FairScene: Learning Unbiased Object Interactions for Indoor Scene Synthesis

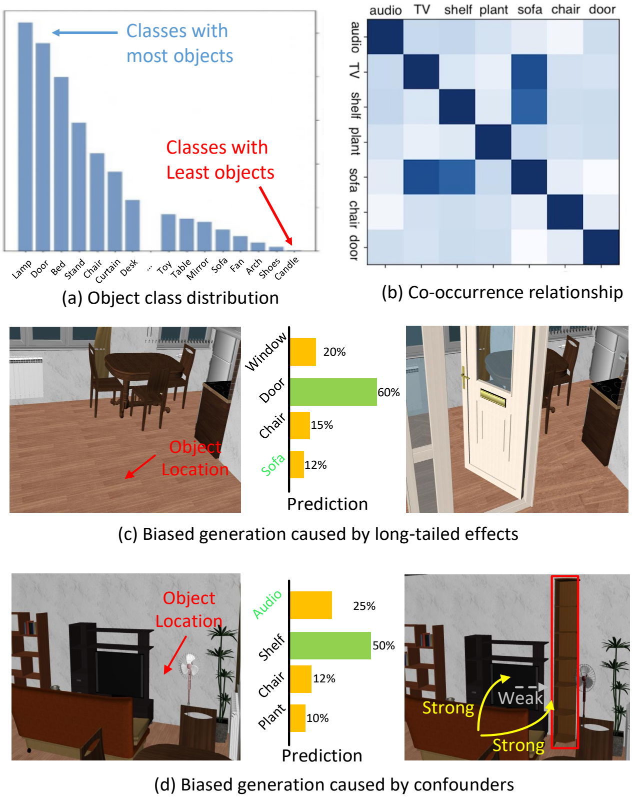
Zhenyu Wu, Ziwei Wang, Shengyu Liu, Hao Luo, Jiwen Lu, and Haibin Yan

***Abstract*—In this paper, we propose an unbiased graph neural network learning method called FairScene for indoor scene synthesis. Conventional methods directly apply graphical models to represent the correlation of objects for subsequent furniture insertion. However, due to the object category imbalance in dataset collection and complex object entanglement with implicit confounders, these methods usually generate significantly biased scenes. Moreover, the performance of these methods varies greatly for different indoor scenes. To address this, we propose a framework named FairScene which can fully exploit unbiased object interactions through causal reasoning, so that fair scene synthesis is achieved by calibrating the long-tailed category distribution and eliminating the confounder effects. Specifically, we remove the long-tailed object priors subtract the counterfactual prediction obtained from default input, and intervene in the input feature by cutting off the causal link to confounders based on the causal graph. Extensive experiments on the 3D-FRONT dataset show that our proposed method outperforms the state-of-the-art indoor scene generation methods and enhances vanilla models on a wide variety of vision tasks including scene completion and object recognition.**

***Index Terms*—Indoor scene synthesis, graph neural networks, causal inference, counterfactuals, intervention.**

I. Introduction

Recent years have witnessed the increasing requirements for virtual models in 3D indoor scenes, because great progress has been made in virtual and augmented reality (VR/AR) [10], [41], [61], robot navigation [11], [16], [6], interior design [8], [19], [64], or can even create synthetic training datasets for other indoor scene understanding tasks [17]. Indoor scene synthesis has aroused extensive interest in computer vision and robotics due to the great efficiency in data collection. Given an empty interior space with geometrical constraint including the floor, ceiling, and walls, the indoor scene synthesis aim to reasonably select the furniture and appliance for arrangement. However, the indoor scene design requires experienced architects to spend a couple of days completing the complex task. Therefore, it is desirable to automatically synthesize indoor scenes with the structural knowledge learned from the annotated 3D indoor scenes, and can significantly reduce the cost of producing datasets for indoor scene understanding tasks.



(a) Object class distribution

(b) Co-occurrence relationship

(c) Biased generation caused by long-tailed effects

(d) Biased generation caused by confounders

Fig. 1. Example of dataset bias visualization. (a) The object class distribution in SUNCG, where the class with most objects contains about as 100× samples as that with least objects. (b) The co-occurrence relationship among different classes, where darker colors mean stronger correlation. (c) The biased generation caused by long-tailed effects shown in (a). Doors are added with the highest probability than the groundtruth class sofa because doors frequently appear in the training set. (d) The biased generation caused by the confounder object sofa, where a speaker should be placed in the object location due to the existence of TV. Because of the strong co-occurrence relationship between sofa and TV and that between sofa and shelf, the shelf is added with the highest probability.

Due to the strong discriminative power and generalization ability of deep learning [18], [15], the deep indoor scene synthesis models [37], [25], [66], [23], [33], [48] have been widely studied and yielded promising results. Those methods utilize convolutional neural network (CNN) to learn the rich visual semantics in annotated data [49], [37] or leverage graph neural network (GNN) to mine the complicated object structures in the training samples [66], [25], [48]. However, conventional methods employ simple relationships such as "co-occurrence" that are insufficient to construct the layout of complex indoor scenes. Some recent approaches further refine the relationships between furniture objects such as "supporting", and "surrounding" to enhance the contextual

Zhenyu Wu, Luo Hao, and Haibin Yan are with the School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: {wuzhenyu, haroll_luo,eyanhaibin}@bupt.edu.cn

Ziwei Wang, Shengyu Liu, and Jiwen Lu are with the Beijing National Research Center for Information Science and Technology (BNRist), and the Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: {wang-zw18, liusheng17}@mails.tsinghua.edu.cn, lujiwen@tsinghua.edu.cn.

information between objects and avoid confusion caused by oversimplified relationships. However, the current approaches have ignored the observation bias caused by the long tail of dataset categories and scene confounding, which will lead to poor realistic generated scenes. Since the 3D indoor scene datasets are collected in natural distribution, unconstrained learning with those annotated data usually leads to biased prediction in two aspects. The first is the imbalanced object categories in long-tailed distribution, which will significantly bias the learned model to higher occurrence categories. Generating room layouts will also overfit with several object categories, which reduces the realism of the synthesized scene. The second is the confounder objects in the complex scenes. Due to biases in dataset annotation, common knowledge between furniture objects can be confounded by observed contextual biases. As shown in Fig. 1 (a) and (c), the number of doors is much more than the sofa, so the learned model tends to select the former with higher probability even without any significant context. Fig. 1 (b) and (d) demonstrate the confounder objects in the complex scenes. Although the causality link between the TV and the shelf is weak, they are strongly correlated due to their co-occurrence with the confounder object sofa. As a result, the two types of biases degrade the plausibility of synthesized indoor scenes with unreasonable object layouts.

In this paper, we present a FairScene method to learn unbiased graph neural networks for automatic indoor scene synthesis. Unlike conventional methods confounded by dataset bias which directly leverage graphical models to represent the object correlation for subsequent furniture and appliance insertion, our method mines the object interaction by causal inference instead of correlation learning further to reduce contextual bias. Therefore, the object category imbalance and the confounder effects are eliminated based on the causal graph, so that unbiased indoor scenes are synthesized with higher plausibility. More specifically, we leverage GNN with message passing to capture the structural information contained in the context, where pre-defined relations including ”supporting”, ”surrounding” and ”co-occurring” are mined with recurrent modules. Then we remove the long-tailed object priors in the learned messages by subtracting the counterfactual messages, which is acquired from the default input defined as average node representation. Finally, we intervene the input scene context by constrained object prediction to eliminate the confounder effects, which is achieved via cutting off the causal link to the confounders. Hence, unbiased indoor scenes are generated with enhanced plausibility for downstream tasks. We conduct extensive experiments on a wide variety of downstream tasks including scene completion, object query and object recognition with the 3D-FRONT dataset [42], and the results show that our FairScene outperforms the state-of-the-art indoor scene synthesis methods by a sizable margin.

In summary, our work consists mainly of the following contributions:

1) We propose an unbiased indoor scene synthesis framework based on causal inference named FairScene, which effectively reduces prediction bias in layout generation.
2) We eliminate the long-tail effect caused by the unbalanced distribution of object classes in the dataset by constructing counterfactual samples and further eliminate the confounding effect through the intervention approach.
3) We evaluated FairScene on the 3D-FRONT public dataset, and the extensive experimental results demonstrate the effectiveness of the proposed approach.

II. Related Work

In this section, we briefly review two related topics, including indoor scene synthesis frameworks and causal inference.

A. Indoor Scene Synthesis

Indoor scene synthesis has received significant attention due to alleviating the cost of complex and tedious handcrafted scenes. According to the approach of parsing the relationship between objects in a room, current frameworks can be divided into three main categories: rule-based indoor scene synthesis, non-parametric model indoor scene synthesis, and parametric model indoor scene synthesis.

