Computational Visual Media
https://doi.org/10.1007/s41095-022-0299-z

Research 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 incorporate interior design knowledge, which has been an essential prior for most indoor scene modeling methods. However, the layout qualities of indoor scene datasets are often uneven, and most existing data-driven methods do not differentiate indoor scene examples in terms of quality. In this work, we aim to explore an approach that leverages datasets with differentiated indoor scene examples for indoor scene modeling. Our solution conducts subjective evaluations on lightweight datasets having various room configurations and furniture layouts, via pairwise comparisons based on fuzzy set theory. We also develop a system to use such examples to guide indoor scene modeling using 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 and direction of an object group and the room configuration. To perform indoor scene modeling, given an empty room, our system first assesses it in terms of the user-specified object groups, and then places associated objects in the room guided by the assessment results. A series of experimental results and comparisons to 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 example; fuzzy sets

1 Introduction

The problem of indoor scene modeling has been extensively studied in the last decade. From early guideline-based approaches [1, 2] to example-based approaches [3, 4], later activity-centric methods [5–9], and deep learning models [10, 11], their capability to generate visually pleasing and functionally-valid 3D indoor scenes has advanced steadily, benefiting many applications including games and interior design.

To obtain plausible indoor layouts and object arrangements, most existing indoor scene modeling approaches rely on either expert-designed guidelines or examples. For data-driven methods, using large datasets of indoor scenes can improve the quality of the synthesized scenes through abundance of examples. However, the layouts of indoor scene examples in large datasets may vary in quality. Due to the lack of metrics to evaluate the quality of indoor scenes based on associated functionality, most existing indoor scene modeling methods give low-quality examples the same weights as high-quality ones. Intuitively, indoor scenes with different qualities, called *differentiated examples* in this work, should play different roles in indoor scene modeling: the impact of high-quality examples should be enhanced while that of others should be reduced.

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 support their assumed uses well. Even for rooms with specially-designed layouts, their furniture layout can still have some common relationships to the room configuration, including room size and shape, positions of windows and 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 common layout relations to provide metrics to differentiate

1 School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications, Beijing, China. E-mail: Q. Fu, fu.john.qiang@gmail.com; S. He, heshuhan@bupt.edu.cn; X. Li, lixm@bupt.edu.cn.
2 School of Creative Media, City University of Hong Kong, Hong Kong, China. E-mail: fuplus@gmail.com.
3 Department of Computer Science, University of Houston, TX, USA. E-mail: zdeng4@uh.edu (✉).
Manuscript received: 2022-03-17; accepted: 2022-06-15

examples in the datasets. Besides handling a variety of room layouts, the evaluation metrics need to be also associated with object functionality. Therefore, examples in an ideal interior design dataset should: (i) be able to be classified into functionality-associated object groups, (ii) be differentiated examples with associated evaluations, and (iii) include sufficient layout variation 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 of indoor scenes is challenging, we leverage fuzzy measures and subjective comparisons to evaluate the layout quality of the differentiated samples. Specifically, we adopt *membership degree*, a concept borrowed from fuzzy set theory [12, 13], to evaluate the samples in the dataset through pair-wise comparisons of the samples. See Fig. 1. Our method collects differentiated examples of indoor scenes to facilitate indoor scene modeling (a). Given input rooms and assigned activity labels representing certain object groups (b), our method uses differentiated examples, which have been given subjective evaluation scores, 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 by 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, thereby synthesizing 3D scenes with plausible indoor layouts (d). Moreover, we provide an ease-of-use tool to assist users in designing indoor scenes. It also allows users to merge multiple rooms into a larger and more complex scene.

In summary, our work makes two novel contributions: (i) a metric to assess indoor scenes through differentiated examples in a dataset, based on fuzzy set theory, and (ii) a framework to model indoor scenes based on 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 to state-of-the-art, data-driven indoor scene synthesis methods [8, 10, 14].

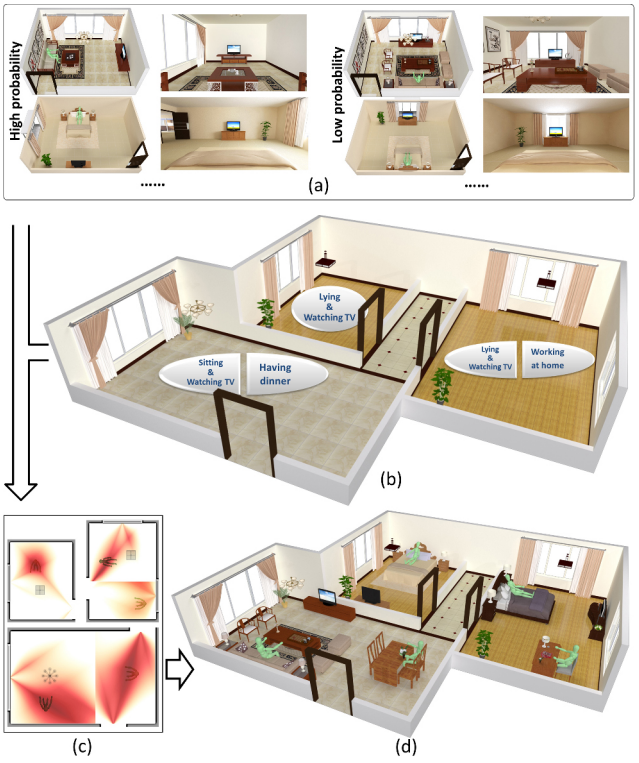


Fig. 1 Elements of our method: (a) datasets, (b) input room and activity labels, (c) room assessment results, and (d) synthesized scene based on the assessment.

2 Related work

Many systems and approaches for indoor scene modeling have been proposed in the past decade. 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., Refs. [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., Refs. [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. 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., Ref. [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, evaluating the quality of indoor scenes is a challenging yet wide-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, may be utilized to measure the quality of an indoor environment [27]. Researchers have revealed that the indoor environment can impact the comfort and cognitive performance of human beings, and acceptable ranges exist 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 indoor illumination. Ochoa and Capeluto [30] proposed a similar analysis of indoor illumination with simulated indoor environments to evaluate visual comfort. In our work, we extract expert knowledge from datasets of differentiated indoor scene examples. Evaluation of differentiated examples is performed through subjective comparisons, and we use the evaluation results, 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 rely on pre-defined guidelines or relations learned from 3D scene datasets (e.g., Refs. [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 contexts 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, which are placed into the given room as a whole. More specifically, our work relies on methods such as those from Ref. [7] to determine relevant object positions and directions in a group (e.g., a group of a 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. Unlike human-centric methods (e.g., Refs. [5, 6]) that directly measure the probability of various activities in a certain region in a room, we leverage 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's indoor scenes weighted by fuzzy metrics to guide 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] employed 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 a normal distribution to the distribution of primary objects in indoor scenes. This work focuses 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 of objects within 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 suited to certain human activities. To describe such relations, we define the room features in terms of the environment-related components like windows and doors. Moreover, unlike these deep-learning-based methods, our method does not rely on a large indoor scene dataset.

3 Data preprocessing

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

3.1 Scene data collection and representation

To verify the usability of differentiated examples as priors for indoor scene modeling, we collected 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: scenes in the same dataset have the same kind of object group. As object groups are generally associated with certain activities, we choose the activity name as the object group label in the user interface of our system.

