Eurographics Symposium on Geometry Processing 2016
Maks Ovsjanikov and Daniele Panozzo
(Guest Editors)

*Volume 35 (2016), Number 5*

# Learning 3D Scene Synthesis from Annotated RGB-D Images

Z. Sadeghipour Kermani[1] Z. Liao[2] P. Tan[1] H. Zhang[1]

[1]Simon Fraser University [2]Zhejiang University

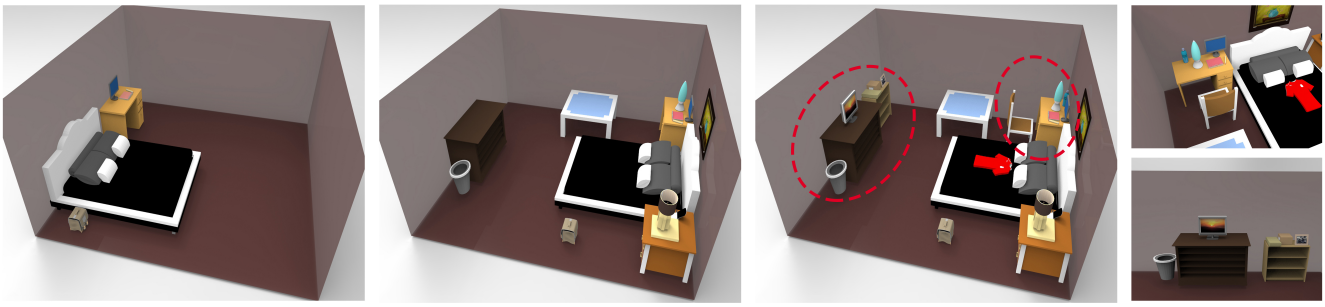

**Figure 1:** *Progressive synthesis of 3D indoor scenes. Starting from an empty room, results from steps 1, 4, and 7 of the synthesis procedure are shown. Two close-ups are given on the side for the last result to highlight the placement of small objects into the scene. Object selection and arrangement are implemented fully automatically based on models learned from a large set of annotated RGB-D Images.*

**Abstract**

*We present a data-driven method for synthesizing 3D indoor scenes by inserting objects progressively into an initial, possibly, empty scene. Instead of relying on few hundreds of hand-crafted 3D scenes, we take advantage of existing large-scale annotated RGB-D datasets, in particular, the SUN RGB-D database consisting of 10,000+ depth images of real scenes, to form the prior knowledge for our synthesis task. Our object insertion scheme follows a co-occurrence model and an arrangement model, both learned from the SUN dataset. The former elects a highly probable combination of object categories along with the number of instances per category while a plausible placement is defined by the latter model. Compared to previous works on probabilistic learning for object placement, we make two contributions. First, we learn various classes of higher-order object-object relations including symmetry, distinct orientation, and proximity from the database. These relations effectively enable considering objects in semantically formed groups rather than by individuals. Second, while our algorithm inserts objects one at a time, it attains holistic plausibility of the whole current scene while offering controllability through progressive synthesis. We conducted several user studies to compare our scene synthesis performance to results obtained by manual synthesis, state-of-the-art object placement schemes, and variations of parameter settings for the arrangement model.*

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—

## 1. Introduction

With the resurgence of VR and AR applications, there has been an increasing demand for 3D scene data, particularly those of indoor environments. Large collections of 3D scenes are also immensely valuable both as training data to support machine learning algorithms for scene analysis and understanding [ZSTX14], as well as model repositories for data-driven 3D scene modeling [CLW*14, FSL*15, KMYG12, LZW*15]. Hence, techniques and tools that are capable of producing realistic virtual 3D scenes in high volume and with large diversity are sought after.

Automated or semi-automated generation of 3D indoor scenes is a relatively new research topic in *geometric modeling*. Methods proposed so far can be roughly classified into three categories in terms of the input and problem formulation: furniture layout optimization, scene modeling and reconstruction, and scene synthesis. Layout optimization involves rearranging a given set of pieces of furniture in a specified room [MSL*11, YYT*11], while scene

modeling is necessarily constrained by a model specification, such as a Kinect capture [FSL*15], an online photo [LZW*15], or a free-hand sketch [XCF*13]. On the other hand, the synthesis problem is typically more open-ended and less constrained. This is the problem we are interested in.

The best known and state-of-the-art method for synthesizing 3D indoor scenes is the work of Fisher et al. [FRS*12]. They take an *example-based* synthesis paradigm where the output scene should bear resemblance to a small number of exemplar 3D scenes provided by the user. To encourage diversity in the synthesized results, a larger database of 3D scenes, serving as priors for the synthesis, is utilized to provide more variations in object occurrences and arrangements, as well as contextual relations between scene objects. They learn a probabilistic model which mixes knowledge from both the exemplars and the background database and produce a synthesized scene by sampling from the probabilistic distribution.

Our synthesis approach is inspired in part by Fisher et al. [FRS*12]. However, instead of taking a holistic view of scene synthesis, targeting overall similarity between the generated scene and the exemplars, we synthesize a 3D scene *progressively* by inserting *one or more* objects into the scene at a time based on a learned probabilistic model. Similar to Fisher et al. [FRS*12], our probabilistic model also consists of two main components: a co-occurrence model to guide *which* objects are to be inserted into the scene and an arrangement model to determine *where* each object should be placed. However, there are two key distinctions related to the probabilistic model and the associated learning process:

1. In addition to considering pairwise object relations, we extract and learn salient *higher-order* relations involving more than two objects, e.g., two nightstands symmetrically surrounding a bed. Clearly, higher-order relations of these kinds do exist in many indoor scenes, and more importantly, such relations are not mere aggregates of a set of pairwise relations. Incorporating such relations into the probabilistic model allows us to insert a group of objects into the current scene following the learned priors.
2. Both co-occurrence and arrangement models consider the *whole* current scene. This way, we do progressive synthesis to offer more controllability of the synthesis process while still ensuring global coherence and plausibility of the synthesized result at each step.

For both the method in [FRS*12] and our work, along with any other data-driven methods, the generality and richness of the data, which forms the prior knowledge for modeling, is critical. Our co-occurrence and arrangement models are not learned from a database of a few hundred user-synthesized 3D scenes, but from a much larger database of depth images capturing real-world indoor scenes. For our work, we utilize the SUN RGB-D database of Song et al. [SLX15], which consists of 10,000+ RGB-D scenes (about 1,300 bedroom scenes alone) in contrast to a total of 130 hand-crafted scenes used by Fisher et al. [FRS*12]. The ensuing challenge however, lies in the extraction of object relations from the far-from-perfect and much noisier RGB-D images.