Rule-based indoor scene synthesis is aimed at parsing room layouts based on handcrafted features and retrieves objects from the dataset to place into the scene. Early work [57] applied the rule-based constraint to generate 3D object layouts, which can quickly rearrange existing scenes by resetting the preferences for placing indoor objects. Merrell *et al.* [26] incorporated interior design principles as terms in the density function and generated layout recommendations by sampling the density function through a Monte Carlo algorithm. Fu *et al.* [13] constructed an activity-related object relationship graph by calculating the coexistence frequency of objects in indoor scenes to generate reasonable scene layout suggestions. Inspired by the work of [5], several works focusing on mining prior knowledge of datasets using co-occurrence analysis and statistical models have been proposed. Weiss *et al.* [52] presented a physically driven framework for continuous scene synthesis, avoiding the slow and inefficient sampling from different layout configurations caused by traditional methods of random optimization. Xiong *et al.* [55] applied various mechanical constraints to initialize the paths of scene objects to target locations and constructed an interior scene system. Zhang *et al.* [63] first proposed a strategy for representing indoor scene layouts with coherent sets and introduced layout properties to build a framework for scene composition with human interests. Zhang *et al.* [62] accomplished employing a spatial randomness (CSR) test to measure the strength of object spatial relationships and generate room layouts accurately based on the discrete prior provided by the sample, the proposed method without the need to fit models.

Non-parametric model indoor scene synthesis aims to capture the relationships between furniture by statistically fitting a scene prior to the dataset. Early work Yu *et al.* [59] extracted the hierarchical and spatial relationships of each furniture object in the current set to optimize realistic furniture layouts by incorporating the statistical relationships between objects into the objective function. In order to learn object priors from the annotated data, data-driven indoor scene synthesis was presented. Chen *et al.* [7] ensured semantic compatibility between the generated object model so that scene layouts can be generated from sparse low-quality RGB-D image inputs. Liu *et al.* [24] further advanced the development of the image-based synthesis of indoor scenes by incorporating segmentation mask information into the scene generation framework, fully utilizing the correspondence between RGB recognition results and 3D models. Subsequent variants exploit the co-occurrence or alignment of furniture objects with various primitive models by different approaches. Yang *et al.* [58] normalized the output of the synthetic model using parameter prior distributions captured from the training dataset, and corrected infeasible furniture layouts by predicting consistency constraints between attributes. Qi *et al.* [36] applied a probabilistic grammar model to resolve the indoor scenes into attribute spaces with or graph S-AOG and sampled the new layout using a Monte Carlo Markov chain.

Parametric indoor scene synthesis primarily focuses on learning the relationship between furniture from the training samples through the deep learning model. With the emergence of large-scale indoor scene datasets (e.g. SUNCG [42] and 3D-FRONT [12]) and the achievements of deep learning [18], deep neural network based methods [23], [25], [33], [37], [48], [66] were widely studied in recent years. Wang *et al.* [49] and Ritchie *et al.* [37] encoded the scene images in top-down views, and decoded the representation to sequentially decide the existence, category, location, orientation and dimension of new objects. Zhang *et al.* [65] utilized the Variational AutoEncoders (VAE) [22] coupled with Generative Adversarial Networks (GAN) [15] to generate an object matrix where each column represented the object location and geometry attributes. Ostonv *et al.* [27] approximated room layout generation as a proximal policy optimization problem, and estimated the reward of placing objects at each iteration with deep reinforcement learning. Paschalidou *et al.* [29] proposed a novel autoregressive transformer architecture that approximates scene synthesis and in-room object generation as sequence and unordered generation. In order to fully leverage the rich structural information in the context, graphical methods were proposed to mine the complex object correlation. Zhou *et al.* [66] and Wang *et al.* [48] proposed a graph neural network for indoor scene synthesis where the edge depicted the spatial and semantic correlation between objects. Gao *et al.* [14] proposed a hierarchical graphical network for synthesizing interior scenes, directly synthesizing the fine geometry of room layouts and furniture at a hierarchical level, and employed functional areas as intermediate agents for rooms and furniture to further ensure rationality.

Nevertheless, directly applying the graphical models to represent object correlation ignores the object category imbalance and confounder effects in the dataset, leading to significantly biased scene generation with unreasonable arrangements.

B. Causal Inference

In recent years, research based on causal inference has been active in various fields. We first briefly introduce the widely utilized latent outcome model and structural causal model, and finally outline the application of causal inference to computer vision.

The core of the potential outcome model is to compare the effects of receiving the intervention and not receiving the intervention for the same research object. Rubin [39] improved the counterfactual inference framework by introducing an allocation mechanism to describe the event. Imbens *et al.* [20] defined potential outcomes as each pair of ”intervention-outcomes”. Rubin [40] defined causal effects as differences in potential outcomes of the same research object. The potential outcome model argues that omitting covariates in observational research can lead to serious causality inference bias. Bickel [4] defined variables that affect the relationship measure between two other variables as confounders and further concluded that confounding effects in the conventional inference framework. Rosenbaum [38] proposed the inverse probability weighted (IPW) estimation method to effectively eliminate the bias due to the different distributions of covariates. In addition, the stratification and matching methods can also eliminate the confounding bias.

Structural causal models can visually represent the causal relationships between multiple variables. Pearl [31], [32] proposed the concept of external intervention based on Bayesian networks, which further extended the formal representation of causal relationships and inspired subsequent methods for mining causal relationships from data. Based on the development of human perception of things, [31] divided causal relationships into three levels: association, intervention, and counterfactual. The association is mainly represented by the statistical correlation defined by the data, i.e., the relationship between the joint distribution probabilities of the individual variables. The intervention is expressed as changing the variable data distribution, [30] proposed using the do operation to represent, which is applied mainly through the Correction Formula, the Backdoor Criterion, and the Frontdoor Criterion. The counterfactual is expressed as retroactive to the estimated relationship between the variables. For counterfactual inputs, structural causal models employ minimal interventions to satisfy the proposed hypotheses and predict outcomes based on past perceptions with added conditions.

The causal analysis has been proven to be very effective in numerous research fields such as removing spurious bias [2], [44], image super-resolution [21], disentangling model effects [3], [67] and acquiring generalizable features [28], [51]. Recently, causal inference has been widely adopted in computer vision. Tang *et al.* [45] debiased the scene graph generation by drawing the counterfactual causality in the defined causal graph, and utilized the Total Direct Effect to remove the long-tailed priors. Wang *et al.* [50] proposed Visual Commonsense R-CNN for unbiased object detection, where the causal link directed to the confounder was cut off and the model learned sense-making knowledge instead of common co-occurrence. Abbasnejad *et al.* [1] explored the bias in the dataset and enhanced the generalization ability in visual question answering (VQA) by counterfactual generation. Zhang *et al.* [60] yielded pixel-level masks by only using the image-level labels via the structural causal model analysis among images, pixels, and labels. Xu *et al.* [56] employed counterfactual interventions to maximize the difference between unintentional and counterfactual intentional behaviors to enhance model

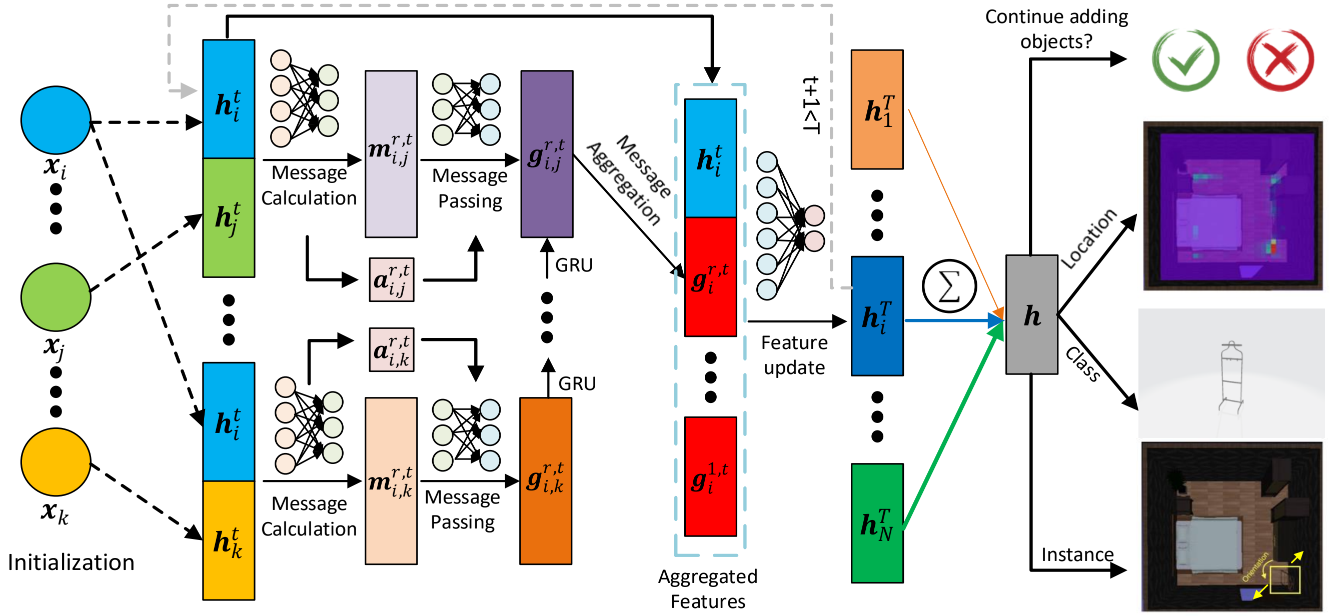


Fig. 2. The pipeline of our FairScene. The object properties are leveraged to initialize the node features, and the concatenation of two node features yields the messages between them and the corresponding attention. The GRU modules aggregate all messages directed to one node via the recurrent cells, and the aggregated features containing messages about various kinds of relationships are applied to predict the updated features for given node features. The feature update stops when achieving the iteration limit, where all final node features are summed with different weights to form the scene features. With the scene features, the existence, location, class, and placement (orientation and size) of additional objects are predicted.

recognition performance.