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. We then place the human agent in the object group to represent its forward direction and position. Note that, for a given room, the user could specify multiple activity labels to generate an indoor scene with more than one object group. In our preliminary work, we only focus on six types of common object groups. As shown in Fig. 3, each object group is associated with a certain activity including lying down and watching TV (a), sitting and watching TV (b), working at home (c), having dinner (d), conferencing (e), and office work (f).

For each type of scene dataset (see Fig. 2), to generate the other scene examples, we first choose a well-designed indoor scene (above-left) and then change its room size, the positions of windows/doors/artificial lights, or the positions/directions of the object groups. This leads 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 bottom-left corner of the floor as the origin of the 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 (up, down, left, and right). We also limit the range of room sizes to avoid too small or too large

Fig. 2 Examples of dataset scenes with different room configurations and furniture layouts.

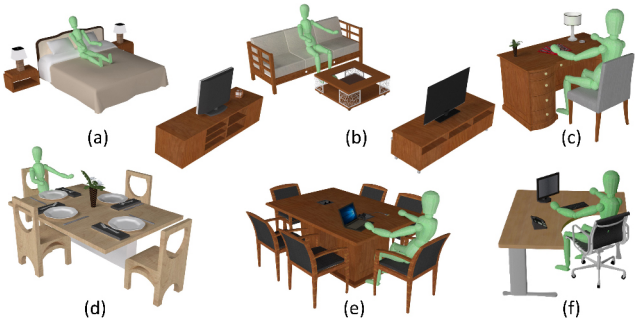


Fig. 3 Object groups and associated agents for various activities.

rooms in the datasets. In our lightweight datasets, we have a total of 48 scene examples of 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 to balancing the quantities between good- and poor-quality scene examples. Since these datasets are small, we can conduct subjective scoring comparisons on them. 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 relationships between furniture layout and room components (wall/window/door), rather than simply use the whole shape of the room. Specifically, we consider four types of room features to establish such relationships. The features of windows and doors are associated with 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 in the object group, and thus impact the subjective perception of humans on the associated activity. Considering symmetry, we use sine-squared functions of the included angles as features. The features of artificial lights and walls are associated with their relative distances to the object group. We directly use 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 Fig. 4(below) shows, we define room features based on the frontal 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 (see Fig. 4(above)); we discuss general cases later. Let $\boldsymbol{d}$ denote the forward

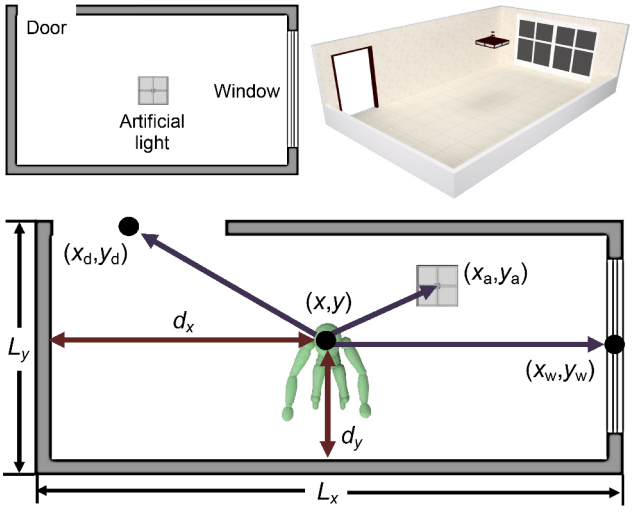


Fig. 4 Above: 2D plan of an empty room (left) and corresponding 3D scene (right). Below: room features and parameters related to the human agent (representing object groups).

direction and (x, y) denote the position of the object group. Let d_x and d_y be the distances from the object group to the right-side and front walls respectively, L_x and L_y be the corresponding side lengths of a rectangular room (or the oriented bounding box of a non-rectangular room), and $(x_\mathrm{w}, y_\mathrm{w}), (x_\mathrm{d}, y_\mathrm{d})$, and $(x_\mathrm{a}, y_\mathrm{a})$ denote the respective positions of the window, door, and artificial light. To describe the relationships between the object group and room components including windows, doors, walls, and lights, the room features comprise the angles between the front direction of the object group and the object-to-window/-door direction that measure the relationships 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 $\boldsymbol{d}$ and the directions from (x, y) to $(x_\mathrm{w}, y_\mathrm{w})$ and from (x, y) to $(x_\mathrm{d}, y_\mathrm{d})$, the distance F_a from (x, y) to $(x_\mathrm{a}, y_\mathrm{a})$, and the relative position F_l of the object group in the room, are described as Eq. (1):

$$
\begin{aligned}
F_\mathrm{w}(x, y, \boldsymbol{d}) &= 1 - \left(\frac{\boldsymbol{d} \cdot (x - x_\mathrm{w}, y - y_\mathrm{w})}{\| \boldsymbol{d} \|_2 \cdot \| (x - x_\mathrm{w}, y - y_\mathrm{w}) \|_2} \right)^2 \\
F_\mathrm{d}(x, y, \boldsymbol{d}) &= 1 - \left(\frac{\boldsymbol{d} \cdot (x - x_\mathrm{d}, y - y_\mathrm{d})}{\| \boldsymbol{d} \|_2 \cdot \| (x - x_\mathrm{d}, y - y_\mathrm{d}) \|_2} \right)^2 \\
F_\mathrm{a}(x, y) &= \| (x - x_\mathrm{a}, y - y_\mathrm{a}) \|_2 \\
F_\mathrm{l}(x, y, \boldsymbol{d}) &= \left(\frac{d_x}{L_x}, \frac{d_y}{L_y} \right)
\end{aligned}
\tag{1}
$$

4 Fuzzy-based subjective comparisons

Due to the lack of quantitative metrics for layout quality evaluation, we apply subjective evaluations to discriminate the indoor scene examples in the datasets. Inspired by fuzzy set theory [13], especially the analytic hierarchy process [12], we employ pairwise comparisons to score the differentiated indoor scene examples by calculating their membership degrees. To reduce the impact of individual biases, we recruited 32 participants to compare 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, choosing the one they preferred. We collected such intuitive but fuzzy comparisons instead of accurate and professional evaluations for two reasons: (i) such comparisons do not require professional interior designers so are easy to conduct, and (ii) the non-expert users can still judge the quality of an indoor scene from their own perspective: even if they might not know exactly how and why high-quality scenes always look visually appealing. In total, we collected 2962 comparisons as follows: sitting and watching TV (496), lying and watching TV (713), having dinner (372), working at home (544), conferencing (310), and office work (527).

Let $\mathcal{S} = \{s_1, s_2, \cdots s_M\}$ be a set of scenes of the same type. We construct the pairwise comparison matrix G as Eq. (2):

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

Unlike Ref. [12], which uses *intensity of importance* (from 1 to 9) to construct the pairwise comparison matrix, the entries of the above matrix in our method are defined such that 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 binary, we simply define the matrix entries as

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

where $p_{s_j}(s_i)$ represents the number of votes where the participants felt scene s_i to be better than scene s_j. Since each participant only compared some 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 square root of the comparison frequency $t(i,j) = \sqrt{n_{i,j}/N}$, where N is the total number of participants and $n_{i,j}$ is the number of comparisons for pair of scenes s_i and s_j. Note $t(i,j) = 0$ if scenes i, j have not been compared (and so $t(i,i) = 0$). Mathematically, for each dataset scene s_i, the membership degree function for scene quality is

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

where $T = \sum_{j=1}^{M} t(i,j)$. 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 $M_C(s_i)$.