The co-occurrence model learned for a particular class of scenes is represented as a *factor graph* [ZJ10] encoding different types of pairwise and higher-order relations between objects, including support, symmetry, distinct orientations, and proximity, along with their probabilities. Furthermore, the pairwise placement patterns among different object categories, expressed as the arrangement model, are learned and clustered by the K-means algorithm. Eventually, in the synthesis step, we start from either an empty or a pre-organized scene and sample a set of objects from the factor graph to insert in the scene which results in a high combination score. Afterwards, according to the arrangement model learned from the RGB-D dataset and additional placement constraints imposed by design guidelines, we optimize the position and the orientation of objects in the scene.

We evaluate our contributions by examining the arrangement plausibility of our method against human efforts to manually design a placement in a user study. The first of the other two tests involve comparison with different settings for higher-order relations and the second one evaluates our approach against Fisher's arrangement model in [FRS*12]. Additionally, we emphasize on compelling elements in our approach which result in a considerable improvement. Particularly, as the major contrast to Fisher et al. [FRS*12], we can accomplish holistic plausibility for the whole scene, not just a dining or study area. The distinction is mostly attributed to the richer source of knowledge which empowers us to extract, learn, and formulate more object-object relations.

## 2. Related Work

There are several ways to produce a virtual 3D scene: furniture layout design or re-arrangements, 3D scene modeling from RGB-D scans or 2D sketches and images, object placement, or synthesis from scratch. Our discussion follows this classification. But first, we start by reviewing two relevant works in the 2D domain.

**Edit and shape placement propagation.** Employing geometric relations as the building blocks, both works in [GJWW14, GJWW15] intend to propagate edit operations in the 2D domain. In [GJWW14], the edit propagation is applied to similar parts based on a set of geometric relationships. By incorporating example 2D scenes for guiding the shape placement, a more recent work [GJWW15] learns a probabilistic model based on the feature sets of geometric relations in example placements, which further enables generating novel similar placements. Although the second method can be extended to 3D scenes, both approaches are originally developed for 2D polygons. Moreover, they only consider pure geometric relations, rather than semantic higher-order relations as in our proposed algorithm.

**Furniture layout optimization.** Focusing on *re-arranging* furniture in a room, the works of Merrell et al. [MSL*11] and Yu et al. [YYT*11] both start from a given set of furniture placed arbitrarily in a room. In [MSL*11], Merrell et al. optimize the furniture layout based on a cost function encoding spatial relationships between objects along with ergonomic factors, such as visibility and accessibility. Yu et al. [YYT*11] also attempt to achieve the same goal in an interactive framework by suggesting furniture arrangements following a set of interior design rules, namely visual balance and alignment. In contrast, our approach enumerates the category and number of instances per category to be inserted in the scene automatically and optimizes the placement in a progressive manner, as opposed to a global optimization.

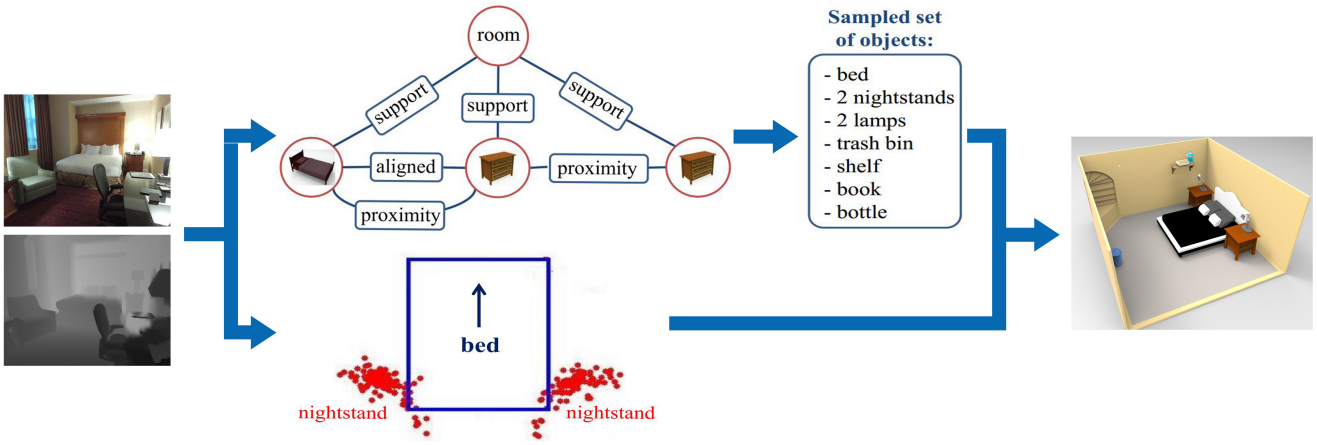


**Figure 2:** *An overview of progressive 3D scene synthesis. From the annotated SUN RGB-D image dataset of real scenes, two general types of relations between objects are obtained; salient higher-order relations to develop a factor graph as the co-occurrence model, which are accompanied by pairwise spatial relations to learn the arrangement model. Objects are sampled from the factor graph to certify a highly probable combination. Subsequently, they are placed in the scene observing a high configuration score computed by the arrangement model.*

**Scene modeling and reconstruction.** To acquire a 3D scene, one can perform reconstruction from freehand sketches [XCF*13] or single-view images [LZW*15]. To model a set of sketched objects, Xu et al. [XCF*13] propose a joint processing technique to co-retrieve and co-place objects in an output scene, according to the structural groups denoting salient relations between objects. In another recent work, Liu et al. [LZW*15] segment an input image to recover object cuboids and obtain the most similar 3D models from the database to model the complete scene in the image. As opposed to their works, we synthesize a 3D scene from scratch guided by knowledge learned from real scenes.

Generating new scenes can also be guided by human activities. Fisher et al. [FSL*15] focus on imitating the functional and geometric properties of the input which is a noisy and incomplete RGB-D scan of a scene. Our model, in contrast, does not consider the viable actions in a scene. Instead, we learn co-occurrence and arrangement relations between objects from RGB-D images. Aiming to extract 3D cuboid arrangements for objects in an RGB-D scan, and similar to our work, Shao et al. [SMZ*14] consider a set of higher-order relations, mainly stability, to construct their model.