In this paper, we extend causality inference to generate input intervention and counterfactuals, through which the long-tailed object category distribution is calibrated and the confounder effects are eliminated in our unbiased indoor scene synthesis. Inspired by the outstanding performance of transformer network architecture on various visual tasks, recent work has started to capture the high-level relationships between objects with cross-attention mechanisms.

III. Approach

In this section, we first introduce the general pipeline of our graph neural networks for indoor scene synthesis. Then we present object category distribution calibration to remove the long-tailed bias and propose confounder effect elimination to obtain the unbiased indoor scene synthesis model. Finally, we demonstrate the application of our FairScene in a wide variety of vision tasks.

A. *Parsing furniture layout*

Following [48], [49], we sequentially add objects for indoor scene synthesis. We define six fine-grained relationships as in [66] to strengthen the structure discrimination ability including ”supporting”, ”supported by”, ”surrounding”, ”surrounded by”, ”next to” and ”co-occurring”. We identify the above six relationships between two objects (A, B) by measuring the 3D bounding boxes (A_{bbox}, B_{bbox}) and category of the furniture object, which are defined as follows:

”supporting” and ”supported by” : The ”supporting” relationship exists when the bounding box of one object is on top of the other. We further determine whether the center point of A is in the top projection of B_{bbox} to finalize the ”$B \quad supporting \quad A$” relationship. ”supported by” is the inverse relationship.

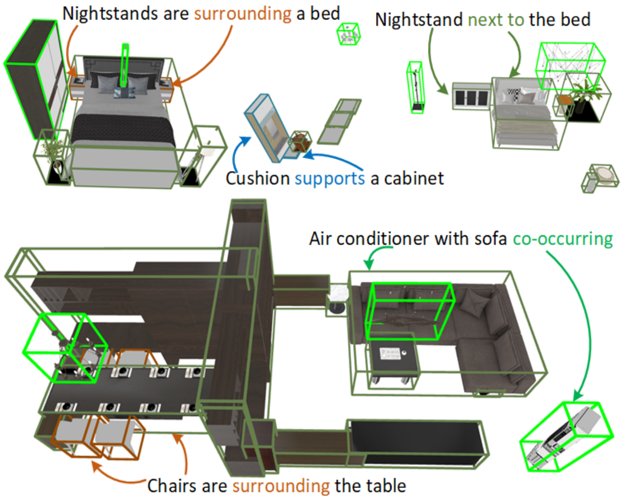

Nightstands are surrounding a bed
Nightstand next to the bed
Cushion supports a cabinet
Air conditioner with sofa co-occurring
Chairs are surrounding the table

Fig. 3. Room furniture layout analysis visualization results. We display the layout of three types of rooms: bedroom, dining room, and living room. The green bounding boxes represent ”co-occurrence” relationships, the brown bounding boxes represent the ”surrounding” relationship and the dark green bounding boxes represent the ”next to” relationship(Best view in color).

”surrounding” and ”surrounded by” : We traverse the whole scene with each object in the room as the center. For object A, we first set the query radius R according to the size of A_{bbox}, which can be described as:

$$\boldsymbol{R} = (W + H)/2 + \sqrt{W^2 + H^2}/2 \tag{1}$$

where W and H are the length and width of A_{bbox}, respectively. We further constrain the center points of the above set of objects to be in the same plane at the same time. Similarly, "surrounded by" is the reversed relationship of "surrounding".

"next to" and "co-occurring" : We finalize the "next to" relationship by determining whether the shortest distance between the vertices of two objects' bounding boxes is less than a threshold value(the setting in this paper is 0.2 meters). As for "co-occurring", all objects in one room can be considered to have this relationship.

Fig. 3 shows the results of the room layout parsing visualization, which further justifies the selected six relationships. Compared with previous methods, our approach precisely parses furniture fine-grained layout relationships. For example, the chairs and the tables are no longer in the usual "next to" or "co-occurring" relationships, but are defined as a more distinctive "surrounding". Meanwhile, for objects with significant "co-occurring", such as a bed and a nightstand, our "next to" relationship provides more precise spatial location guidance compared to "co-occurring.

B. Graph Networks for Indoor Scene Synthesis

Fig. 2 shows the general pipeline of the graph neural networks for indoor scene synthesis, which relies on message passing graph convolution [48], [66]. At the t_{th} step for sequential object addition, the i_{th} node the in graph represents the state of the i_{th} object with node feature $\boldsymbol{h}_i^t$ in the scene. The edge between the i_{th} and j_{th} nodes is denoted as e_{ij}^t that demonstrates the correlation between the i_{th} and j_{th} objects at the t_{th} step. The node feature is initialized with the function f_{init} as follows:

$$\boldsymbol{h}_i^0 = f_{init}(\boldsymbol{x}_i, \boldsymbol{w}_{init}) \tag{2}$$

where $\boldsymbol{x}_i \in \mathbb{R}$ is the input feature representing the object category, location, orientation, and size of the i_{th} object in the scene, and $\boldsymbol{w}_{init}$ means the parameters of the initialization function. For information propagation in the graph, we first calculate the edge message of the r_{th} relation from the i_{th} to the j_{th} node at the t_{th} step:

$$\boldsymbol{m}_{i,j}^{r,t} = f_{msg}(\boldsymbol{h}_i^t, \boldsymbol{h}_j^t; \boldsymbol{w}_{msg}) \tag{3}$$

where f_{msg} demonstrates the message function with the parameters $\boldsymbol{w}_{msg}$. The attention weights of the message $\boldsymbol{m}_{i,j}^{r,t}$ is denoted as $\boldsymbol{a}_{i,j}^{r,t}$, which is calculated by the attention function f_{att} with the parameters $\boldsymbol{w}_{att}$:

$$\boldsymbol{a}_{i,j}^{r,t} = f_{att}(\boldsymbol{h}_i^t, \boldsymbol{h}_j^t; \boldsymbol{w}_{att}) \tag{4}$$

Then the messages to each node are aggregated with GRU modules in the following way:

$$\boldsymbol{g}_{i,j}^{r,t} = f_{GRU}(\boldsymbol{g}_{i,j-1}^{r,t}, \boldsymbol{m}_{i,j}^{r,t}, \boldsymbol{a}_{i,j}^{r,t}; \boldsymbol{w}_{GRU}) \tag{5}$$

where $\boldsymbol{g}_{i,j}^{r,t}$ stands for the aggregated messages of the r_{th} relation from the i_{th} to the j_{th} node at the t_{th} step, and f_{GRU} means the aggregating function with the parameters $\boldsymbol{w}_{GRU}$. The last cell of the GRU module yields the aggregated features of the r_{th} relation for the i_{th} node at the t_{th} step, which is denoted as $\boldsymbol{g}_i^{r,t}$. Finally, the node feature $\boldsymbol{h}_i^t$ is updated via the following transition:

$$\boldsymbol{h}_i^{t+1} = f_{upd}(\boldsymbol{h}_i^t, \{\boldsymbol{g}_i^{r,t}\}_r; \boldsymbol{w}_{upd}) \tag{6}$$

where $\{\boldsymbol{g}_i^{r,t}\}_r$ is the concatenation of the aggregated features across all relationships to the i_{th} node at the t_{th} step. f_{upd} illustrates the node feature update function with the weight $\boldsymbol{w}_{upd}$. The features $\boldsymbol{h}$ for the scene containing N objects are acquired via $\boldsymbol{h} = \sum_{i=1}^N v_i \boldsymbol{h}_i^T$, where v_i means the importance weight of the i_{th} object and T is the maximum iteration number of graph update in each round of object addition. The existence, location, class, and placement (orientation and size) of additional objects are predicted based on the scene features.

However, due to the object category imbalance and the confounder objects in the dataset, directly learning the graphical model from the training data usually causes significantly biased prediction. Our goal is to learn unbiased graph neural networks with object category distribution calibration and confounder effect elimination via causal inference.