5 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.

5.1 Room assessment

The continuous variations of room configuration form a space containing all possible scenes for a certain group of objects, the domain $\mathcal{U}$, and the differentiated scene examples in our datasets can be considered as some sparsely sampled examples in $\mathcal{U}$. The mapping $A(\mathcal{U}) \to \mathcal{V}$ is used to assess scenes in $\mathcal{U}$ and output $\mathcal{V} \in [0,1]$. The aforementioned user study provides 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 a mapping A for all scenes in $\mathcal{U}$.

Since the scene examples in our datasets are characterized by 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 $\{F_{\mathrm{w}}, F_{\mathrm{d}}, F_{\mathrm{l}}, F_{\mathrm{a}}\}$ in Eq. (1), let $f_n(x, y, \boldsymbol{d})$ be the value of the n-th type of feature at position (x, y) in the room with the direction $\boldsymbol{d}$

for 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 position. To reuse the assessment of our example scenes on the given room, we define the assessment energy at position (x, y) with direction $\boldsymbol{d}$ as Eq. (5):

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

where K is the number of example scenes for the assigned activity, and $D(k, \boldsymbol{d})$ is the sum of feature distances for all possible positions in the room, for normalization. As a result of Eq. (5), areas in the given room with similar features to the example scenes have low energy due to the second term in E_n. If these similar example scenes have high assessment scores, the areas will have lower energies than other scenes, due to the first term in E_n. Note that this calculation is conducted on the 2D floor plan, and is equivalent to employing a greedy strategy to traverse all sampled positions in the given room for assessment.

Note that the given room might have multiple components (windows/doors/artificial lights) and so be more complex than the scenes in our datasets. Imagine adding a new window to a room that already has one. The new window might have no influence on a certain object group if it is too far away, or add a further influence in conjunction with that of the original window. The latter case leads to the assessment of a certain position for object placement to be a weighted sum of the assessments for the two windows. Approximately, we may use the mean assessment score as the combined result, relying on two assumptions when using the above energy function: (i) for large rooms, room components that are not close to the object group will not impact the assessment, and (ii) the influences of nearby room components on the assessment score are additive and linear. Thus, a compound assessment can be performed by using the energies in Eq. (5) to find the proper position and direction for placing the object group in a room. Since different room features may have different effects on the assessment, weights are needed to balance the scores, assuming that the feature that is more correlated to its assessment score should have a larger weight. For each pair of dataset scenes, s_i and s_j, we calculate the correlation coefficient W_n between the feature difference $\| \tilde{f}_n(s_i) - \tilde{f}_n(s_j) \|_2$ and the assessment difference $\| A(s_i) - A(s_j) \|_2$, 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 to achieve better effects in practice. Overall, we obtain a compound assessment for the position (x, y) and the direction $\boldsymbol{d}$ as Eq. (6):

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

Since we traverse all sampled positions in the given room to calculate the assessment energy, we can easily obtain the minimum value of Eq. (6) from these sampled positions. As a result of the weight W_n, the four kinds of room features in Eq. (3) have different impacts on different activity-related object groups. In our implementation, we observe that the weight of F_l (light-to-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 (door-to-human angle) is significantly smaller than other features for activities of *conferencing* and *office work*. 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 additive and linear, the assessment for each type of feature can be done independently. Figure 5(middle) visualizes

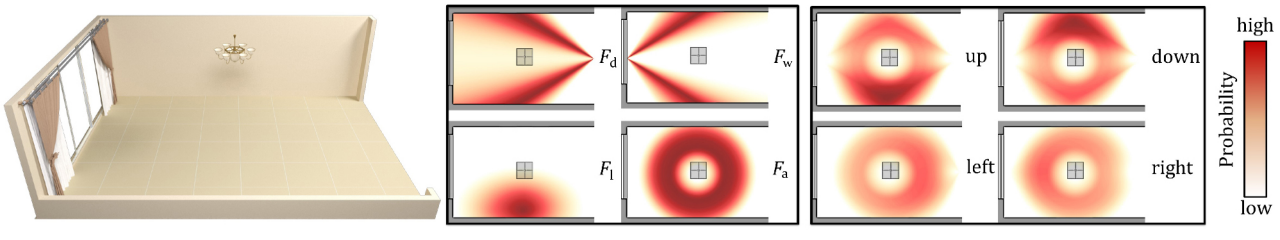


Fig. 5 Left: A room with activity label of *sitting and watching TV*. Middle: Assessment results with respect to the up direction of the associated object group for different room features (see Eq. (5)). Right: Compound assessment results for four different directions (see Eq. (6)). We conjointly normalize the four maps to reveal that the proper direction (up or down in this case) has extremal probability.

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 (darker red) that has lower energy is a better place for the objects. Assessment with different directions can also ascertain the proper direction of the placed object group (see Fig. 5(right)). Our method can also 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 Fig. 6, the room with multiple windows in Fig. 6(c) can be decomposed into two rooms with a single window (Figs. 6(a) and 6(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 Fig. 6(c). Analogously, rooms with multiple windows/doors can be assessed by our method.

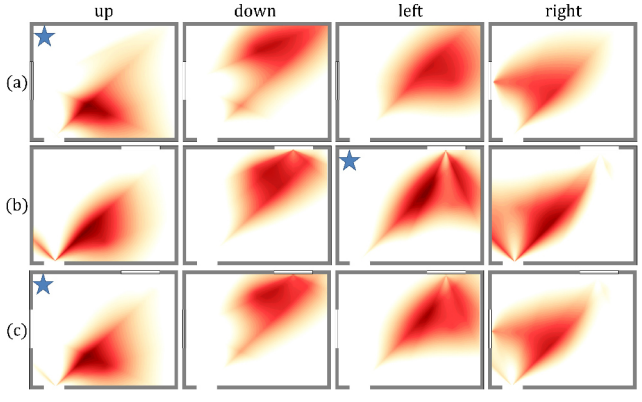


Fig. 6 Assessment results of activity *sitting and watching TV* for the four directions for three given rooms. The suggested direction in each row is marked by a blue star.

5.2 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., a downlight), and then assigning one or multiple activity labels to each room. Based on the assessment of an input room, our system can find appropriate positions and directions for the object group, as well as its member objects, with proper areas from our object database (collected from the well-known 3D Warehouse [32]). Our system can then generate 2D floor plans by applying 2D projections of these objects. We can easily transfer 2D plans to 3D scenes, using the 3D information for the objects in the database. Some furniture types like beds and TV sets are snapped to the wall near the suggested position to constrain the layout.

An input room can have multiple activity labels for the placement of more than one object group. As illustrated in Fig. 7, once an object group has been placed based on the first assigned activity label, areas that have been occupied or are 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 our method so far focuses on single room modeling, large indoor scenes with multiple rooms can also be tackled by our method room by room. Moreover, to relieve the workload of manually specifying activity labels, we can adopt a similar strategy for adaptive indoor scene modeling to Ref. [8], in which the area ratio between objects and room is used to measure whether more objects could be placed in a room. In this way, the user could make a long list of activity labels, but how many labels are available depends on the size of the given room: 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 6.