**Object Placement.** The works of Jiang et al. [JLZS12, JLS12] learn object arrangement models from 3D scene data. Specifically, in [JLZS12], for a noisy scan of multiple objects and prospective supporting areas, they deduce the surface that each object should be placed on along with its feasible position and orientation. To tackle this problem, they train a probabilistic graphical model based on various criteria including stability, semantic preference, and stacking objects together. Although they follow a learning algorithm for object-object relations, contextual information is not encoded in their model. The work [JLS12] focuses on human-object relations in compliance with the observation that an object arrangement is normally controlled by its affordance and reachability. Similar to our work, they allow placing relevant objects together with the distinction of grouping them according to human poses they share.

**Example-based synthesis.** Our approach is inspired in part by the work of Fisher et al. [FRS*12]. In their framework, they learn a probabilistic model to place and arrange the objects from a mix of user-provided examples and a 3D scene database. Results were only reported for the synthesis of partial scenes, such as those surrounding an office desk or dining table, not of larger-scale scenes such as those of a whole bedroom. In addition, their use of a limited number of training scenes appears to restrict the variety in the final set of outputs. In our work, we learn from many more depth images of real-world scenes to attain more diverse arrangement patterns which are observed over whole rooms. As well, our synthesis algorithm is progressive, allowing a higher-degree of modeling granularity. Technically, one may view our synthesis paradigm as a special case of theirs with the exemplars given a weight of zero in the mixed model. In terms of the object co-occurrence and arrangement models, we add higher-order relations and consideration of holistic scene plausibility into the learning and synthesis pipeline.

## 3. Overview

In this work, we aim to synthesize plausible 3D indoor scenes given the scene type and the desired number of objects to be added in each step. The challenge is first, *which* objects feature a plausible

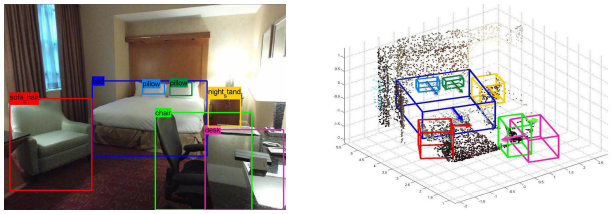

**Figure 3:** *Annotations for images in SUN RGB-D. Left: 2D bounding boxes with category labels for objects. Right: Corresponding 3D bounding boxes along with object orientations.*

*combination* and second, how each object should be *placed* and *orientated* to produce a realistic configuration. Additionally, the synthesized scenes should observe diversity and variety.

We start with a set of annotated RGB images along with their depth data as the main source of knowledge to learn two models in the proposed approach; the extracted higher-order relations between objects are encoded in a factor graph as the probabilistic model and they are augmented with the pairwise relations to state the recurrent patterns for relative arrangements of different object categories. To avoid repeated scenes in a group of final outputs, we apply sampling methods both for generating the list of objects for the scene and also while solving for their placements. As one of the key points of our approach, we benefit from utilizing the SUN RGB-D dataset of more than 10,000 depth images of real-world scenes which contains annotations for 2D and 3D bounding boxes of objects along with their category labels and orientations, as displayed in Figure 3.

The rest of the paper is organized as follows; in Section 4, we explain various types of higher-order relations in more detail and the logic behind the choice of factor graphs as the probabilistic model for our problem. Section 5 provides a closer look at the pairwise relations and side-to-side constraints contained in the arrangement model. Further explanations about utilizing aforementioned models in a sampling procedure to synthesize scenes are included in Section 6. Section 7 presents a comprehensive discussion on results and the evaluation method. Lastly, we conclude and suggest some future directions in Section 8.

## 4. Co-occurrence Model

To control the category and number of instances per category of objects to be inserted in a scene, the co-occurrence model prioritizes the ones with salient relations with other objects which are present in the scene. The prominence of each relation is denoted as a probability in a factor graph. We learn these probabilities from the scenes in RGB-D images and construct a global scene graph for a specific scene category.

### 4.1. Types of relations between objects

The co-occurrence model incorporates several classes of relations between objects. In addition to supporting relations in previous work [FRS*12], two or more objects can be grouped according to a set of higher-order relations, representing their co-occurrences. The significance of the obtained relations is deduced from their frequencies in the training data. We preserve the ones with instance counts higher than a specific threshold. These thresholds are specified separately for each type of relations in Table 1.

**Support Relations.** Every object in a scene, excluding the room itself, requires to be supported by a certain surface of another object. The support could be either from below or behind. To illustrate, a book might be supported by either a desk, a shelf or a nightstand from below while a mirror is usually supported by a wall from behind (Figure 4(a)). We extract and count these relationships from the labels in the image database.

**Symmetry Relations.** Being ubiquitous at every scale in the physical world, geometric symmetry is an inseparable factor from any scene synthesis approach. As an example, in a dining set, the chairs are placed symmetrically around the table. Multiple instances of an object category is regarded as a symmetric group if their 3D models are identical. Since it is challenging to detect similarity between the 3D models in images, we assume that having the same size and the same object category would suffice for being symmetric. Furthermore, the items in a symmetric group might share the same distance and orientation with respect to an object with a different class label, which introduces a different set of relationships in our model.

| Type of relation | Threshold |
|---|---|
| support | 10 |
| symmetry | 0.5 |
| proximity | 5 |
| orientation | 5 |
| side-to-side | 5 |

**Table 1:** *The threshold for recognizing salient relations, stated as the percentage of scenes that include a specific relation*

**Proximity Relations.** The saliency of the relations between objects can be evaluated based on their proximity. Following the prior work on organizing scene collections [XMZ*14], we utilize frequent substructure mining for scene graphs to acquire groups of two or more objects within a proximity threshold of one another. We start with modeling each scene in the RGB-D images dataset as a local graph. The graph comprises the objects as nodes as well as a set of proximity edges connecting each object to the nearest match (Figure 5). Thus, applying a standard algorithm for frequent substructure mining, namely gSpan in [YH02], on the collection of all the local scene graphs would supply recurrent proximity relations between objects.