C. Object Category Distribution Calibration

Due to the long-tailed distribution of dataset categories, conventional indoor scene synthesis methods are severely biased in capturing the relationships between objects, i.e., the models prefer to predict the more high-frequency categories, which obscures the basic common knowledge between objects. Meanwhile, the explicit prior knowledge among furniture objects will be hard to identify due to the solidified room layout in the dataset, and the common-sense relationships among objects will be confounded by observation bias. Fig. 4(a) and 4(d) illustrates the causal graph of indoor scene synthesis for conventional methods and our FairScene. For a given object, X represents the furniture and appliance that provide beneficial context for sequential object addition, which shows the clear causal link between the existing and added objects instead of the trivial correlation. Y is the object to be added and Z means the implicit confounder in indoor scene generation. A stands for the attention of the message related to X and the given object, which demonstrates the importance of their correlation. The causal graph of conventional biased indoor scene synthesis leads to two kinds of prediction bias:

- Since the 3D indoor scene datasets are collected in natural distribution, the number of objects in each category is significantly imbalanced. The indoor scene synthesis model tends to add objects that appear more frequently in the training dataset even without clear context, because the biased prediction decreases the training loss faster.
- Despite the beneficial causal links $X \to Y$ and $X \to A \to Y$, X may also impose influence on Y via the confounder Z through the backdoor path $X \leftarrow Z \to Y$. Even though the causal link between X and Y is weak, the correlation learned by the conventional graphical model can be very strong when the causal links $Z \to X$ and $Z \to Y$ are both significant. The backdoor path enforces the model to learn the illusion that X and Y are clearly related.

Both the imbalanced object category distribution and the confounder effect cause sizably biased scene generation. To address these, we remove the long-tail effects via counterfactual sample construction and eliminate the confounder effects by input context intervention. Intervention and counterfactual are two widely adopted techniques for causal analysis [32]. Intervention deletes all incoming links to a variable and assigns a certain value, which means the variable is no longer affected by its parents. Counterfactual sets the variable as the situation that would never happen, and takes one more step further than intervention.

However, the bias caused by category long-tail effects can potentially help the model filter out implausible predictions (e.g. toilets rarely appear in bedrooms). Therefore, we focus on measuring biases that are positive for predicting object Y. According to the causal graph shown in Fig. 4(a), the observed object prediction with the given input x and implicit confounder z in conventional indoor scene synthesis methods is written as $Y_x(z)$. We denote the corresponding counterfactual prediction with the implicit confounder z as $Y_{\bar{x},a}(z)$ with the causal graph shown in Fig. 4(b), which is obtained by setting input x as the mean features $\bar{x}$ of all objects in the training set and remaining the attention a unchanged. The counterfactual prediction comes from the blank input with the object priors, which are completely learned from the training set with long-tail object category distribution. For example, we directly set some nodes of X as a zero vector in the training phase. Therefore, the prediction without biased object priors lies in the difference between $Y_x(z)$ and $Y_{\bar{x},a}(z)$, which is dubbed as Total Direct Effect (TDE) in causal inference [46], [47]. The TDE in the causal inference domain measures the proportion of A in the effect of $X \rightarrow Y$. Meanwhile, contextual bias can be naturally eliminated between facts and counterfactuals. We acquire the prediction $\hat{Y}_x(z)$ without biased object priors of input context x in the following:

$$\hat{Y}_x(z) = Y_x(z) - Y_{\bar{x},a}(z) \tag{7}$$

The applied TDE does not introduce any additional parameters and can be used as an off-the-shelf module in a wide range of indoor scene synthesis methods. The resulting causal graph after long-tailed category calibration is shown in Fig. 4(c), where X^e depicts the beneficial context object with the balanced class distribution.

D. Confounder Effect Elimination

Realistic scene datasets are subject to unavoidable observational biases due to human annotation. Previous methods focus only on mining the relationships between furniture objects in the room layout and ignore the presence of confounding factors. Although subtracting the counterfactual prediction removes the bias caused by long-tailed category distribution, the result from the causal graph shown in Fig. 4(c) still suffers from the confounder biases. Conventional indoor scene synthesis methods predict the probability for object addition in the following conditional probability, by Bayesian theorem we can obtain:

$$P(Y|X,A) = \sum_z P(Y|X,A,z)P(z|X) \tag{8}$$

where the confounder effects are brought by the observational bias $P(z|X)$. The conventional method learns the relationship $X \rightarrow Y$ by the likelihood $P(X|Y)$ only but significantly suffers from the likelihood $P(z|X)$ of the confounding factor z with X. If z is frequently associated with X, Y, the original common knowledge between X and Y will be covered by z, which leads to serious bias in network learning and prediction. Let us take Y as the nightstand and X as the pillow in co-occurrence relationship mining for example. Since $P(z = bed|X = pillow)$ is very large, the most contribution to the likelihood in (8) comes from $P(Y = nightstand|X = pillow, z = bed)$, where A is omitted for simplification. Therefore, the estimation of $P(z = bed|X = pillow)$ actually focuses on beds instead of pillows and leads to biased object addition prediction. We can consider the ordered triad $\{X, Z, Y\}$ as a structural causal model (SCM), where Z is an exogenous variable and X and Y are endogenous variables. To explore the relationship between X and Y, we need to introduce the concept of intervention in causal inference, which can be written as $do(\cdot)$. The intervention operation can modify the value of a node in the causal graph, e.g. $do(X = x)$, while removing all paths to that node. The causal relationship between X and Y can be determined by observing the effect of a change in the value of X on Y. In order to eliminate the confounder bias, we intervene the input object feature by cutting off the causal link between X and Z, and the prediction probability is formulated as follows:

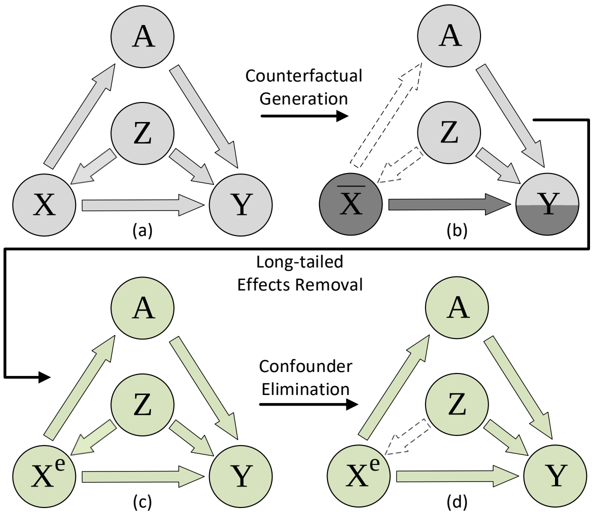


Fig. 4. The causal graph of (a) conventional scene synthesis, (b) the counterfactual prediction, (c) scene generation with long-tailed effect removal, and (d) unbiased scene generation with further confounder elimination. For a given object, X means the furniture and appliance that provide beneficial context with clear causal links for object addition, and Y is the object to be added. A stands for the attention of messages related to X, and Z represents the implicit confounder in indoor scene generation. X^e depicts X with balanced object category distribution.

$$P(Y|do(X),A) = \sum_z P(Y|X,A,z)P(z) \tag{9}$$

where $do(X)$ means the intervention operation on X. By applying the backdoor adjustment, the object prediction treats each z fairly by considering the statistical prior $P(z)$ instead of $P(z|X)$. Consequently, the confounder effect is eliminated thoroughly, where the corresponding causal graph is depicted in Fig. 4(d).

Since the likelihood in (9) is intractable, we apply neural networks to parameterize the probability. As shown in Fig. 2,

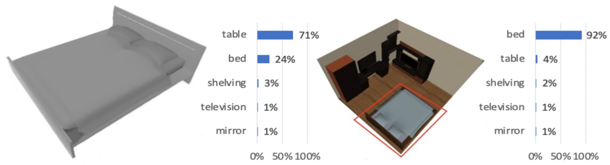


Fig. 5. Object recognition in scenes. **Left:** Object recognition using conventional 3D shape processing architectures. **Right:** Context-based object recognition that fuses the shape and context information with enhanced robustness.

$P(\hat{Y}|X = \boldsymbol{x}_i, A, z = \boldsymbol{x}_j)$ is parameterized by $\boldsymbol{g}_{i,j}^{r,t}$ which fuses the information carried by the beneficial contextual objects X, the message attention A and the confounder Z in the scene. The conventional biased scene synthesis performs expectation of $\boldsymbol{g}_{i,j}^{r,t}$ across all j via GRU module to aggregate messages with z sampled from $p(z|X)$, so that the likelihood in (8) depicts biased object addition prediction. In order to eliminate the confounder effects, we adjust the original message aggregation process in the following by combining (8) and (9):

$$\boldsymbol{g}_{i,j}^{r,t} = f_{GRU}(\boldsymbol{g}_{i,j-1}^{r,t}, \boldsymbol{m}_{i,j}^{r,t}, \boldsymbol{a}_{i,j}^{r,t}) \cdot \frac{p(z = \boldsymbol{x}_j)}{p_r(z = \boldsymbol{x}_j|X = \boldsymbol{x}_i)} \qquad (10)$$

where $p(z = \boldsymbol{x}_j)$ means the probability that the category of $\boldsymbol{x}_j$ appears in the dataset, and $p_r(z = \boldsymbol{x}_j|X = \boldsymbol{x}_i)$ demonstrates the probability that $\boldsymbol{z}_j$ is correlated with $\boldsymbol{x}_i$ via the r_{th} relationship in all training scenes. Both probabilities are obtained via calculating the statistics in the training dataset.

