Propose-and-Complete: Auto-regressive Semantic Group Generation for Personalized Scene Synthesis

Shoulong Zhang[1,2]
shoulong.zhang@buaa.edu.cn
Shuai Li[1,2,*]
lishuai@buaa.edu.cn
Xinwei Huang[1,3]
e2lighten@buaa.edu.cn
Wenchong Xu[1]
19375332@buaa.edu.cn
Aimin Hao[1]
ham@buaa.edu.cn
Hong Qin[4,*]
qin@cs.stonybrook.edu

[1] Beihang University, Beijing, P.R. China
[2] Zhongguancun Laboratory, Beijing, P.R. China
[3] Kuaishou Technology, Beijing, P.R. China
[4] Stony Brook University (SUNY), Stony Brook, USA

Abstract

Our research goal is to build a novel scene synthesis framework enabling the flexible generation of individualized indoor virtual environments. Current deep methods only learn the layout patterns from training scene samples, affording only partial co-occurrence possibilities while ignoring any user intent. In contrast, this paper devises a novel framework by flexibly combining and generating function-oriented semantic object groups while accommodating strong user intent. Conforming to this group-centric design paradigm, we consider different strategies for proposing group-level locations and completing semantic clusters with intra-group relationships. The entire framework hinges upon two technical innovations. First, we design a conditional normalizing flow-based *ProposeNet* to learn the exact distribution of semantic groups, by which we sample potentially plausible group-level locations constrained by user-desirable room functionalities. Second, we design a conditional graph variational auto-encoder, *CompleteNet*, to instantiate each semantic group with the user-specific complexity (e.g., graph size). With the complete groups readily available, we then recursively select the most plausible proposals and optimize the final layout subject to a collision-free, accessible room space and an arbitrary floor plan. Comprehensive experiments have confirmed that our new framework can produce personalized and versatile unseen 3D scenes from a more expansive design space than conventional domains delimited by training data.

* Corresponding authors.

1 Introduction

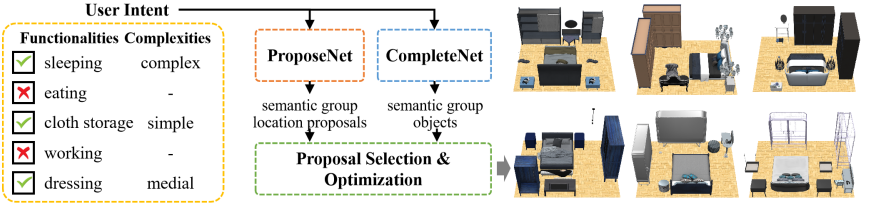

Figure 1: We present a novel scene synthesis framework combining compact semantic object groups based on user intent. The ProposeNet proposes plausible group-level locations, and the CompleteNet generates the intra-group detailed arrangement. Our propose-and-complete strategy flexibly explores a more expanded design space than training data (e.g., bedroom, living room) and produces versatile scenes given a personalized setting.

Problem Statement and Background. As a human-built environment, indoor space is created to support versatile desired functions to facilitate our daily activities. This demand requires generative methods to produce realistic, plausible, and functional indoor scenes that must be flexible and personalized. Recent deep-learning methods have achieved impressive results in scene synthesis. Most previous works learn the scene-level layouts consistently and memorize frequent room patterns from the trained samples. Nonetheless, after training with typical domestic scenes (e.g., bedroom, living room), the deep models can hardly produce a creative combination with personalized settings. The CNN-based methods [24, 29, 37] adopt a top-down heatmap or 2.5D image to guide the 3D layout creation. Another line of research uses GNN-based [4, 30, 38] or transformer-based [23, 31] architectures to learn scene-level prior and generate scene graphs or object sequences. Although there are human-centric approaches [2, 7, 10, 11, 14, 25, 36] that enhance the flexibility by sampling various active agents, the human-involved data preparation is quite time-consuming. In short, the main challenge is to design a powerful model with high versatility that also supports custom-made functional indoor scenes, possibly beyond the design space delimited by the training data.