**Distinct Orientation Relations.** To determine the categories for objects to generate a plausible scene, their relative poses are likely to provide more guidance. For instance, although a combination of a TV set and a sofa is repeated in indoor scenes, they are generally placed too far apart to be considered as a proximity pair. Also, they

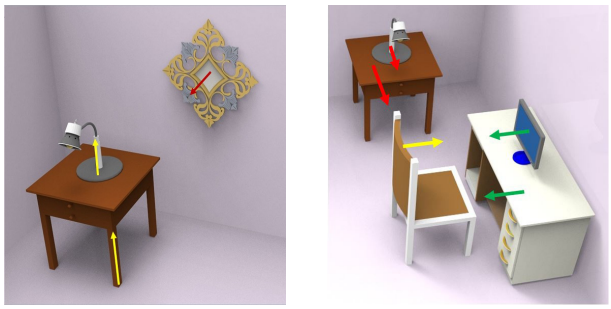

**Figure 4:** *(a) The lamp is supported from below by the nightstand which is itself supported by the floor. The mirror is supported from behind by the wall. (b) The monitor and the desk are relatively aligned and they are both oriented oppositely to the chair. The orientations of these three objects are perpendicular to the poses of the lamp and the nightstand. (Note that all of these relations are extracted from RGB-D images and the 3D representation in this figure is for better visualization.)*

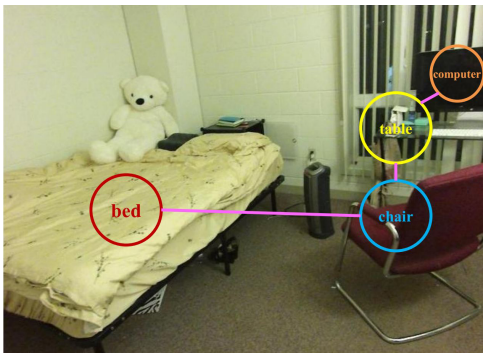


**Figure 5:** *A local scene graph constructed for a scene in the SUN dataset. Each object, represented as a node, is connected to its closest counterpart.*

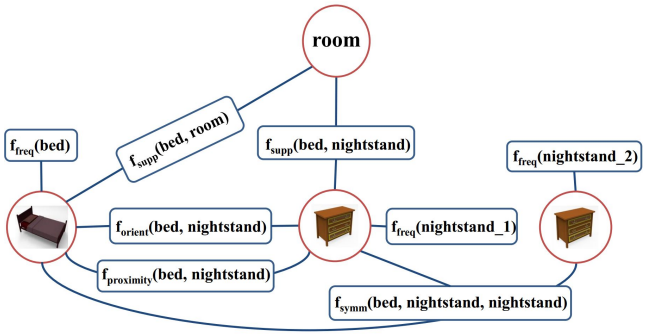


**Figure 6:** *A part of the global scene graph containing three nodes: bed, nightstand_1, and nightstand_2, along with all the salient relations between them, encoded in factors.*

are not categorized as a symmetry or support group, yet the pair follows a common pattern of placement with opposite orientations. Observing similar cases, we encode a number of specific orientation relations in the model, including pairs of objects with aligned, perpendicular, and opposite orientations (Figure 4(b)).

Note that in our current implementation, we only focus on symmetry, proximity, and distinct orientation as the higher-order relations extracted and applied, while other types of grouping relations can also be considered. In particular, groups of objects often appear together in a scene, with specific spatial configurations, due to the functionalities they each serve and collectively perform. For example, a laptop, keyboard, monitor, mouse, mousepad, and printer are typically placed in a predictable manner. We currently do not attempt to learn such complex relations as the analysis would require sophisticated functionality-aware analysis, e.g., [HZvK*15]; we leave that for future work.

### 4.2. Global Scene Graph

To allow an efficient estimation of the plausibility score for a scene configuration, we exploit probabilistic graphical models as in prior work on example-based scene synthesis [FRS*12]. However, our model differs from their approach in several respects, the most notable of which is instead of a Bayesian network, we represent the acquired knowledge from the data in a *factor graph* [FKLW97]. The principal rationale for this modification is the requirement of combining various types of relations in one model. We desired both conditional relations, such as support, and unconditional ones, namely proximity, to be denoted as a single global graph for a scene type. Inspired by the approach in [ZJ10], among several graphical models, we employed factor graphs as they correspond to our goal the most.

Factor graphs consist of two kinds of nodes; variables and factors. For example, in our implementation, the global scene graph for bedrooms contains 30 variables and 108 factor nodes. More details about the global scene graph, including a list of variables, the scope of each factor, and learned parameters are provided in the supplementary material.

**Variables.** In the global factor graph, the variables denote disparate object categories along with the possibly maximal number of instances per category. To put it another way, for each different number of instances of a category, a distinct binary variable is incorporated in the model. By way of illustration, if we do not detect a scene in the dataset with more than two chairs, we preserve two variable nodes for chairs, each with two possible values; one demonstrating the presence of the first chair and the second variable for the second chair. Obviously, the second variable is conditioned on the first one.

**Factors.** The factor nodes in the co-occurrence model primarily indicate the relations between objects and their probabilities. Hence, every factor is associated with two or more variable nodes. As an example, in Figure 6, $f_{orient}(bed, nightstand)$ denotes the frequent aligned orientation between the bed and the nightstand and $f_{symm}(bed, nightstand, nightstand)$ stores the probability information for a group of two nightstands symmetrically placed on two sides of a bed. Additionally, to demonstrate the occurrence probability of one object, a set of single-variable factor nodes are contained in the graph, indicated as $f_{freq}$ in Figure 6.

## 5. Arrangement Model

The plausibility of a scene is contingent on the layout of the scene and its comprising objects; i.e. where each object is placed and which orientation it indicates. Thus, in addition to the co-occurrence model for determining the categories of objects in a scene, the synthesis scheme entails an arrangement model to ascertain a realistic placement for the selected set of objects.

The principal components of the arrangement model is augmenting the higher-order relations with the pairwise spatial relations between different objects. However, we observed that K-means clustering results in a more robust description of these relations than the Gaussian mixture models in their method, primarily owing to the fact that GMMs tend to suffer from *overfitting* in the case of a small amount of training data. Moreover, instead of considering only the centroids of objects, our model respects the side-to-side relations between them as well to fulfill some arrangement constraints such as pushing objects to the walls.

### 5.1. Pairwise Spatial Relations

To model the arrangement of objects, in addition to the higher-order relations, we extract the spatial relations between every pair of objects present in all the instances of a particular scene category in

the dataset. Both position and orientation factorize the placement of an object. As a consequence, the pairwise relations integrate three variables $(x, y, \theta)$; $x$ and $y$ are the projected coordinates of the center of the object 3D bounding box (as in Figure 7(a)), and $\theta$ depicts the yaw angle for the object orientation which measures how much the object is rotated about the $z$ axis, indicating the direction it faces. To reduce the number of parameters of the model, we do not include the $z$ coordinate for it can be estimated by knowing the supporting surface.