E. Applications

In this section, we show the application of our FairScene in a wide variety of vision tasks including scene completion and object recognition in scenes.

Scene completion: Our FairScene generates the deficient furniture and appliance in a scene based on the existing objects and their relationship. For incomplete scenes, we first mine the unbiased object correlation to form the scene features, through which the existence, location, category, and placement of the additional objects are predicted. The completeness of the synthesized scenes can be controlled by the probability from the existence predictor. Once the model predicts a low probability for the existence of additional objects, the scene can be regarded to be completed with sufficient objects. Users can tune the furniture or appliance arrangement during the iterative scene completion, which enables the interaction between our scene synthesis model and human designers.

Object recognition in scenes: Recognizing objects in 3D scenes is challenging because objects across different categories may share similar appearances and mislead the classifier. Conventional methods extract the volumetric [54], multi-view [43] or point cloud representations [34] via the 3D shape processing architectures, which are sensitive to the appearance variance of objects. On the contrary, we recognize the objects in 3D scenes via the context predictions from our FairScene. We remove the object to be recognized in the scene and query the location for object addition. The predicted object category probability can be regarded as the object recognition results only based on the scene context. By multiplying the object probability predicted by the conventional 3D shape processing architectures and that yielded by our context predictions, we obtain robust object recognition in 3D scenes. Fig. 5 shows the comparison between our conventional and context-based object recognition in 3D scenes, where the latter achieves higher robustness.

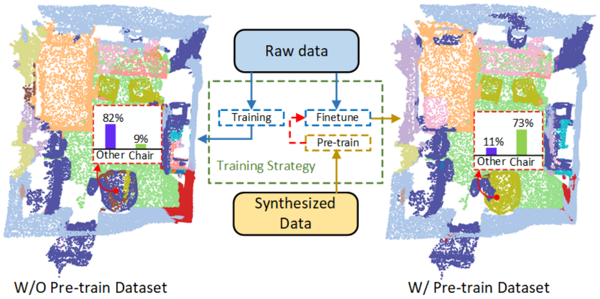


Fig. 6. Synthetic data pre-training visualization process. The 3D visual segmentation model is first pre-trained with a synthetic dataset, which allows the model to learn some object feature information, and then fine-tuned on the public dataset to meet downstream tasks. Compared with the original dataset alone, the pre-training of the synthetic dataset can improve the accuracy of object segmentation.

Pre-training dataset synthesis: The process is illustrated in Fig. 6. Our proposed FairScene can synthesize new scenes based on a priori knowledge of the dataset to meet the demand for large training data for 3D visual understanding tasks. Since it is synthesized by a simulator, our approach can be quickly deployed in unmanned system simulation sessions such as robot indoor cruising and exploration perception. Pre-training the visual perception model with synthetic datasets helps to further improve the model performance. Compared to traditional datasets, synthetic datasets have the following advantages: (1) Synthetic datasets can be automatically labeled, significantly reducing the cost of dataset labeling while further improving labeling accuracy and reducing the difficulty of model learning features. (2) Influenced by the prior knowledge of the dataset and causal inference, our proposed FairScene can alleviate the long-tail effect of traditional training samples and generate a more reasonably distributed training set by artificially setting the ratio. (3) Synthetic datasets can generate a large number of training samples, which can effectively support deep learning model pre-training.

IV. Experiments

In this section, we first introduce the dataset and implementation details in our experiments and then evaluate our FairScene via qualitative visualization and quantitative perceptual study of the generated scenes. Moreover, we compare our method with the state-of-the-art scene synthesis methods in a wide variety of vision tasks including object query and object recognition in scenes.

A. Datasets and Implementation Details

We carried out all experiments on the 3D-FRONT [12] dataset. 3D-FRONT is a large-scale synthetic 3D scene dataset with dense volumetric annotations, which contains 6813 houses and has approximately $18,797$ rooms. To meet the diversity of scenes, 3D-FRONT offers 7302 pieces of furniture

TABLE I
PERCENTAGE (WITH STANDARD ERROR) OF FORCED-CHOICE COMPARISONS ON WHICH SCENES GENERATED BY FAIRSCENE ARE JUDGED TO BE MORE PLAUSIBLE THAN OTHER METHODS.

Room Type	FairScene vs.	
	SceneGraphNet	Real
Bedroom	54.1±2.7	43.9±3.8
Living Room	58.2±3.8	39.1±4.2
Library	53.7±3.2	52.5±2.6
Dining Room	57.9±2.9	48.2±3.5

with high-quality textures. As for training and testing, We applied rooms of four different types (Living room, Bedroom, Library, and Dining room) in 3D-FRONT. After dataset preprocessing and filtering out rooms with non-conforming categories or object texture vestiges, we ended up with 1848 living rooms, 4526 bedrooms, 1585 libraries, and 5780 dining rooms. To meet the training requirements, we further merged the categories of some furniture instances to simplify the training process. We finally selected 45 furniture categories for the living room, 43 for the bedroom, 41 for the library, and 45 for the dining room. Following [66], we trained independent models for each room type and excluded the rooms without four walls in the rectangle in our experiments. For a fair comparison, we split the preprocessed dataset into $8:1:1$ for training, validation, and testing respectively.

We trained all MLPs in our framework jointly, where the dimension of the encoded vector and hidden layers is set to be 100 and 300, respectively. The node features were updated for three iterations before fusing them to obtain the scene features. By randomly removing the objects in the training scenes, we constructed the input scenes for training and enforced our model to correctly predict the existence, location, category, and placement of additional objects. We utilized cross-entropy loss to train the binary classifier for existence and the multi-class classifier for category prediction. The L_2 distance between the prediction and the ground truth is minimized to train the location and size predictors. For the orientation in object placement, we discretized the angles in $[0, 2\pi]$ into 16 values with equal angular difference, which was predicted by a multi-class classifier. The number of iterations set for each training is 50. We utilized the Adam optimizer with the starting learning rate 0.001, which decayed twice at the 10_{th} and 20_{th} epoch out of 30 training epochs by multiplying 0.1. The batch size was set to 1 in all experiments.

B. Comparisons with State-of-the-art Methods

In this section, we compared our FairScene with the state-of-the-art graph network-based scene synthesis model SceneGraphNet [66]. As for training and testing, the original SceneGraphNet used SUNCG [42] dataset to train the model. Unfortunately, the SUNCG dataset was no longer available at the time we performed the experiment. Thus, we retrained SceneGraphNet on the 3D-FRONT dataset. We perform extensive experiments on the qualitative synthesized scene visualization and quantitative perceptual study on plausible scene selection. Moreover, we show the performance of our FairScene in a wide variety of tasks including object query and object recognition in 3D scenes, where ablation study w.r.t. the presented techniques was conducted on object query.

Scene completion: For incomplete scenes, we iteratively add furniture or appliances to design a plausible scene with the automatic scene synthesis frameworks. Fig. 7 demonstrates the synthesized scenes with different completeness across various methods. Although SceneGraphNet can effectively mine the object correlation via convolutional neural networks and graphic models, it ignores the causality among them and the prediction is significantly biased. For example, as shown in the 4_{th} column of the 1_{st} row of Fig. 7, SceneGraphNet deviates to generate the bedside nightstand next to the bed, but violates the spatial location constraint, resulting in a reduced scene realism. Meanwhile, the 5_{th} column of the 4_{th} row of Fig. 7 demonstrates that SceneGraphNet generates plant pots at bedside locations due to confounding effects, resulting in biased predictions and reduced scene confidence. On the contrary, the generated scenes of our FairScene acquire higher plausibility compared with SceneGraphNet since we remove the bias caused by long-tailed category distribution and the confounder effects, especially for scenes with many objects since the complex object interaction is usually biased. For example, our approach weakens certain ”co-occurring” relationships between bed and nightstand to dynamically generate scenes with higher confidence.