Method Motivation and Overview. We propose a novel method that generates and combines user-specific functional semantic groups to tackle the aforementioned challenge. In particular, our method needs to answer two questions: (1) how the semantic groups support an expansive indoor design space, and (2) how to ensure the user intent controls the generative process. For the first question, we observe that the semantic function is a natural clustering that compactly fills the space delicately [6, 32, 33]. Conversely, we can enrich an empty room by adding various semantic groups with flexible complexities. Thus, we adopt a propose-and-complete strategy. The *ProposeNet* aims to learn the group-level location distributions, where we sample new group position and rotation proposals. Regardless of room type (e.g., bedroom, living room), we can obtain the candidates of any arbitrary combination of the semantic groups. The *CompleteNet* generates the instances inside the groups respecting the intra-group relationships and respective complexities (i.e., object numbers). This strategy enables exploring a more expansive indoor design space beyond training data by supporting theoretically infinite combinations of room functionalities, continuous group locations, and semantic group complexities. For the second question, ProposeNet takes the user-specific semantic group categories as the condition to learn the group-level location distributions by a conditional normalizing flow model. Similarly, CompleteNet also takes the

user-desired object number to guide a conditional graph variational auto-encoder to generate the exact instances of the group. Once the group location proposals and the semantic groups are generated, we recursively selected the most proper proposal for each group by scoring them with potential collision, object accessibility, area of the free zone, and compatibility with the floor shape. Since the selected proposals may still be unqualified, we adopt an extra optimization process with the same scoring criteria to produce the final room layout.

Key Contributions. The primary contributions of this paper comprise: (1) We propose a novel flexible propose-and-complete scene generation framework that supports personalized semantic function settings; (2) We design a normalizing flow-based ProposeNet for learning the group-level position and rotation distributions, where we sample the plausible semantic group locations conditioning on user-specific functionalities; (3) We devise a conditional graph variational auto-encoder, CompleteNet, to generate entire semantic groups with user-desired group complexities. Our novel generation strategy and networks successfully produce plausible indoor designs respecting personalized settings beyond the training set. Moreover, our model can also achieve comparable performance with the SOTA method while generating typical domestic indoor rooms. From the application perspective, our extensive experiments and thorough evaluations validate all the claimed advantages of our novel approach toward the effective production of custom-made 3D indoor scenes with high versatility and personalized functionality.

2 Related Work

Image-guided Scene Generation. Deep learning in the image process has achieved impressive performance, which is also widely used in scene synthesis. Wang et al. proposed a method to add a new instance based recursively on 2D convolutional prior from informative top-down masks [29]. Similarly, Ritchie et al. introduced a fast and flexible scene synthesis method (FastSynth) using four convolutional sub-modules predicting the object category, location, orientation, and scales [24]. Luo et al. designed a scene instantiation model based on an abstract scene graph and a 2.5D sketch guidance [21]. Zhang et al. combined handcrafted and learned descriptors as a hybrid representation [37] to generate realistic layouts with a generative adversarial network using the top-down image [1, 12]. Yang et al. learned scene volume from a collection of 2.5D partial observations [34]. Since the 2D image guidance has difficulty representing specific 3D configurations, our method directly regresses 3D bounding boxes to produce satisfactory scene layouts.

Graph and Sequence based Scene Generation. Scene graph [17, 26, 28], tree, and sequence structures expedite the representation and learning of the relationships among instances. Li et al. proposed a recursive auto-encoder (GRAINS) [19]. It encodes various relations as intellectual nodes and generates the following hierarchy by trained variational auto-encoders (VAE) [5]. Wang et al. adopted a graph auto-regressive model that uses spatial edges to constrain the scene graphs [30]. Zhou et al. proposed a graph-based scene augmentation model (SceneGraphNet) [38], which predicts the most plausible properties of the newly-added instance. Dhamo et al. devised a detailed scene edition method by manipulating latent graph features [4]. Recently, researchers adopted transformer [3] architecture to generate a scene as a sequence. Wang et al. proposed the SceneFormer to auto-regressively produce an ordered sequence based on the layout image features [31]. Similarly, Paschalidou et al. introduced an auto-regressive transformer (ATISS) that generates rooms as unordered sets of objects, supporting scene completion and failure case detection [23]. Our method

adopts a graph-based generative model to learn the intra-group relationships and sample newly semantic groups.