6 Results and discussion

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

6.1 Modeling results

In Fig. 8, we show synthesized indoor scenes along with assessment results of the corresponding rooms computed by our method, with respect to the user-specified activity labels. In Fig. 8(a), we choose assessment results with four different directions to place the same object group in a large room. We can see from the energy map (Fig. 8(middle)) that the four directions have different probability distributions

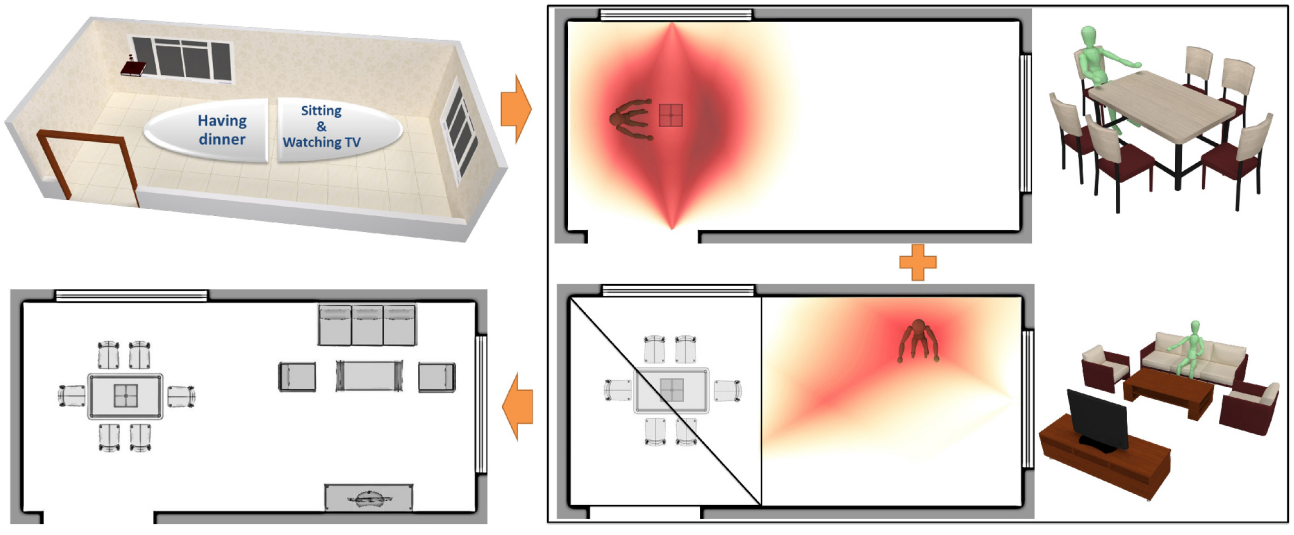


Fig. 7 Scene synthesis with multiple activity label inputs. After the object group for the first activity label is placed in the room (above right), the occupied area is then masked for assessment for the next activity label (below right).

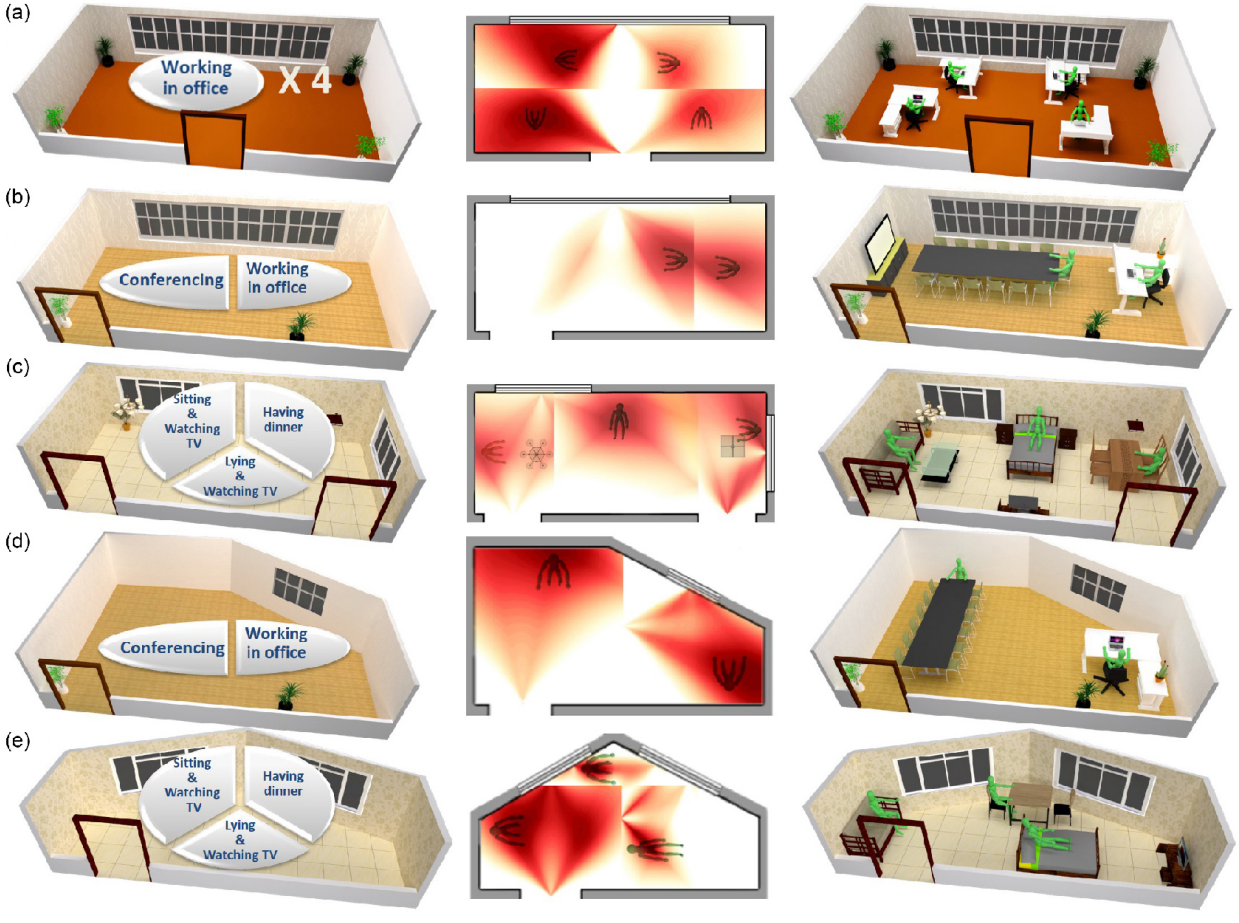


Fig. 8 Gallery of synthesized indoor scenes. In each case we show the input empty room with the assigned activity label(s), assessment results of the suggested directions in 2D projection, and the 3D scene generated by our method.

and suggested positions. Note that we intend to show the relations between the suggested positions of objects and the four directions we considered in this case. For the synthesis of large rooms with multiple objects of the same kind, symmetry should be considered, e.g., mirroring half or quarter of the