We also conducted a perceptual study on plausible scene selection. The participants were asked to select the most plausible scenes in a triplet consisting of generated scenes from different methods and the training scenes, where the order of scenes in each triplet is random. We employed 21 participants which were equally divided into three groups including vision researchers, non-vision researchers and non-technical subjects, and each participant was presented with 31 triplets. Fig. 8 demonstrates the statistics of the perceptual study. The scenes generated by our FairScene were voted to have higher plausibility than other scene synthesis methods across all three groups of participants, and even achieve slightly higher plausibility than the training scenes that are designed by humans in the group of non-vision researchers. Meanwhile, our proposed FairScene outperforms the baseline in all participant groups, which demonstrates the effectiveness of bias elimination for scene synthesis. Prior knowledge of dataset object distribution and counterfactual inference can effectively improve synthetic scene realism. Moreover, we directly compare the plausibility of the scenes generated by FairScene and other methods by forced-choice experiments. The participants are forced to select the scenes with higher plausibility from those generated by the two approaches. Table I demonstrates the percentage with standard error of forced-choice experiments, where the percentage larger than 50% means the preference of our method, and the comparison with the real data in 3D-FRONT is also demonstrated. Our method outperforms the compared baseline methods in all room types and even shows competitive plausibility with the real data from 3D-FRONT. The living room and dining room scenes contain most objects compared with other room types, where the complex object correlation causes obvious scene bias. Therefore, the plausibility for bedrooms is significantly enhanced by our unbiased scene generation.

Object query: Object query means predicting the category

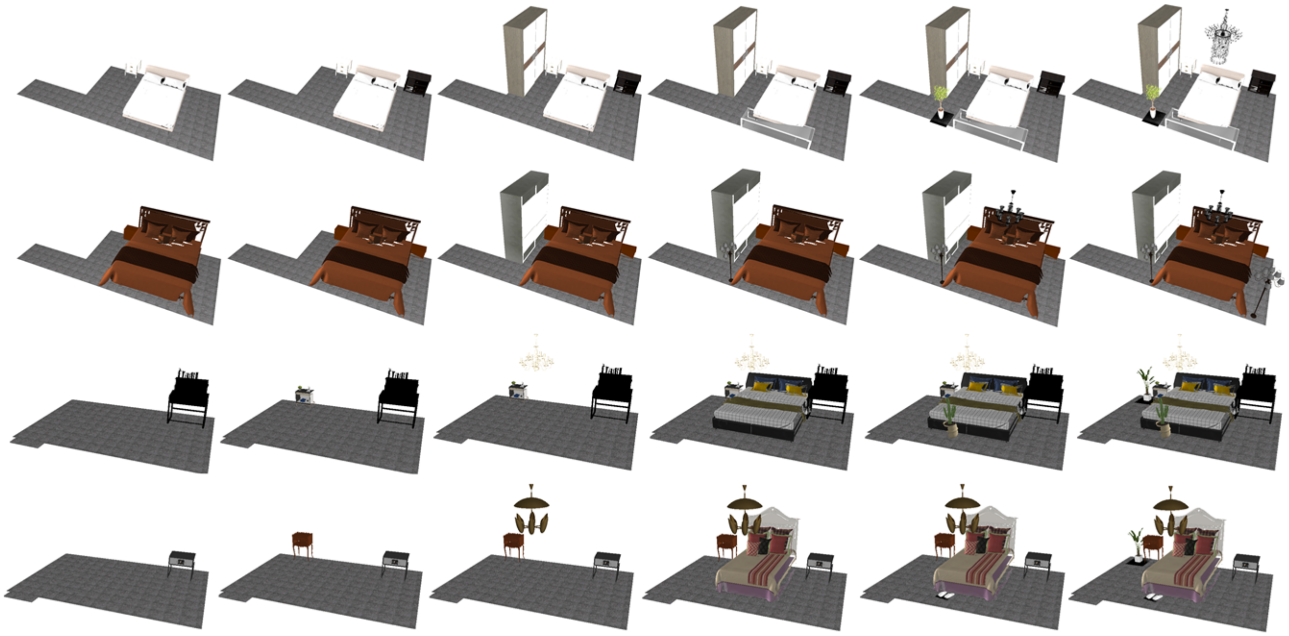


Fig. 7. Scene synthesis visualization results. Scenes synthesized by SceneGraphNet (the first and third row), and our FairScene (the second and fourth row) in different completeness. Although various methods achieve similar results in the early stage, scenes generated by our FairScene obviously obtain higher plausibility with the object increase, because removing the long-tail effects and confounders benefits indoor scene synthesis with complex object interactions.

TABLE II

Top-K accuracy (%) averaged over all scenes in the living room, bedroom, library, and dining room of different methods for the object query. We remove counterfactual predictions, and intervention operations to explore the effectiveness of each module separately, and also conduct ablation experiments on the iteration numbers and message passing modules.

Methods	Living Room			Bedroom			Library			Dining Room			Average		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Baseline	58.70	78.78	85.00	67.09	82.00	87.38	55.15	75.49	83.65	68.29	83.41	88.61	62.31	79.92	86.16
Baseline-K2	59.58	77.74	85.08	67.44	81.86	87.06	51.73	75.98	84.04	69.42	83.98	89.48	62.04	79.89	86.42
Baseline-K4	59.08	77.90	84.08	65.66	82.91	88.25	54.01	77.25	84.52	66.65	83.52	89.65	61.35	80.40	86.63
FairScene-K2	59.09	77.76	84.94	68.08	82.79	88.45	54.06	77.73	84.96	68.40	84.57	90.47	62.41	80.71	87.21
FairScene-K4	61.93	79.64	86.01	68.02	84.17	89.13	53.05	77.46	85.08	68.83	83.98	88.69	62.96	81.31	87.23
Maxpool	65.21	83.10	88.67	64.52	79.35	85.60	48.18	72.42	82.64	65.00	80.15	86.96	60.73	78.76	85.97
CatRNN	59.50	78.01	85.20	67.56	82.52	87.74	48.88	74.53	82.77	68.30	84.14	90.29	61.06	79.80	86.50
Sum	63.31	82.39	88.59	66.34	80.34	86.67	47.44	72.42	82.42	62.63	79.34	87.19	59.93	78.62	86.22
FairScene-NL	59.97	78.64	85.59	67.89	83.03	87.78	55.55	77.03	83.95	69.13	83.95	89.49	63.14	80.66	86.70
FairScene-NC	59.46	78.74	85.79	68.78	84.28	89.62	53.27	78.22	85.46	69.55	84.67	89.82	62.77	81.48	87.67
FairScene-NI	56.41	75.57	82.01	64.00	83.43	88.73	47.96	73.21	82.51	67.84	83.25	89.51	59.05	78.87	85.69
FairScene	62.94	80.65	87.49	69.24	85.65	90.24	54.52	79.00	86.51	70.19	85.07	90.45	64.22	82.59	88.67

of objects to be added for a query location in incomplete scenes. We randomly removed an object in the scene, whose location was then queried in the scene synthesis model for object addition. We calculate the classification accuracy that the predicted category is consistent with the ground truth of the removed objects. We also measure the Top-K accuracy which means the ground truth category is comprised in the K most probable classes because some objects (a speaker beside a TV) can be substituted by another (a plant beside a TV) in a scene without plausibility degradation. We ran the officially released code of the baseline methods to obtain their performance. For a fair comparison, all 3D scenes are projected to 2D top-down view scenes. Table II demonstrates the Top-K (K=1, 3, 5) accuracy averaged over all scenes in each room type of different methods for the object query. Compared with the state-of-the-art method SceneGraphNet, our FairScene achieves increases in the average Top-1, Top-3, and Top-5 accuracy by 1.92%, 2.67%, and 2.51%, respectively. For specific room categories, our proposed FairScene improves the Top-1 accuracy by 4.24%, 2.15%, -0.63%, and 1.90% respectively compared to the baseline. The performance improvement is more significant for the living room, bedroom, and dining room which room layouts contain more distinctive features. For example, chairs or tables are often distributed around the table which denotes ”surrounding” relationships. The correlation becomes more complex and the generated scenes are more biased in conventional methods. The contrast is with libraries that have fewer objects and simpler scenes. Since the library contains only the main furniture such as bookshelves, chairs, tables, etc., it cannot form complex relationships due to the limitation of the number of furniture. Our FairScene underperforms SceneGraphNet on the samples from the library because there is no significant bias from the layout of the library scene.

Table II also demonstrates the performance of FairScene variants. We evaluated the FairScene without object category

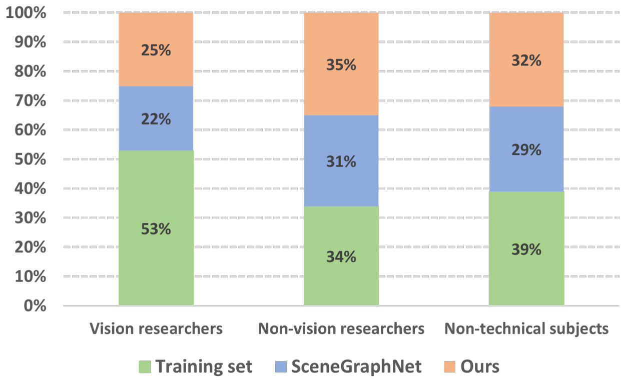


Fig. 8. Perceptual study on plausible scene selection, where the participants were divided into three groups: vision researchers, non-vision researchers, and non-technical subjects. The scenes generated by our FairScene were considered to have much higher plausibility than other scene synthesis methods, and even slightly higher than the training scenes in the group of non-vision researchers.