During locating an object in the scene, we examine its position and orientation with respect to other objects. Accordingly, for a pair of object categories $C_1, C_2$, the acquired triples $(x, y, \theta)$ encode the position of the instances of category $C_1$ in the frame of category $C_2$ objects and the angle between their orientation vectors as annotated in the SUN RGB-D dataset. Thereafter, to account for varied rational settings of a pair, we apply the K-means clustering algorithm to obtain several feasible modes. Since the number of clusters changes for each pair, we evaluate the silhouette values [Rou87] to estimate the appropriate number.

### 5.2. Side-to-side Constraints

A group of critical arrangement standards which mainly require certain pairs of objects to have the minimal distance could not be perfectly achieved through a model based on the centroids of objects. To name a few instances, nightstands are generally found adjoining to the bed as illustrated in Figure 7(b) or nearly all of the pieces of furniture supported by the floor, rested against the walls, which demands counting walls as objects in the model. Consequently, we investigate distances between disparate pairs of sides of two objects and treat it as a placement guideline if it is less than a definite threshold, which is set to 30$cm$ in our implementation, and occurs frequently enough (Table 1). The collection of side-to-side constraints leads to creating real-world scene layouts by means of rejecting undesirable samples.

## 6. Scene Synthesis

Our scene synthesis framework achieves flexibility through the choice of starting from either an empty scene of a particular category or a prearranged one. Additionally, the desired number of extra objects for the input scene can be specified by the user. To maintain plausibility while reducing the optimization parameters, we adopt a progressive approach to place objects in the scene, one group at a time. The results in Section 7 demonstrate that our algorithm generates scenes with holistic coherency.


**Figure 7:** *(a) The relative positions of the nightstand, the table, the desk, and the chair in the coordinate frame of the bed (illustrated by the red arrows) to learn pairwise relations. (b) An example of side-to-side constraints between two sides of bounding boxes of the bed and the nightstand. (Note that all of these relations are extracted from RGB-D images and the 3D representation in this figure is for better visualization.)*

### 6.1. Sampling Object Categories

To determine the categories and the number of instances for each category of objects, we draw samples from the factor graph formulated in Section 4. To attain both variety in the synthesized scenes and soundness in the sampled set, we perform the Markov Chain Monte Carlo (MCMC) method [Bis07] to obtain multiple plausible combinations of objects.

The sampling step begins with a random collection and proceeds by enabling, disabling, or replacing a factor in each iteration. When we enable a factor, we assign the value one to the associated variables, denoting the presence of the corresponding objects in the scene. Disabling refers to the reverse act of setting the binary variables to zero and integrating the two aforementioned acts introduces a separate type of movement in the sampling process as factor replacement. The probability of selecting each kind of moves is adjusted according to the number of existing objects. As an example, if the user aimed for five new objects and the current set contains six, the probability of disabling factors becomes higher than the other two types.

Every possible move is evaluated according to the joint probability of the factor graph, which is interpreted as the *objects combination score*. If the scores for current and new combinations are denoted by $s$ and $s'$ respectively, the proposed move is accepted with probability:

$$\alpha = min(\frac{s'}{s}, 1) \tag{1}$$

We repeat sampling for a fixed number of iterations to produce varied reasonable groups of objects.

### 6.2. Object Placement

The above sampling procedure is followed by placing each object in the set with respect to the existing ones. The order in which multiple objects are placed depends on their sizes and the number of their salient relations with other objects. Larger objects with more relations limit the arrangement of others, which induces their precedence for being placed in the scene.

We utilize a discrete optimization approach by taking samples from the pairs in the training data. When an object is *being* inserted in the scene, the algorithm searches for a current object with the most salient relationship with the new one. To identify saliency, besides the occurrence probability of every relation in a particular class, various types of relations should be ranked in accordance with their importance. Our experiments revealed that prioritizing in the order of support, symmetry, and distinct orientations, leads to promising results.

Similar to the previous step, we employ MCMC sampling to avoid copying arrangements in the training data. The samples are generated for the position and orientation of the new object with respect to its paired object from the acquired clusters in Section 5, weighted by the number of available data points in each cluster. To

evaluate each sample, we compute the *scene configuration score* which equals the sum of K-means scores for all the pairs of objects in the scene which are related to the new object through at least one of the higher-order relations.

### 6.3. Holistic Plausibility Constraints

The scene configuration score is not sufficient for ensuring the plausibility of the final scene. We reject the sample placements contravening the *holistic plausibility constraints* including side-to-side constraints, objects not colliding, and sufficient space on the supporting surface. Among the remaining samples, we pick the one with the highest scene configuration score.

## 7. Results and Evaluation

Considering the nonexistence of a quantitative criteria for analyzing the plausibility of a scene arrangement, we entrusted human subjects to judge the quality of our results. We first compared our synthesized scenes against manually arranged ones in a pairwise manner. Second, we produced a set of results by removing one or multiple types of relations between objects from the model and investigated their impact on the final placement. We also compared our approach against the object arrangement model by Fisher et al. [FRS*12].

The user studies were distributed among 35 graduate students with diverse majors. User study results reveal (a) our results are reasonably close to human-level arrangement, (b) the effectiveness of relation models in our method, and (c) a clear advantage of our learned object arrangement model against the state-of-the-art [FRS*12].

### 7.1. Datasets

For extracting the knowledge for our models, we mainly exploit the images in SUN RGB-D dataset [SLX15, NSF12, JKJ*13, XOT13]. We retrieve the scenes with the same label as the target category and learn the co-occurrence and arrangement models conforming to them. The dataset incorporates 10,335 images from 47 scene categories in total. Existing objects are from about 800 various categories. To illustrate, there are around 1300 bedroom images. In our method, we focus on 2D and 3D bounding boxes and orientations of objects among the annotations in the dataset.

To model support relations, since SUN RGB-D does not include labels for supporting and supported surfaces, we concentrated on NYU Depth dataset V2 [NSF12] which is also included in SUN dataset. NYU dataset comprises 1449 images including 383 samples of bedroom category.

For the synthesis of scenes, we utilized the 3D models collected in the scene database in [FRS*12]. Furthermore, to display the output, we employ a slightly modified version of their scene viewer.

### 7.2. Comparison with Human Performance

To be aligned with real-world data, the synthesized scene should express an object arrangement which is feasible and plausible, i.e. in addition to a set of hard constraints, such as no collision between objects or sufficient supporting area for an object, there is a group of guidelines for placing objects in a room that are challenging to be stated concisely. The comprehensive evaluation intends to assess the plausibility of our synthesized results subject to these constraints and guidelines. The control group is a set of manually arranged scenes created by non-expert users in a similar setting.