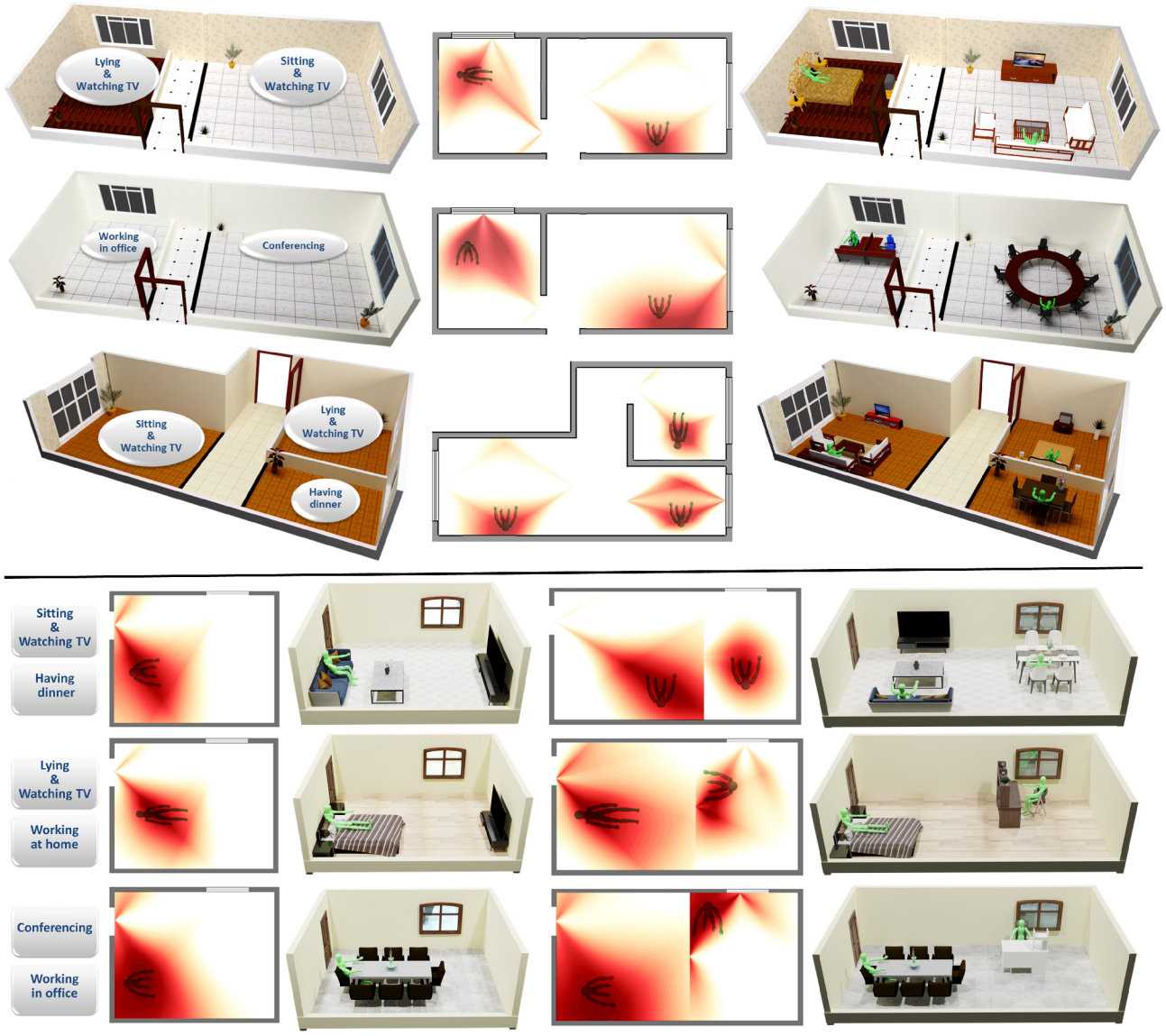


Fig. 9 Above: large indoor scenes produced by combining per-room modeling results. Below: indoor scene modeling with adaptive groups of objects. In each case, we show activity labels and synthesized scenes with small and large room inputs.

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 (Figs. 8(d) and 8(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 to synthesise plausible non-rectangular scenes.

In Fig. 9(above), using three cases we show how to combine multiple rooms into larger indoor scenes. For each case, indoor scene modeling is conducted per room. Fig. 9(below) 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 (Fig. 9(left)) can be used for indoor scene modeling. Note that the TV set was manually removed in the large room of the middle case to provide a better passageway. In this manner, our method can be used for both interactive and automated indoor scene modeling.

Our metrics for the room assessment with respect to certain object groups can also be used to evaluate indoor scenes. For example, in Fig. 10, we normalize the compound assessment result of the object group in each scene in the range of 0–1. The scenes on the

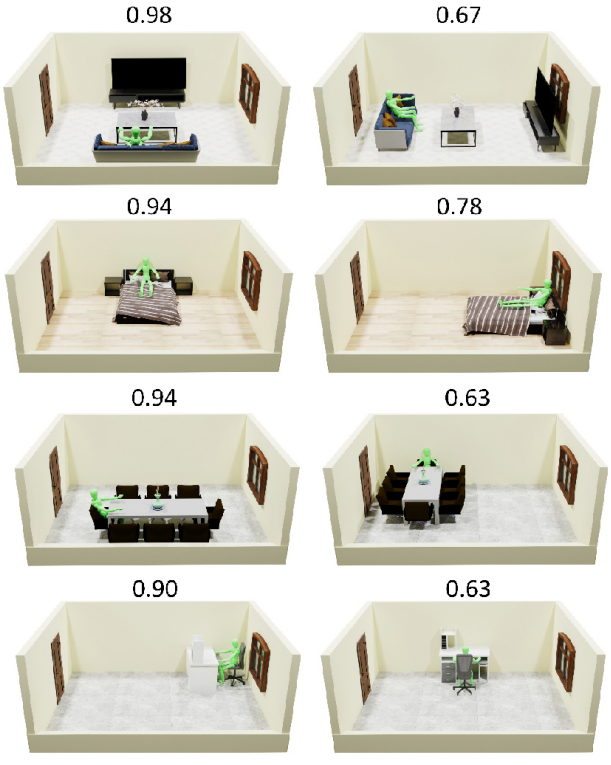


Fig. 10 Indoor scenes evaluated by our metrics, and the assessment scores, with respect to the layouts of the groups of objects.

left have better paths according to the door positions (the first and third cases), do not block the windows (the first and second cases), and better privacy (the last office). We can see that scenes with good layouts have high scores. This demonstrates that our method can refine large-scale indoor scene datasets by filtering out examples with low assessment scores. On average, the generation of one indoor scene takes 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 a 1.80 GHz Intel Core i7-8550U CPU and 8 GB RAM.