TABLE III
COMPARISON OF OBJECT CLASSIFICATION ACCURACY (%) WITH 3D SHAPE RECOGNITION METHODS AND CONTEXT-BASED RECOGNITION OF STATE-OF-THE-ART SCENE SYNTHESIS METHODS. LIVING., BED. AND DINING. MEAN LIVING ROOM, BEDROOM, AND DINING ROOM RESPECTIVELY.

Methods	Living.	Bed.	Library	Dining.	Average
PointNet++	41.31	37.55	36.89	44.61	40.09
SceneGraphNet	61.82	70.37	58.25	71.49	65.48
FairScene	64.61	72.93	59.27	73.82	67.66

TABLE IV
COMPARISON RESULTS OF 3D SCENE SEMANTIC SEGMENTATION MODELS ON DIFFERENT TRAINING DATA.

Methods	Point acc	Class acc	Point mIoU
PointNet++	74.15	58.37	47.33
PointNet++ add SG	76.16	58.84	48.70
PointNet++ add Ours	77.59	61.94	51.45
PointConv	71.69	63.15	42.18
PointConv add SG	72.55	64.34	42.56
PointConv add Ours	73.28	64.82	43.41

distribution calibration (NL), without confounder removal (NC), and without importance weights for node feature merging (NI), with two iterations (K2) and four iterations (K4) for node feature update. We also perform ablation experiments on multiple message transfer units to explore the optimal network configuration, the original GRU module is replaced with CatRNN, Maxpool, and Sum respectively. By comparing FairScene-NL, FairScene-NC, and FairScene, we conclude that long-tailed effects removal and confounder elimination both enhances the model plausibility, and integrating them further increases the top-k accuracy as the bias is thoroughly removed. NL and NC have made considerable contributions to the elimination of synthetic scene bias from different perspectives, respectively. By observing results from FairScene-NI, we draw the conclusion that lacking importance weights for node feature fusion when generating scene features significantly degrades the performance, since the contribution of various objects to the scene is very different. Node feature fusion weights provide more discriminative information to the network compared to NL and NC. Constructing only through dataset prior knowledge and counterfactual samples cannot deeply extract the overall contextual information of the scene. The performance of FairScene-K2 and Baseline-K2 indicates that insufficient iterations for node feature updates fail to pass informative messages among different nodes. Although the smaller number of iterations improves the accuracy in the living room scenario, the average accuracy of Top-1 decreases by 0.27% and 1.81%, respectively. We also found that the number of iterations is not as large as possible, the increase in the number of single iterations significantly increases the computational cost of the model while the accuracy improvement is weak. The performance of FairScene-K4 and Baseline-K4 indicates that more iterations lead to a 0.96% and 1.26% decrease in Top-1 accuracy, respectively. Considering both computational cost and model performance, we set the number of iterations(K) to 3. Comparing CatRNN and FairScene, we know that the update gates and reset gates in GRU sizably strengthen the informativeness of feature aggregation in node feature updates. Comparing Maxpool, Sum, and CatRNN, we can obtain that the simple feature information transfer operation loses scene context information, resulting in a 0.33% and 1.33% decrease in accuracy Top-1, respectively. Although Maxpool is efficient in updating information with Sum and achieves the highest accuracy in the living room scenario with 2.27% higher Top-1 accuracy than the GRU module, its Top-1 accuracy decreases by 6.34% for simple scenarios such as the library. Combining the performance of several scenes, we selected GRU as the information transfer unit model.

Object recognition in 3D scenes: Since recognizing objects in 3D scenes by 3D shape processing architectures is sensitive to object appearance variance, considering the context in the scenes enhances the robustness of object recognition in 3D scenes. By multiplying the category probability predicted by conventional 3D shape processing architectures and that produced by our object query, we acquire the robustness of object recognition results in 3D scenes. For each object in our dataset, we utilized PointNet++ [35] to obtain the category probability predicted by 3D shape processing architectures, and then we leveraged different scene synthesis methods to predict the context-based object probability by removing the objects in the scene. As for training, we first iterate through all the mesh files for each category of rooms based on the parsed room furniture dictionary. Then, we sampled each furniture instance mesh file using the farthest point sampling provided by PointNet++ to obtain the point cloud data and compose the training and test datasets. Table III illustrates the average accuracy of different methods for various room types. Compared with the selected baseline, SceneGraphNet significantly improves the accuracy by employing the rich object interaction information in complex scenes, with an average improvement of 25.39%. Integrating our FairScene and PointNet++ yields the highest accuracy due to the category distribution calibration and confounder elimination with an average improvement of 25.57% recognition accuracy, which is ignored in SceneGraphNet with biased object predictions. The above experimental results further demonstrate that the dataset prior knowledge and counterfactual reasoning can eliminate the confusion effect of indoor complex scenes.

Synthetic dataset for pre-training: Since training 3D scene semantic segmentation requires a large number of

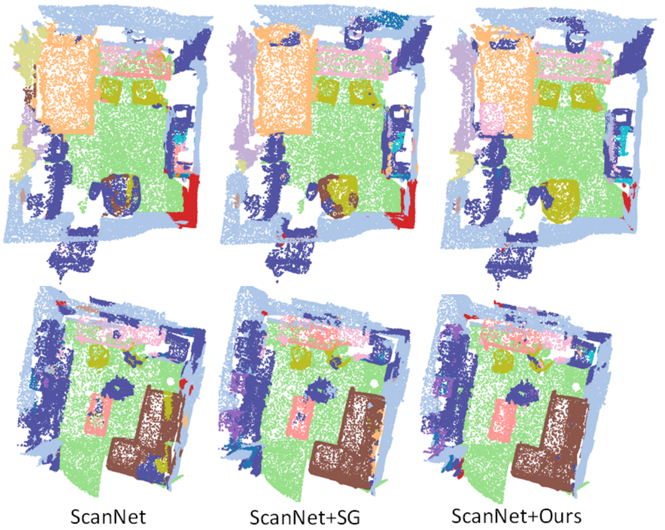


Fig. 9. Visualization results of PointNet++ with different training data settings. We selected a total of three settings: original ScanNet data, SceneGraphNet synthetic data with ScanNet (ScanNet + SG), and FairScene synthetic data with ScanNet (FairScene + Ours). Using pre-trained model fine-tuning can further improve the classification accuracy of the model.

training samples to meet the deep neural network learning requirements, we employ indoor scene synthesis to form a pre-training dataset to further improve the performance of the models on other publicly available datasets. We selected the widely adopted PointNet++ and PointConv [53] as the baseline networks. Meanwhile, we selected the ScanNet [9] dataset to fine-tune the pre-training model. For the synthetic dataset, we first restore the synthetic room structure and reconstruct the entire scene based on the predicted categories, poses, and poses with the CAD model provided by 3D-FRONT. Then, we generate the boundary structure information such as floor and wall for the newly synthesized room based on the original structure of the room provided by 3D-FRONT. In the end, we obtained the entire scene point cloud from the rendered CAD model by furthest point sampling and labeled the categories according to furniture labels, thus forming a synthetic indoor scene understanding dataset. The selected baseline network is first pre-trained on the synthetic dataset and then fine-tuned on the ScanNet dataset to verify the performance of the synthetic dataset. Table IV demonstrates the performance of the pre-training models, where $add\ SG$ represents using SceneGraphNet to synthesize the pre-training dataset, $add\ Ours$ means using FairScene to synthesize the dataset. No addition means that the original ScanNet is adopted to train the models. Compared with the original dataset, the dataset generated using SceneGraphNet resulted in a 1.37% and 0.38% improvement in Point mIoU for PointNet++ and PointConv, respectively. Meanwhile, using FairScene improved 4.12% and 1.23%, respectively. The above experimental results illustrate that the synthetic scene pre-training model can improve the model performance, and also point out that the layout generated by our proposed method is more in line with the real scene. Fig. 9 demonstrates the results of PointNet++ on different settings for the ScanNet segmentation task. For example, the second and third columns of the visualization in the first row of Fig. 9 visually illustrate that the pre-training model can provide rich contextual information to further constrain the prediction, the chair in the bedroom is correctly identified by the model with the help of the synthetic dataset. The second row of Fig . 9 illustrates that the accuracy of the sofa segmentation in the living room is improving influenced by the contextual information resulting from the pre-training dataset. Meanwhile, the use of FairScene synthetic data can provide accurate a priori information guidance (higher accuracy of chair segmentation), which illustrates that the scene layout generated by our proposed FairScene is more similar to the real scene.