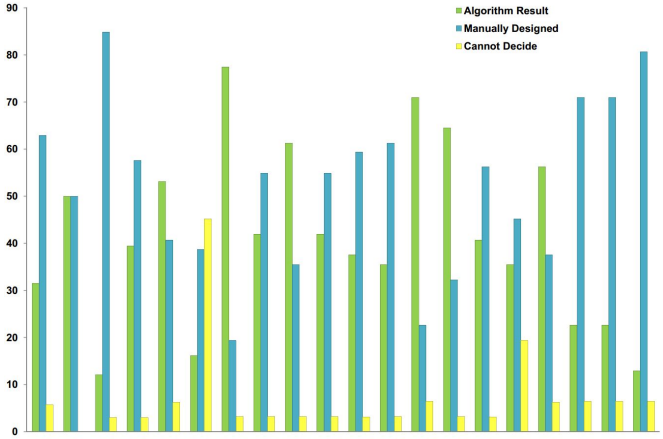


**Figure 8:** *The results for the comprehensive evaluation for 20 scenes. The y axis states the percentage of votes for our method, manually designed, and the 'Cannot Decide' option as green, blue, and yellow bars, respectively.*

In this study, for a collection of 20 scenes, randomly chosen among the ones generated by our approach, we asked a user to arrange the same set of objects for each room without providing him/her any prior examples of plausible scenes. Next, we ask users to rate between pairs of results, one by our algorithm and the other by a human, which one is more plausible. Figure 8 summarizes the results. On average, 52% of the votes were assigned to manually designed scenes while object arrangements produced by our method were selected 41% of times, leaving 7% for the cases that the user could not recognize which scene is more plausible. These numbers demonstrate we are very close to human-level performance.

### 7.3. Higher-order Relations and Constraints Evaluation

To demonstrate the effectiveness of salient relations and constraints learned in our model, we performed a second user study focusing on different modes of higher-order relations. The synthesized scenes in the previous study are split into four subgroups. For each subgroup, a comparing arrangement is reproduced with one or two types of salient relations being removed. In the experiment, participants are presented with two views of each scene in a pair and asked to select the more plausible one. The results are plotted in Figure 9. It shows that integrating each class of higher-order relations results in a significant improvement in the plausibility test, especially for the "symmetry and orientation", and "side-to-side" relations.

**Symmetry Relations.** Excluding symmetry from the set of salient relations in the model does not considerably affect the plausibility of the final arrangement; however, it might lead to choosing different 3D models for symmetric objects. We observed that objects in a symmetry group are normally associated through other types of higher-order relations, which results in considering them while determining the position and the orientation of one of the objects in the group. To illustrate, in the first column of Figure 10, although the nightstands are not placed perfectly symmetrically around the

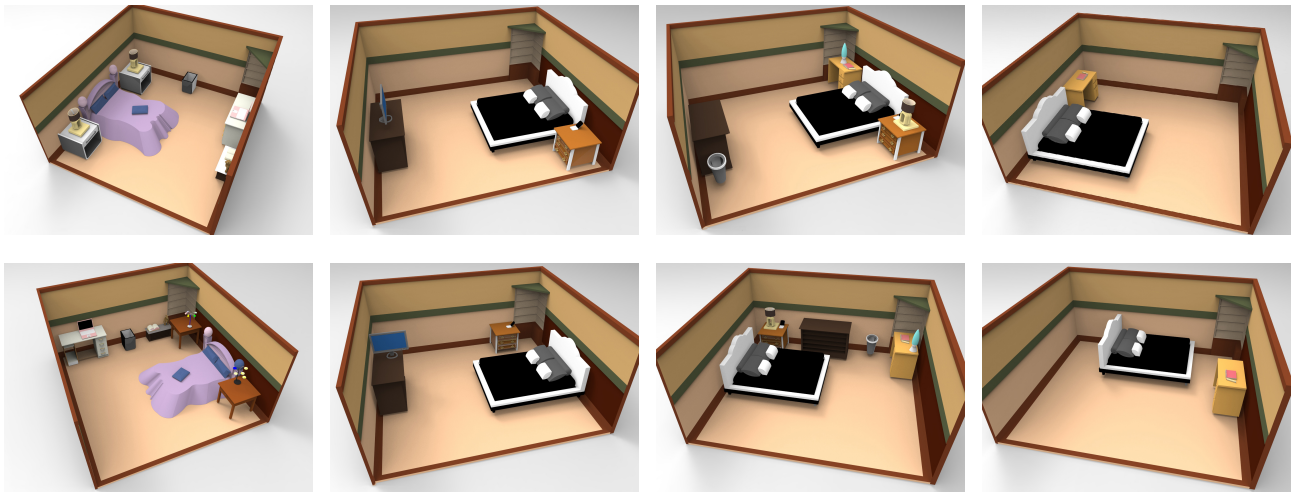

**Figure 10:** *Comparison with different parameter settings. In each column, the top row represents the arrangement produced by our algorithm and the bottom row is related to discarding (a) symmetry, (b) orientation, (c) symmetry and orientation, (d) side-to-side relations, from left to right, respectively.*

bed, they are arranged on two sides of the bed and relatively close to the wall.

**Distinct Orientation Relations.** To represent the influence of this category, consider the second column in Figure 10 in which an incorrect orientation is assigned to the TV. To go into detail, the pair *(bed, TV)* is not incorporated in optimization steps for the TV as an outcome of removing the oppositely oriented relation from the model.

**Symmetry and Orientation Relations.** If we cross out both symmetry and orientation from the list of salient relations, all the other objects will be taken into account to obtain an ideal arrangement for an object. As demonstrated in the third column in Figure 10, this behavior might induce undesirable output such as the orientation of the nightstand in this room. In this example, the desk and the shelf are inserted before the nightstand because of their larger sizes and consequently, pairs *(desk, nightstand)* and *(shelf, nightstand)* lead to the improper pose for the nightstand. On the other hand, in our approach the above pairs do not influence the arrangement process for the nightstand since there are no salient symmetry or orientation relations between them.