6.2 Evaluation

In Eq. (5), we set $A(s_i) = M_C(s_i)$ to encourage the high-quality indoor scene examples to play a more important role than 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 Fig. 11, energy maps in the first row were directly calculated from Eq. (5), while energy maps in the second row were calculated by setting $A(s_i) = 0$ in Eq. (5). The four energy maps in each row correspond to up, down, left, and right directions for placing object groups, respectively. In

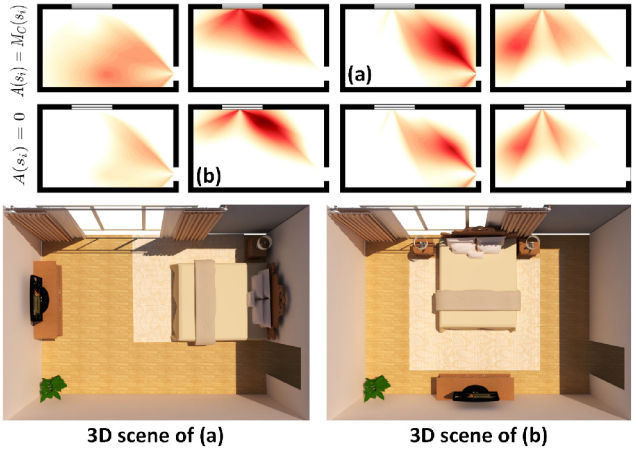


Fig. 11 Above: calculated energy maps for four directions, given the same room in an ablation experiment. Below: 3D scenes corresponding to energy maps (a) and (b).

Fig. 11(below), we show two 3D scenes corresponding to energy maps (a) and (b). Since the position of the window might lead to a backlighting problem in scene (b) when watching TV, scene (a) has a relatively better layout than scene (b). If we do not weight the dataset indoor scenes (i.e., set $A(s_i) = 0$), Eq. (5) is unlikely to generate scenes like (a) when there are fewer scene examples similar to (a) than similar to (b): our method that weights the dataset is effective for the small dataset with differentiated examples.

We further conducted a user study to demonstrate that the dataset weights (the fuzzy measurement $M_C(s_i)$ in Eq. (1)) are consistent with subjective perception. Since our method can provide commonly-seen layouts in the real world for most residential scenes, we only conducted the user study on office scenes that have varied layouts. We recruited 10 volunteers (postgraduate students) as two groups (5 participants in each group) to evaluate two office scenes (see Fig. 12(left)) in terms of six different positions for the activity *office work*, 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 to the participants' scores, we proportionally mapped our results to the range between the maximum and minimum participants' scores. Our assessment results and the participants' are shown in Fig. 12(right). Even though the perceptions are too subjective to be precisely measured, the comparison suffices to show

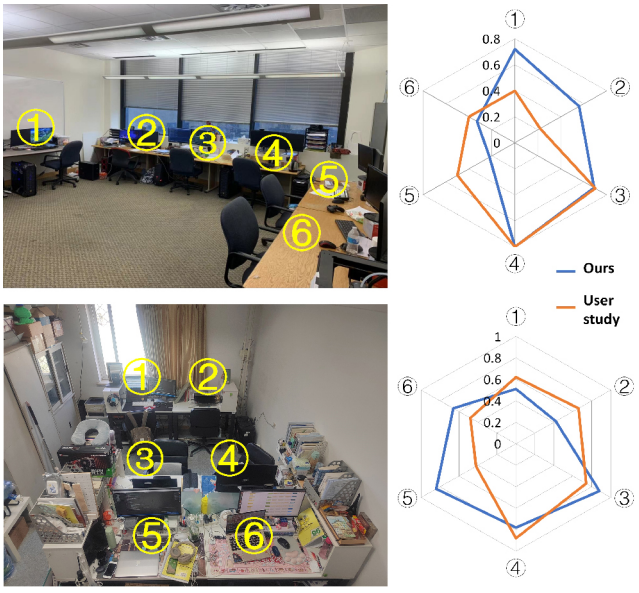


Fig. 12 Human subjective assessments of two offices with six desks in the real-world and the corresponding assessment results of our method.

that the priors extracted from our dataset are similar to those obtained from real-world experience.

We further compared our modeling results with those from other state-of-the-art indoor scene modeling methods [8, 10, 14] to demonstrate the effectiveness of our method. In Fig. 13(above left), we compare our method to an activity-centric method [8]. For two given rooms (one small, one large), we used *lying and watching TV + working at home* and *bed + bookshelf* as the input labels for our method and the method of Ref. [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 Ref. [8] uses activity-associated object relation graphs to determine appropriate object categories in the room, while our method simply uses the order of the specified activity labels as priorities for choosing the associated pre-defined object groups. Limited by the pre-defined object groups, some indirect interactive objects (e.g., a bookshelf) are not in any object groups considered by our method, so the explored objects in our results are fewer than in Ref. [8]. However, the method of Ref. [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 Fig. 13(right), we compare our results to those of Ref. [10]. Its model is based on a convolutional neural network and trained using a large-scale indoor scene database [33]. The comparison shows that both methods can generate plausible indoor layouts for the given room. Note that the given room may not be similar to any scene examples in the datasets used in the two methods. In Fig. 13(below left), using rectangular rooms, we compare indoor scenes synthesized by our method and by the one in Ref. [14]. Since the latter does not consider the impact of windows or doors, even though their generated scenes may have richer content, our results appear more appropriate for the environment of the given room. Comparing layouts from our methods and those in Refs. [10, 14] in terms of the given environment, furniture in their methods may block the window while in ours that does not happen, e.g., see Fig. 13(right, 2nd row) and Fig. 13(below left, first column). Our results can have better paths to the doorway that do not impact the activity of the object groups, e.g., Fig. 13(right, 3rd row), Fig. 13(below left, 2nd column). 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 can be more flexible than directly using hard constraints such as setting the relevant positions/directions between object groups and windows/doors. However, the methods of Refs. [10, 14] can make more layout variations in their synthesized indoor scenes, benefiting from the large-scale training datasets.

We further conducted a user study to compare the quality of the generated layouts from our method and the above two methods [10, 14]. 17 participants (undergraduate students in digital media technology) were recruited to compare 16 pairs of scenes (8 pairs for our method and Ref. [10], and 8 for our method and Ref. [14]), and asked to choose the better layout from each pair (order of the scenes in each pair was randomized). In the first case, in 136 comparisons (17×8) our results were preferred 67 times (49.3%), and those of Ref. [10], 69. In the second case, our results were preferred 70 times (51.5%) in 136 comparisons, and those of Ref. [14], 66. These results demonstrate that our method can generate indoor layouts with comparable quality to the two