Semantic group-based Scene Analysis. There are also semantic group-based approaches analyzing highly-related instance clusters as the primary element to construct the scene space. Fisher et al. executed a very early approach that clustered the interchangeable objects in a dataset based on their neighborhood similarity [6]. Xu et al. proposed the "structural group" concept as a functional object set with reliable relationships, which enabled sketch-based object co-retrieval and co-placement [33]. Xu et al. introduced an extraction method of representative substructure, named "focal point", which is used for characterizing and comparing scene collections [32]. Unlike previous works that extract and analyze the functional substructure of scenes, our novel method directly takes the annotated groups as known prior and learns the distribution of intra-group patterns to organize personalized indoor space.

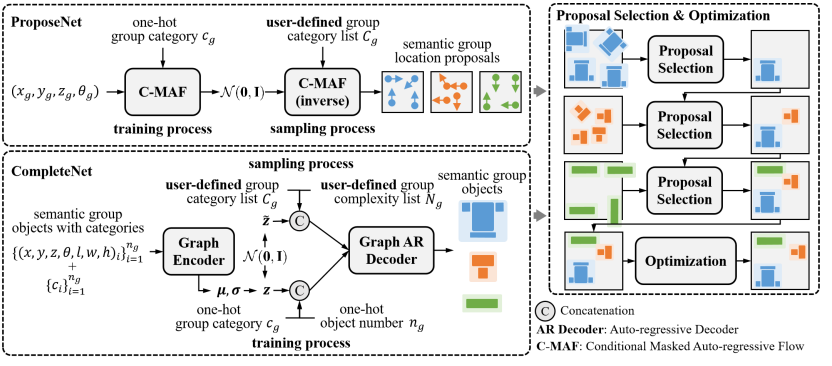

Figure 2: The overall framework. In this example, we take three semantic groups and top-down 2D representation for better illustration, while the real outputs are in 3D space. The ProposeNet takes the group category as condition and learns the group-level locations distributions by a C-MAF model. We can sample possible group locations conditioning on user-specific room functionalities. The CompleteNet is a variational graph auto-encoder, which conditions on both group category and complexity (i.e., node number). Given the possible group locations and the instantiated groups, we recursively select the *best* proposal by fitting the group to all proposed locations and ranking them with plausibility criteria. We adopt the same criteria for the final optimization process.

3 Novel Approach

The overall framework structure is illustrated in Figure 2. Given the user-defined semantic groups with desired complexity, we first sample group location proposals from the ProposeNet (Sec 3.1), and instaintiate the specific intra-group objects by the CompleteNet (Sec 3.2). Finally, we choose the most proper group locations and optimize them with our proposed selection and optimization algorithm (Sec 3.3) given the outputs of the networks.

ProposeNet. We decompose a room space into a compact set of semantic groups, which are defined as the set of objects for executing certain daily activities. Each group location has a descriptive vector $\mathbf{x} = (x_g, y_g, z_g, \theta_g)$ of position and orientation around its vertical axis with all values normalized in [-1, 1]. The group position is the average of the object centers,

and the orientation coincides with the most dominant object. We assume that the group location of category c_g follows a conditional density $p(\mathbf{x}|c_g)$. In order to learn this density and sample new group proposals, we design a conditional normalizing flow model, ProposeNet, based on the masked auto-regressive flow (MAF) [22]. As preliminary, the normalizing flow models [16, 18] infer a density $p(\mathbf{x})$ as an invertible differentiable transformation f of a basic density $p_u(\mathbf{u})$ (e.g., $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$), which is $\mathbf{x} = f(\mathbf{u})$ where $\mathbf{u} \sim p_u(\mathbf{u})$. Based on the invertibility of the transformation f, the desity $p(\mathbf{x})$ can be represented as:

$$p(\mathbf{x}) = p_u(f^{-1}(\mathbf{x}))\left|\det\left(\frac{\partial f^{-1}}{\partial \mathbf{x}}\right)\right|. \tag{1}$$