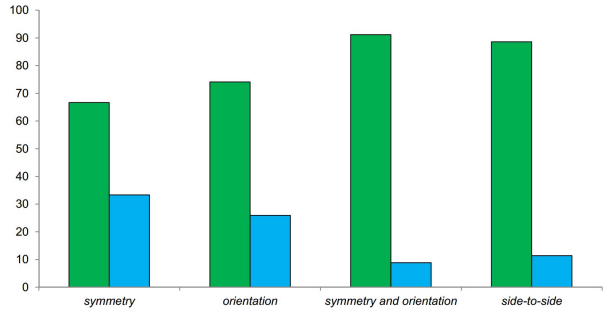


**Figure 9:** *For each mode of higher-order relations, the green bar demonstrates the output for the general setting while the blue bar is for the percentage of votes for eliminating the specific relation from the model.*

**Side-to-Side Constraints.** Although the salient relations provide a solution for relative placement of objects, the holistic arrangement is not counted as plausible unless side-to-side constraints are added to the model. Even for a simple scene like the one shown in the last column in Figure 10, the position of the bed is not entirely suitable since its back side is too far from a wall and it is almost in the middle of the room. In contrast, side-to-side relations constrain almost a zero distance between the back of a bed and a wall.

### 7.4. Comparison with the Arrangement Model in [FRS*12]

The work that comes the closest to ours is Fisher's object arrangement model in [FRS*12]. Note that although the general models share a set of similarities in two methods, they differ when it comes to the details and applying the models. To indicate that these variations make a notable difference in the final result, we attempted to compare the two methods in a similar setting. Since justifying plausibility for two scenes with both different sets of objects and different placements is more tricky, we only compared the arrangement models of two approaches for the same set of objects. We also factor out the influence of the source of data by adapting their arrangement model to utilize SUN RGB-D dataset. Furthermore, we trained a set of GMMs for pairwise spatial relations and implemented hill climbing to optimize the object placement. We produced arrangements for half of the scenes in the first study by their approach and sought users' opinions on the plausibility of each pair of scenes. Since we do not have any exemplar scenes as the input in our algorithm, we assumed a zero weight for examples in their work.

The user study results indicate that the users significantly prefer our method over Fisher's arrangement model for holistic plausibility. Quantitatively, our results were favored 93% of times while their approach received 7% of users' preferences. We inspected that stating pairwise relations only as coordinates of the centroids of

objects in Fisher's model causes the issue of not pushing objects to walls when it is essential. To illustrate, consider the arrangements in the top row of Figure 11. Although Fisher's arrangement model succeeds in finding a plausible relative arrangements between objects, the scene is not voted as holistically plausible, since objects are not pushed to walls. As explained in Section 5.2, side-to-side constraints in our model prohibits similar cases.

A closer investigation reveals that for a continuous optimization such as hill climbing in Fisher's approach, a richer source of data is required to avoid unclear distributions for pairwise relations. We instead applied MCMC sampling to provide a more clear image of a plausible arrangement. Furthermore, when all the pairs are considered, a large number of free parameters is involved in the optimization task which might fail to find a plausible solution, namely the placement for the pink stool and the orientation for the brown dresser in the bottom row of Figure 11. In contrast, we only consider pairs with at least one of the salient higher-order relations to prevent misleading the optimization process.

## 8. Discussion and future work

We propose a data-driven 3D indoor scene synthesis scheme to automatically select and place objects progressively. The salient higher-order object-object relations learned by our models, along with the rich source of annotated RGB-D images utilized during learning, distinguish our work from previous example-based and probabilistic learning methods for 3D scene synthesis. Subjective evaluations validate our argument that although progressive synthesis offers more local controllability, it ensures global coherence and holistic plausibility of the scene as well.

The key observation which motivated our work at the start is that all the automatically synthesized scenes from existing works appeared to be quite simple and clean, while real scenes which surround us can be rather cluttered and untidy. Since untidy placement of many objects does not necessarily exhibit clear global patterns, it may not be easy to learn a clean global probabilistic model. To this end, a progressive placement based on local object-object relations, like the one adopted by our work, may be more promising.

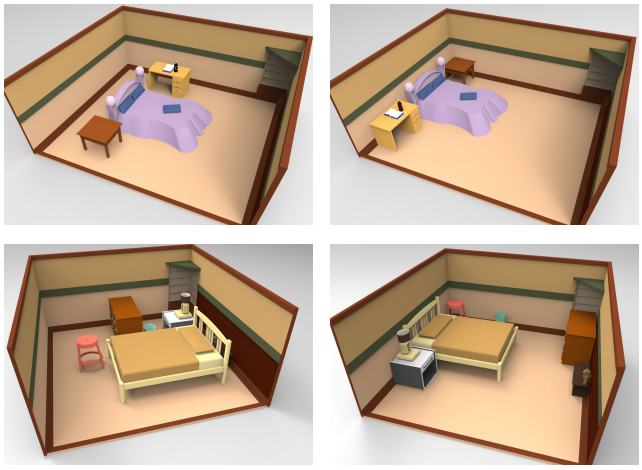

**Figure 11:** *Comparison of Fisher's arrangement model (the left column), against our approach (the right column).*



**Limitations and Future Work.** Ultimately, the success of all data-driven approaches critically depends on the richness of the data source. Relying on a relatively small number of hand-crafted 3D scenes, which are mostly clean to start with, does limit the variability and messiness of the synthesized scenes. Our work learns the synthesis model from real scenes, in the form of RGB-D images, and from a much larger number of them compared to previous attempts. Even so, we believe that we are still far from being able to synthesize highly clustered and untidy scenes, and as the results demonstrate, the set of selected objects are not diverse enough to be accounted a rich dataset. One possible future work is to tap into the vast amount of scene photographs; some of these, with annotations, are available from the Microsoft COCO [LMB*14] and the Visual Genome datasets [KZG*16]. However, learning 3D object placements from photographs offers a different set of challenges.

In addition to the small number of detailed items in the synthesized scenes, there are other limitations that differentiate the current set of results from real-world data. Particularly, our algorithm is designed for perfect rectangular rooms and is not applicable to other shapes, e.g. L-shaped rooms. As another instance, consider the chair and the desk in the right scene in Figure 1. Normally, a chair is partially occluded by a desk or a table; however, the hard constraint of intersection between bounding boxes of objects prevent such a setting. The algorithm might encounter failure in some cases. This can be seen when the algorithm attempts to insert a bed and a pair of night stands in the scene and orientation relations are removed from the model, which yields to a higher priority for the symmetry pair of two night stands over the bed and since the system is not able to determine a plausible location for the bed, it is removed in spite of its discriminating characteristics for a bedroom.