V. Conclusion and Discussion

In this paper, we have presented an unbiased graph neural network learning method called FairScene for indoor scene synthesis. The proposed FairScene calibrates the long-tailed category distribution by subtracting the counterfactual predictions of default input and eliminates the confounder effects in the datasets by backdoor path adjustment so that unbiased object interaction is mined for plausible scene generation. Extensive experiments on a wide variety of vision applications including scene completion, object query, and object recognition in scenes show the superiority of our proposed FairScene. Meanwhile, the proposed approach can also synthesize pre-training data for 3D scene understanding tasks. Though the indoor scene bias becomes significant for rooms with complex object entanglement, our method may not be applicable for scenes with fewer objects such as the library. Meanwhile, only considering the causality between two objects cannot eliminate the bias completely, since the high-order interaction among a group of objects may also contribute to the scene bias. We consider further work to reduce the bias of the synthesis scene by incorporating more furniture information (e.g. material, style), while further enhancing the realism of the synthetic scenes.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant U22B2050, Grant 61976623 and Grant 62125603.

References

[1] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. v. d. Hengel. Counterfactual vision and language learning. In *CVPR*, pages 10044–10054, 2020.

[2] E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108, 2012.

[3] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *ICLR*, 2020.

[4] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.

[5] A. Chang, M. Savva, and C. D. Manning. Learning spatial knowledge for text to 3d scene generation. In *EMNLP*, pages 2028–2038, 2014.

[6] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020.

[7] K. Chen, Y.-K. Lai, Y.-X. Wu, R. Martin, and S.-M. Hu. Automatic semantic modeling of indoor scenes from low-quality rgb-d data using contextual information. *ToG*, 33(6), 2014.

[8] Q. Chen, Q. Wu, R. Tang, Y. Wang, S. Wang, and M. Tan. Intelligent home 3d: Automatic 3d-house design from linguistic descriptions only. In *CVPR*, pages 12625–12634, 2020.

[9] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017.

[10] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *CVPR*, pages 4578–4587, 2018.

[11] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Embodied question answering. In *CVPR*, pages 1–10, 2018.

[12] H. Fu, B. Cai, L. Gao, L.-X. Zhang, J. Wang, C. Li, Q. Zeng, C. Sun, R. Jia, B. Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *ICCV*, pages 10933–10942, 2021.

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

[14] L. Gao, J.-M. Sun, K. Mo, Y.-K. Lai, L. J. Guibas, and J. Yang. Scenehgn: Hierarchical graph networks for 3d indoor scene generation with fine-grained geometry. *PAMI*, 2023.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[16] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi. Iqa: Visual question answering in interactive environments. In *CVPR*, pages 4089–4098, 2018.

[17] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla. Understanding real world indoor scenes with synthetic data. In *CVPR*, pages 4077–4085, 2016.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[19] Y. He, Y. Cai, Y.-C. Guo, Z.-N. Liu, S.-K. Zhang, S.-H. Zhang, H.-B. Fu, and S.-Y. Chen. Style-compatible object recommendation for multi-room indoor scene synthesis. *arXiv preprint arXiv:2003.04187*, 2020.

[20] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[21] T. Katsuki, A. Torii, and M. Inoue. Posterior-mean super-resolution with a causal gaussian markov random field prior. *TIP*, 21(7):3182–3193, 2012.

[22] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[23] M. Li, A. G. Patil, K. Xu, S. Chaudhuri, O. Khan, A. Shamir, C. Tu, B. Chen, D. Cohen-Or, and H. Zhang. Grains: Generative recursive autoencoders for indoor scenes. *ToG*, 38(2):1–16, 2019.

[24] M. Liu, K. Zhang, J. Zhu, J. Wang, J. Guo, and Y. Guo. Data-driven indoor scene modeling from a single color image with iterative object segmentation and model retrieval. *TVCG*, 26(4):1702–1715, 2018.

[25] A. Luo, Z. Zhang, J. Wu, and J. B. Tenenbaum. End-to-end optimization of scene layout. In *CVPR*, pages 3754–3763, 2020.

[26] P. Merrell, E. Schkufza, Z. Li, M. Agrawala, and V. Koltun. Interactive furniture layout using interior design guidelines. *ToG*, 30(4):1–10, 2011.

[27] A. Ostonov, P. Wonka, and D. L. Michels. Rlss: A deep reinforcement learning algorithm for sequential scene generation. In *WACV*, pages 2219–2228, 2022.

[28] G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf. Learning independent causal mechanisms. In *ICML*, pages 4036–4044, 2018.

[29] D. Paschalidou, A. Kar, M. Shugrina, K. Kreis, A. Geiger, and S. Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *NIPS*, 34:12013–12026, 2021.

[30] J. Pearl. *Causality*. Cambridge university press, 2009.

[31] J. Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

[32] J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[33] P. Purkait, C. Zach, and I. Reid. Sg-vae: Scene grammar variational autoencoder to generate new indoor scenes. In *ECCV*, pages 155–171, 2020.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NIPS*, 30, 2017.

[36] S. Qi, Y. Zhu, S. Huang, C. Jiang, and S.-C. Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *CVPR*, pages 5899–5908, 2018.

[37] D. Ritchie, K. Wang, and Y.-a. Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *CVPR*, pages 6182–6190, 2019.

[38] P. R. Rosenbaum. Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394, 1987.

[39] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[40] D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[41] M. Shibuya, S. Sumikura, and K. Sakurada. Privacy preserving visual slam. In *ECCV*, pages 102–118, 2020.

[42] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, pages 1746–1754, 2017.

[43] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, pages 945–953, 2015.

[44] K. Tang, J. Huang, and H. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NIPS*, 33, 2020.

[45] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020.

[46] T. VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[47] T. J. VanderWeele. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2):224, 2013.

[48] K. Wang, Y.-A. Lin, B. Weissmann, M. Savva, A. X. Chang, and D. Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ToG*, 38(4):1–15, 2019.

[49] K. Wang, M. Savva, A. X. Chang, and D. Ritchie. Deep convolutional priors for indoor scene synthesis. *ToG*, 37(4):1–14, 2018.

[50] T. Wang, J. Huang, H. Zhang, and Q. Sun. Visual commonsense r-cnn. In *CVPR*, pages 10760–10770, 2020.

[51] Y. Wang, F. Liu, Z. Chen, Y.-C. Wu, J. Hao, G. Chen, and P.-A. Heng. Contrastive-ace: Domain generalization through alignment of causal mechanisms. *TIP*, 32:235–250, 2022.

[52] T. Weiss, A. Litteneker, N. Duncan, M. Nakada, C. Jiang, L.-F. Yu, and D. Terzopoulos. Fast and scalable position-based layout synthesis. *TVCG*, 25(12):3231–3243, 2018.

[53] W. Wu, Z. Qi, and L. Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, pages 9621–9630, 2019.

[54] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015.

[55] G. Xiong, Q. Fu, H. Fu, B. Zhou, G. Luo, and Z. Deng. Motion planning for convertible indoor scene layout design. *TVCG*, 27(12):4413–4424, 2020.

[56] J. Xu, G. Chen, J. Lu, and J. Zhou. Unintentional action localization via counterfactual examples. *TIP*, 31:3281–3294, 2022.

[57] K. Xu, J. Stewart, and E. Fiume. Constraint-based automatic placement for scene composition. In *Graphics Interface*, volume 2, pages 25–34, 2002.

[58] H. Yang, Z. Zhang, S. Yan, H. Huang, C. Ma, Y. Zheng, C. Bajaj, and Q. Huang. Scene synthesis via uncertainty-driven attribute synchronization. In *ICCV*, pages 5630–5640, 2021.

[59] L. F. Yu, S. K. Yeung, C. K. Tang, D. Terzopoulos, T. F. Chan, and S. J. Osher. Make it home: automatic optimization of furniture arrangement. *ToG*, 30(4):1–12, 2011.

[60] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun. Causal intervention for weakly-supervised semantic segmentation. *NIPS*, 33, 2020.

[61] J. Zhang, C. Zhu, L. Zheng, and K. Xu. Fusion-aware point convolution for online semantic 3d scene segmentation. In *CVPR*, pages 4534–4543, 2020.

[62] S.-H. Zhang, S.-K. Zhang, W.-Y. Xie, C.-Y. Luo, Y.-L. Yang, and H. Fu. Fast 3d indoor scene synthesis by learning spatial relation priors of objects. *TVCG*, 28(9):3082–3092, 2021.

[63] S.-K. Zhang, H. Tam, Y. Li, K.-X. Ren, H. Fu, and S.-H. Zhang. Scenedirector: Interactive scene synthesis by simultaneously editing multiple objects in real-time. *TVCG*, 2023.

[64] W. Zhang, Y. Zhang, R. Song, Y. Liu, and W. Zhang. 3d layout estimation via weakly supervised learning of plane parameters from 2d segmentation. *TIP*, 31:868–879, 2021.

[65] Z. Zhang, Z. Yang, C. Ma, L. Luo, A. Huth, E. Vouga, and Q. Huang. Deep generative modeling for scene synthesis via hybrid representations. *ToG*, 39(2):1–21, 2020.

[66] Y. Zhou, Z. While, and E. Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *ICCV*, pages 7384–7392, 2019.

[67] S. Zhu, I. Ng, and Z. Chen. Causal discovery with reinforcement learning. In *ICLR*, 2019.