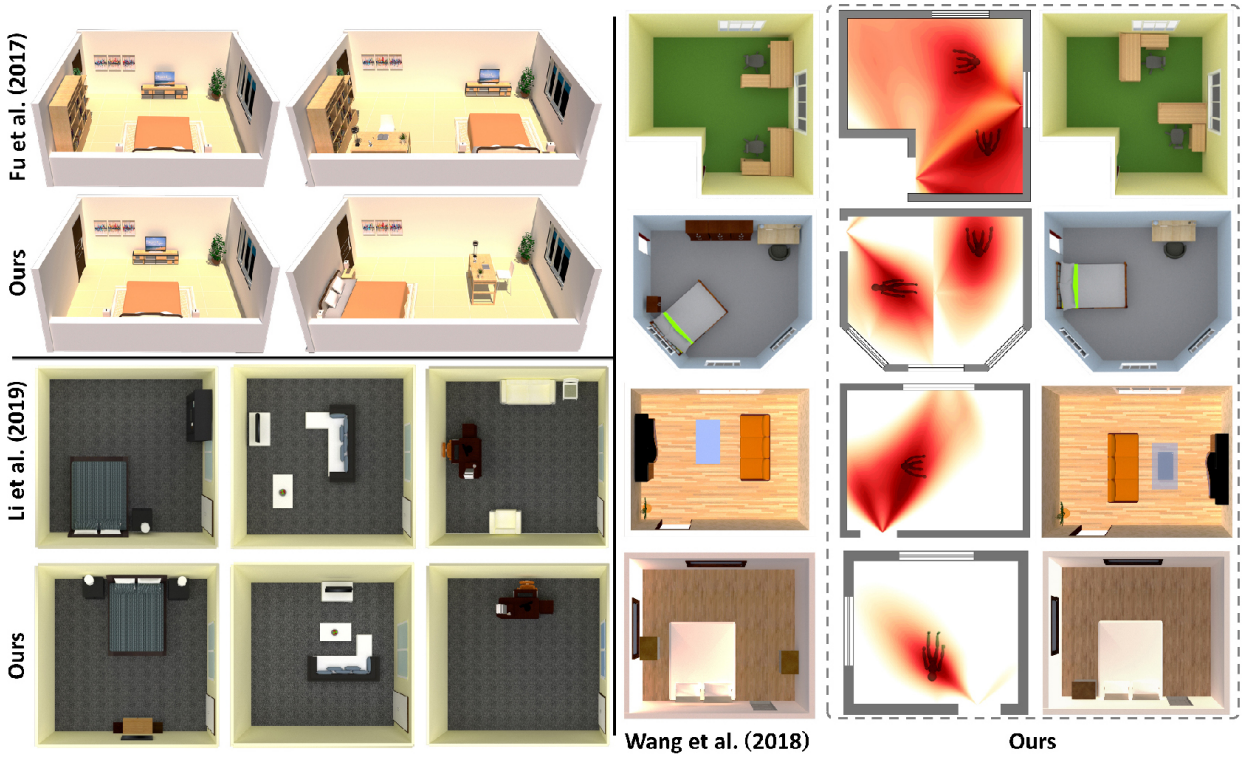


Fig. 13 Comparisons of our results to synthesized indoor scenes of Fu et al. (2017) [8] (above left), Wang et al. (2018) [10] (right), and Li et al. (2019) [14] (below left).

indoor scene synthesis methods in Refs. [10, 14]. It is noteworthy that, while state-of-the-art methods generally train deep-learning-based models using large indoor scene datasets, our method only utilizes much smaller datasets of differentiated examples for indoor scene synthesis, yet can produce comparable results.

6.3 Limitations

Our current approach has several limitations. Firstly, our method is based on several simplifying assumptions: utilizing doors and windows as spots in the room features, ignoring the cross-effects between object groups when collecting indoor scene datasets, and assuming the impacts of windows/doors/artificial lights on room assessment are additively linear.

Secondly, 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. In future, we plan to study transferability of subjective evaluations between different object groups. Our method furthermore needs pre-defined relationships within an object group. User assistance is still needed to manually place some decorative items like pot plants and murals 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 (relative positions/directions) for these objects in a group. To address the placement of such objects and make more variations of in-group object arrangement, we could use flexible object relations that contain multiple examples of in-group object arrangement. For example, in Fig. 14(above), we could pre-set four object relations for adding the bookshelf to the object group associated with *working at home*, while the user decides which one is better for the synthesized indoor scene.

Thirdly, since the activity labels are assigned by users, improper user-specified object groups might lead to unnatural results (e.g., a dining room with a bed). This can be avoided by employing priors such as relation graphs which indicate the co-existing possibilities between different object groups.

Lastly, since we only consider four different directions, our current implementation can only generate axis-aligned layouts even for non-rectangular rooms (see Figs. 8(d) and 8(e)). However, for non-axis-aligned rooms or rooms with curved walls, better

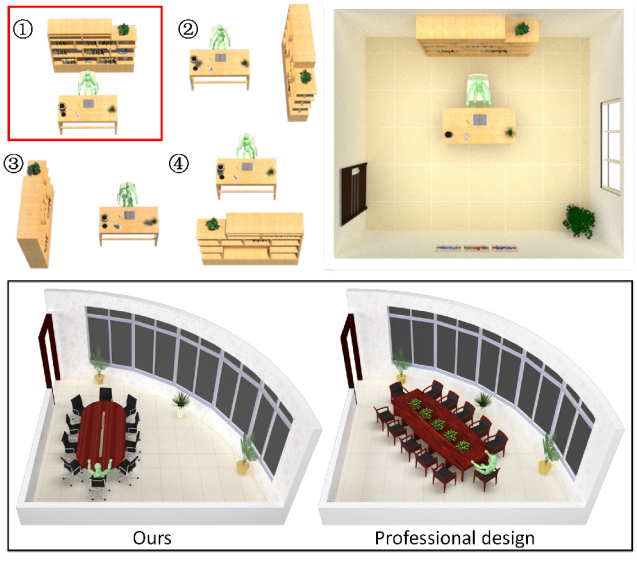


Fig. 14 Above: example that adds a bookshelf to the *working at home* group with four pre-set relations (left). Users can select one of them (e.g., in the red box) to synthesize the indoor scene (right). Below: our result and professional design for a room with a curved side.

object arrangement solutions may exist than axis-aligned layouts. For example, given a room with a curved bay window and the assigned activity of conferencing, we show a scene synthesized by our method (Fig. 14(below left)) and one provided by an interior designer (Fig. 14(below right)). Our method suggests a small conference table, 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 allows the user to slightly adjust the position and direction of objects to refine the indoor layout.

7 Conclusions

In this paper, we have presented a new approach to using datasets of differentiated examples to support indoor scene modeling. To construct such datasets, we perform subjective comparisons on specially-designed indoor scenes with different room features for 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 appropriate positions and directions for the placement of the object groups 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. Our ideas can open up new research opportunities in example-driven indoor scene modeling.

In future, we plan to further extend our room features to tackle more types of factors including color, decoration, and furniture styles, to handle rooms with more complex shapes (e.g., with curved walls), and to enable the automation of designing more ergonomic indoor scenes for target activities. We are also interested in exploring cross-activity influences on indoor layouts to increase the practicality of our method, especially in scenarios in which an object group is associated with multiple activities. This involves prioritizing different activities, which could also be determined from 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 use of a limited number of labeled data for more types of indoor scenes.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work was partially supported by grants from the National Natural Science Foundation of China (61902032), Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11237116), and City University of Hong Kong (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, L. F.; Yeung, S. K.; Tang, C. K.; Terzopoulos, D.; Chan, T. F.; Osher, S. J. Make it home: Automatic optimization of furniture arrangement. *ACM Transactions on Graphics* Vol. 30, No. 4, Article No. 86, 2011.

[2] Merrell, P.; Schkufza, E.; Li, Z. Y.; Agrawala, M.; Koltun, V. Interactive furniture layout using interior design guidelines. In: Proceedings of the ACM SIGGRAPH 2011 Papers, Article No. 87, 2011.

[3] Chen, X. W.; Li, J. W.; Li, Q.; Gao, B.; Zou, D. Q.; Zhao, Q. P. Image2Scene: Transforming style of 3D room. In: Proceedings of the 23rd ACM International Conference on Multimedia, 321–330, 2015.

