7] Savva M, Chang AX, Hanrahan P, Fisher M, Nießner M. PiGraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics*, 2016, 35(4): 139:1–139:12.

[8] Fu Q, Chen X, Wang X, Wen S, Zhou B, Fu H. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics*, 2017, 36(6): 201:1–201:13.

[9] Qi S, Zhu Y, Huang S, Jiang C, Zhu SC. Human-centric Indoor Scene Synthesis Using Stochastic Grammar. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 5899–5908.

[10] Wang K, Savva M, Chang AX, Ritchie D. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics*, 2018, 37(4): 70:1–70:14.

[11] Wang K, Lin YA, Weissmann B, Savva M, Chang AX, Ritchie D. PlanIT: Planning and Instantiating Indoor Scenes with Relation Graph and Spatial Prior Networks. *ACM Transactions on Graphics*, 2019, 38(4): 132:1–132:15.

[12] Saaty TL. *The Analytic Hierarchy Process*, McGraw-Hill1980.

[13] Klir GJ, Yuan B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice Hall1995.

[14] Li M, Patil AG, Xu K, Chaudhuri S, Khan O, Shamir A, Tu C, Chen B, Cohen-Or D, Zhang H. GRAINS: Generative Recursive Autoencoders for INdoor Scenes. *ACM Transactions on Graphics*, 2019, 38(2): 12:1–12:16.

[15] Fisher M, Hanrahan P. Context-based Search for 3D Models. *ACM Transactions on Graphics*, 2010, 29(6): 182:1–182:10.

[16] Fisher M, Savva M, Hanrahan P. Characterizing Structural Relationships in Scenes Using Graph Kernels. *ACM Transactions on Graphics*, 2011, 30(4): 34:1–34:12.

[17] Sharf A, Huang H, Liang C, Zhang J, Chen B, Gong M. Mobility-Trees for Indoor Scenes Manipulation. *Computer Graphics Forum*, 2014, 33(1): 2–14.

[18] Xu K, Ma R, Zhang H, Zhu C, Shamir A, Cohen-Or D, Huang H. Organizing Heterogeneous Scene Collections Through Contextual Focal Points. *ACM Transactions on Graphics*, 2014, 33(4): 35:1–35:12.

[19] Liu T, Chaudhuri S, Kim VG, Huang Q, Mitra NJ, Funkhouser T. Creating Consistent Scene Graphs Using a Probabilistic Grammar. *ACM Transactions on Graphics*, 2014, 33(6): 211:1–211:12.

[20] Zhang SH, Zhang SK, Xie WY, Luo CY, Yang Y, Fu H. Fast 3D Indoor Scene Synthesis by Learning Spatial Relation Priors of Objects. *IEEE Transactions on Visualization and Computer Graphics*, 2021, Early Access: 1–1, doi:10.1109/TVCG.2021.3050143.

[21] Ma R, Li H, Zou C, Liao Z, Tong X, Zhang H. Action-Driven 3D Indoor Scene Evolution. *ACM Transactions on Graphics*, 2016, 35(6): 173:1–173:13.

[22] Ma R, Patil AG, Fisher M, Li M, Pirk S, Hua BS, Yeung SK, Tong X, Guibas L, Zhang H. Language-Driven Synthesis of 3D Scenes from Scene Databases. *ACM Transactions on Graphics*, 2018, 37(6): 212:1–212:16.

[23] Xu K, Chen K, Fu H, Sun WL, Hu SM. Sketch2Scene: Sketch-based Co-retrieval and Co-placement of 3D Models. *ACM Transactions on Graphics*, 2013, 32(4): 123:1–123:15.

[24] Chen K, Lai YK, Wu YX, Martin R, Hu SM. Automatic Semantic Modeling of Indoor Scenes from Low-quality RGB-D Data Using Contextual Information. *ACM Transactions on Graphics*, 2014, 33(6): 208:1–208:12.

[25] Shao T, Monszpart A, Zheng Y, Koo B, Xu W, Zhou K, Mitra NJ. Imagining the Unseen: Stability-based Cuboid Arrangements for Scene Understanding. *ACM Transactions on Graphics*, 2014, 33(6): 209:1–209:11.

[26] Zhang S, Li Y, He Y, Yang Y, Zhang S. MageAdd: Real-Time Interaction Simulation for Scene Synthesis. In *ACM International Conference on Multimedia*, 2021, 965–973.

[27] Frontczak M, Wargocki P. Literature survey on how different factors influence human comfort in indoor environments. *Building & Environment*, 2011, 46(4): 922–937.

[28] Huang L, Zhu Y, Ouyang Q, Cao B. A study on the effects of thermal, luminous, and acoustic environments on indoor environmental comfort in offices. *Building and Environment*, 2012, 49: 304–309.

[29] Konis K. Predicting visual comfort in side-lit open-plan core zones: results of a field study pairing high dynamic range images with subjective responses. *Energy and Buildings*, 2014, 77: 67–79.

[30] Ochoa CE, Capeluto IG. Evaluating visual comfort and performance of three natural lighting systems for deep office buildings

in highly luminous climates. *Building and Environment*, 2006, 41(8): 1128–1135.

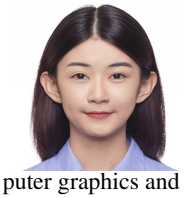
[31] Zhang Z, Yang Z, Ma C, Luo L, Huth A, Vouga E, Huang Q. Deep Generative Modeling for Scene Synthesis via Hybrid Representations. *ACM Transactions on Graphics*, 2020, 39(2): 17:1–17:21.

[32] 3D Warehouse. https://3dwarehouse.sketchup.com/.

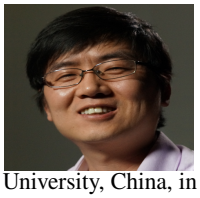
[33] Song S, Yu F, Zeng A, Chang AX, Savva M, Funkhouser T. Semantic Scene Completion from a Single Depth Image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 190–198.

Author biography

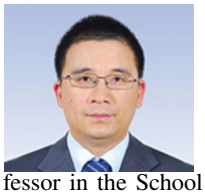
Qiang Fu is an Associated Professor in the School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications. He received the Ph.D. degree in computer science from Beihang University, China, in 2018 and the B.E. degree in automation from Beihang University, China, in 2011. His research interests include Computer Graphics and Virtual Reality.



Shuhan He received the B.E. degree in Digital media technology from Beijing Forestry University. She is a postgraduate student in the School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications. Her research interests include computer graphics and machine learning.



Hongbo Fu is a Professor in the School of Creative Media, City University of Hong Kong. He received the PhD degree in computer science from the Hong Kong University of Science and Technology in 2007 and the BS degree in information sciences from Peking University, China, in 2002. His primary research interests fall in the fields of computer graphics and human computer interaction. He has served as an associate editor of The Visual Computer, Computers & Graphics, and Computer Graphics Forum.



Xueming Li received the B.E. degree in electronics engineering from the University of Science and Technology of China, in 1992, and the Ph.D. degree in electronics engineering from the Beijing University of Posts and Telecommunications in 1997. He is a professor in the School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications. His research interests include digital image processing, video coding, and multimedia telecommunication.

Zhigang Deng received the BS degree in mathematics from Xiamen University, China, the MS degree in computer science from Peking University, China, and the PhD degree in computer science from the Department of Computer Science, University of Southern California, in 2006. He is Moores Professor of computer science with the University of Houston. His research interests include computer graphics, computer animation, and HCI.