Another future direction stemming from this work is to render each generated scene from multiple viewpoints to obtain various RGB-D images. This will lead to supplying more accurate and noise-free training data for a set of computer vision tasks related to scene understanding. Technically, the synthesis algorithm can be improved, e.g., to speed up the sampling process and to learn more granular object-object relations, as well as a richer variety of higher-order structural or functional relations.

## Acknowledgments

We thank all the reviewers for their insightful comments and valuable suggestions. We owe our gratitude to students who took part in the user studies. We also acknowledge Jianxiong Xiao, Shuran Song, and Yinda Zhang for providing the SUN RGB-D dataset and the initial discussions. Thanks also go to Zhaopeng Cui and Warunika Ranaweera for the constructive discussions on this project and Ibraheem Alhashim, Ruizhen Hu, and Jaime Vargas-Trujillo for their assistance with proofreading the paper. This work is supported by grants from NSERC Canada (611370, 611649).

## References


[Bis07] BISHOP C.: Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn, 2007. 6

[CLW*14] 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 Trans. on Graphics (Proc. of SIGGRAPH Asia) 33*, 6 (2014), 208:1–12. 1

[FKLW97] FREY B. J., KSCHISCHANG F. R., LOELIGER H.-A., WIBERG N.: Factor graphs and algorithms. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing* (1997), vol. 35, UNIVERSITY OF ILLINOIS, pp. 666–680. 5

[FRS*12] FISHER M., RITCHIE D., SAVVA M., FUNKHOUSER T., HANRAHAN P.: Example-based synthesis of 3d object arrangements. *ACM Trans. on Graphics 31*, 6 (2012), Article 135. 2, 3, 4, 5, 7, 8

[FSL*15] FISHER M., SAVVA M., LI Y., HANRAHAN P., NIESSNER M.: Activity-centric scene synthesis for functional 3d scene modeling. *ACM Trans. on Graphics 34* (2015), 212:1–10. 1, 2, 3

[GJWW14] GUERRERO P., JESCHKE S., WIMMER M., WONKA P.: Edit propagation using geometric relationship functions. *ACM Transactions on Graphics (TOG) 33*, 2 (2014), 15. 2

[GJWW15] GUERRERO P., JESCHKE S., WIMMER M., WONKA P.: Learning shape placements by example. *ACM Transactions on Graphics (TOG) 34*, 4 (2015), 108. 2

[HZvK*15] HU R., ZHU C., VAN KAICK O., LIU L., SHAMIR A., ZHANG H.: Interaction context (icon): Towards a geometric functionality descriptor. *ACM Transactions on Graphics (Special Issue of SIGGRAPH) 34*, 4 (2015), Article 83. 5

[JKJ*13] JANOCH A., KARAYEV S., JIA Y., BARRON J. T., FRITZ M., SAENKO K., DARRELL T.: A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 141–165. 7

[JLS12] JIANG Y., LIM M., SAXENA A.: Learning object arrangements in 3d scenes using human context. *arXiv preprint arXiv:1206.6462* (2012). 3

[JLZS12] JIANG Y., LIM M., ZHENG C., SAXENA A.: Learning to place new objects in a scene. *The International Journal of Robotics Research* (2012), 0278364912438781. 3

[KMYG12] KIM Y. M., MITRA N. J., YAN D.-M., GUIBAS L.: Acquiring 3d indoor environments with variability and repetition. *ACM Trans. on Graphics 31*, 6 (2012), 138:1–138:11. 1

[KZG*16] KRISHNA R., ZHU Y., GROTH O., JOHNSON J., HATA K., KRAVITZ J., CHEN S., KALANTIDIS Y., LI L.-J., SHAMMA D. A., BERNSTEIN M., FEI-FEI L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. URL: http://arxiv.org/abs/1602.07332. 9

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÃAR P., ZITNICK C. L.: Microsoft coco: Common objects in context. In *Proc. Euro. Conf. on Computer Vision* (2014), pp. 740–755. 9

[LZW*15] LIU Z., ZHANG Y., WU W., LIU K., SUN Z.: Model-driven indoor scenes modeling from a single image. In *Proc. of Graphics Interface* (2015), pp. 25–32. 1, 2, 3

[MSL*11] MERRELL P., SCHKUFZA E., LI Z., AGRAWALA M., KOLTUN V.: Interactive furniture layout using interior design guidelines. *ACM Trans. on Graphics 30*, 4 (2011), 87:1–10. 1, 2

[NSF12] NATHAN SILBERMAN DEREK HOIEM P. K., FERGUS R.: Indoor segmentation and support inference from rgbd images. In *ECCV* (2012). 7

[Rou87] ROUSSEEUW P.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math. 20*, 1 (Nov. 1987), 53–65. URL: http://dx.doi.org/10.1016/0377-0427(87)90125-7, doi:10.1016/0377-0427(87)90125-7. 6

[SLX15] SONG S., LICHTENBERG S. P., XIAO J.: SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition* (2015), pp. 567–576. 2, 7

[SMZ*14] 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 Trans. Graph. 33*, 6 (2014), 209–1. 3

[XCF*13] XU K., CHEN K., FU H., SUN W.-L., HU S.-M.: Sketch2Scene: Sketch-based co-retrieval and co-placement of 3D models. *ACM Trans. on Graphics (Proc. of SIGGRAPH) 32*, 4 (2013), 123:1–10. 2, 3

[XMZ*14] XU K., MA R., ZHANG H., ZHU C., SHAMIR A., COHEN-OR D., HUANG H.: Organizing heterogeneous scene collection through contextual focal points. *ACM Trans. on Graphics (Proc. of SIGGRAPH) 33*, 4 (2014), 35:1–12. 4

[XOT13] XIAO J., OWENS A., TORRALBA A.: Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 1625–1632. 7

[YH02] YAN X., HAN J.: gspan: Graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on* (2002), IEEE, pp. 721–724. 4

[YYT*11] YU L.-F., YEUNG S. K., TANG C.-K., TERZOPOULOS D., CHAN T. F., OSHER S.: Make it home: automatic optimization of furniture arrangement. *ACM Trans. on Graphics (Proc. of SIGGRAPH) 30*, 4 (2011), 86:1–12. 1, 2

[ZJ10] ZHANG L., JI Q.: Image segmentation with a unified graphical model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 32*, 8 (2010), 1406–1425. 2, 5

[ZSTX14] ZHANG Y., SONG S., TAN P., XIAO J.: PanoContext: A whole-room 3D context model for panoramic scene understanding. In *Proc. Euro. Conf. on Computer Vision* (2014), vol. 6, pp. 668–686. 1