[4] Fisher, M.; Ritchie, D.; Savva, M.; Funkhouser, T.; Hanrahan, P. Example-based synthesis of 3D object arrangements. *ACM Transactions on Graphics* Vol. 31, No. 6, Article No. 135, 2012.

[5] Jiang, Y.; Lim, M.; Saxena, A. Learning object arrangements in 3D scenes using human context. In: Proceedings of the 29th International Conference on International Conference on Machine Learning, 907–914, 2012.

[6] Fisher, M.; Savva, M.; Li, Y. Y.; Hanrahan, P.; Nießner, M. Activity-centric scene synthesis for functional 3D scene modeling. *ACM Transactions on Graphics* Vol. 34, No. 6, Article No. 179, 2015.

[7] Savva, M.; Chang, A. X.; Hanrahan, P.; Fisher, M.; Nießner, M. PiGraphs: Learning interaction snapshots from observations. *ACM Transactions on Graphics* Vol. 35, No. 4, Article No. 139, 2016.

[8] Fu, Q. A.; Chen, X. W.; Wang, X. T.; Wen, S. J.; Zhou, B.; Fu, H. B. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM Transactions on Graphics* Vol. 36, No. 6, Article No. 201, 2017.

[9] Qi, S. Y.; Zhu, Y. X.; Huang, S. Y.; Jiang, C.; Zhu, S. C. Human-centric indoor scene synthesis using stochastic grammar. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5899–5908, 2018.

[10] Wang, K.; Savva, M.; Chang, A. X.; Ritchie, D. Deep convolutional priors for indoor scene synthesis. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 70, 2018.

[11] Wang, K.; Lin, Y.; Weissmann, B.; Savva, M.; Chang, A. X.; Ritchie, D. PlanIT: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics* Vol. 38, No. 4, Article No. 132, 2019.

[12] Saaty, T. L. *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.

[13] Klir, G. J.; Yuan, B. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, 1995.

[14] Li, M. Y.; Patil, A.; Xu, K.; Chaudhuri, S.; Khan, O.; Shamir, A.; Tu, C. H.; Chen, B. Q.; Cohen-Or, D.; Zhang, H. GRAINS: Generative recursive autoencoders for INdoor scenes. *ACM Transactions on Graphics* Vol. 38, No. 2, Article No. 12, 2019.

[15] Fisher, M.; Hanrahan, P. Context-based search for 3D models. *ACM Transactions on Graphics* Vol. 29, No. 6, Article No. 182, 2010.

[16] Fisher, M.; Savva, M.; Hanrahan, P. Characterizing structural relationships in scenes using graph kernels. *ACM Transactions on Graphics* Vol. 30, No. 4, Article No. 34, 2011.

[17] Sharf, A.; Huang, H.; Liang, C.; Zhang, J. P.; Chen, B. Q.; Gong, M. L. Mobility-trees for indoor scenes manipulation. *Computer Graphics Forum* Vol. 33, No. 1, 2–14, 2014.

[18] Xu, K.; Ma, R.; Zhang, H.; Zhu, C. Y.; Shamir, A.; Cohen-Or, D.; Huang, H. Organizing heterogeneous scene collections through contextual focal points. *ACM Transactions on Graphics* Vol. 33, No. 4, Article No. 35, 2014.

[19] Liu, T. Q.; Chaudhuri, S.; Kim, V. G.; Huang, Q. X.; Mitra, N. J.; Funkhouser, T. Creating consistent scene graphs using a probabilistic grammar. *ACM Transactions on Graphics* Vol. 33, No. 6, Article No. 211, 2014.

[20] Zhang, S. H.; Zhang, S. K.; Xie, W. Y.; Luo, C. Y.; Yang, Y. L.; Fu, H. B. Fast 3D indoor scene synthesis by learning spatial relation priors of objects. *IEEE Transactions on Visualization and Computer Graphics* Vol. 28, No. 9, 3082–3092, 2022.

[21] Ma, R.; Li, H. H.; Zou, C. Q.; Liao, Z. C.; Tong, X.; Zhang, H. Action-driven 3D indoor scene evolution. *ACM Transactions on Graphics* Vol. 35, No. 6, Article No. 173, 2016.

[22] Ma, R.; Patil, A. G.; Fisher, M.; Li, M. Y.; Pirk, S.; Hua, B. S.; Yeung, S. K.; Tong, X.; Guibas, L.; Zhang, H. Language-driven synthesis of 3D scenes from scene databases. *ACM Transactions on Graphics* Vol. 37, No. 6, Article No. 212, 2018.

[23] Xu, K.; Chen, K.; Fu, H.; Sun, W. L.; Hu, S. M. Sketch2Scene: Sketch-based co-retrieval and co-placement of 3D models. *ACM Transactions on Graphics* Vol. 32, No. 4, Article No. 123, 2013.

[24] Chen, K.; Lai, Y. K.; Wu, Y. X.; Martin, R.; Hu, S. M. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information. *ACM Transactions on Graphics* Vol. 33, No. 6, Article No. 208, 2014.

[25] Shao, T.; Monszpart, A.; Zheng, Y.; Koo, B.; Xu, W.; Zhou, K.; Mitra, N. J. Imagining the unseen: Stability-based cuboid arrangements for scene understanding. *ACM Transactions on Graphics* Vol. 33, No. 6, Article No. 209, 2014.

[26] Zhang, S. K.; Li, Y. X.; He, Y.; Yang, Y. L.; Zhang, S. H. MageAdd: Real-time interaction simulation for scene synthesis. In: Proceedings of the 29th ACM International Conference on Multimedia, 965–973, 2021.

[27] Frontczak, M.; Wargocki, P. Literature survey on how different factors influence human comfort in indoor environments. *Building and Environment* Vol. 46, No. 4, 922–937, 2011.

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

[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* Vol. 77, 67–79, 2014.

[30] Ochoa, C. E.; Capeluto, I. G. Evaluating visual comfort and performance of three natural lighting systems for deep office buildings in highly luminous climates. *Building and Environment* Vol. 41, No. 8, 1128–1135, 2006.

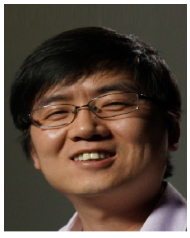
[31] Zhang, Z. W.; Yang, Z. P.; Ma, C. Y.; Luo, L. J.; Huth, A.; Vouga, E.; Huang, Q. X. Deep generative modeling for scene synthesis via hybrid representations. *ACM Transactions on Graphics* Vol. 39, No. 2, Article No. 17, 2020.

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

[33] Song, S. R.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; Funkhouser, T. Semantic scene completion from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 190–198, 2017.

Qiang Fu is an associate professor in the School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications. He received his Ph.D. degree in computer science from Beihang University, China, in 2018 and his B.E. degree in automation from Beihang University in 2011. His research interests include computer graphics and virtual reality.

Shuhan He received her B.E. degreee 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 his Ph.D. degree in computer science from Hong Kong University of Science and Technology in 2007 and his B.S. 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 his B.E. degree in electronics engineering from the University of Science and Technology of China, in 1992, and his Ph.D. degree in electronics engineering from 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 his B.S. degree in mathematics from Xiamen University, China, his M.S. degree in computer science from Peking University, and his Ph.D. 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.