In order to maintain tractability, the MAF adopts an auto-regressive framework with Gaussian conditions. The i-th latent channel is conditional by previous variables. For more discussion and proofs, we refer readers to the work of MAF model [22]. We extend the original MAF model towards a conditional version based on our conditional density $p(\mathbf{x}|c_g)$ assumption. Thus, we modify the auto-regressive model by considering the extra conditions as $p(x_i|\mathbf{x}_{1:i-1}, c_g) = \mathcal{N}(x_i|\mu_i, (\exp\alpha_i)^2)$. The calculations of the mean and the log standard in each recursion are:

$$\mu_i = \phi_{\mu_i}(\mathbf{x}_{1:i-1}, c_g),\ \alpha_i = \phi_{\alpha_i}(\mathbf{x}_{1:i-1}, c_g). \tag{2}$$

We train the ProposeNet with the same loss function design as in MAF [22]. In the sampling process, given the user-specified category list C_g, we sample N_p group location proposals for each semantic group category. The ProposeNet outputs all the location candidates as $\mathcal{P} = \{\{\mathbf{x}_{c_g}^i\}_{i=1}^{N_p}\}_{c_g \in C_g}$. Since the user selects the room functionalities as the conditioning feature, our ProposeNet supports sampling the arbitrary combination of group location proposals regardless of the semantic co-occurrence of the training rooms.

CompleteNet. We build the CompleteNet with a variational graph auto-encoder structure to generate the group objects, conditioning the group semantics and its target complexity. The semantic group is modeled as a directed, fully-connected graph $G = (V, E)$ considering the intra-group relationships. The vertex set V includes all the bounding boxes of the group objects. Each bounding box has a descriptor of position (x, y, z), orientation θ around its vertical axis, scale (l, w, h), and object category c. We define the group complexity as the graph node number. All the objects are normalized and aligned in the group coordinate system, where the original point coincides with the group's center. The encoder is a three-layer graph neural network. The initial node feature is the encoded bounding box descriptor through a two-layer MLP. We aggregate the node features by using the message passing algorithm [13, 20] with the attention mechanism [27]. The message $\mathbf{m}_v$ received by the node $\mathbf{n}_v$ is computed as:

$$\mathbf{m}_v = \frac{1}{N_v}\sum_u w_{uv}\text{LeakyRelu}(\mathbf{n}_u - \mathbf{n}_v), \tag{3}$$

where N_v is the number of edges where $\mathbf{n}_v$ is the target node. The w_{uv} is the attentive weight computed by a three-layer MLP f_w following by sigmoid function. The node features are updated by the gated recurrent unit (GRU). We use the sum pooling for the node and the graph-wise features. We decode the graph structure in an auto-regressive style. We first sample a new code from a Gaussian distribution and reparameterize it with predicted mean and variance as a latent code $\mathbf{z}$. The input latent code of the decoder is the concatenation of $\mathbf{z}$, the one-hot-encoded group category c_g, and the node number n_g. At step t, the newly added node

feature is initialed by the last node feature and the latent code as $\mathbf{n}_t = \phi(\text{cat}(\mathbf{n}_{t-1}, \mathbf{z}, c_g, n_g))$, where ϕ is a two-layer MLP. The $\mathbf{n}_t$ is then updated through a graph neural network with the same architecture as the encoder. At step $t = 0$, we replace the last node feature with the input latent code. For the final bounding box output, we discretize the continuous position (x, y, z) and the scale (l, w, h) into 64 units and the orientation angle θ into 8 units, respectively. We employ a group of MLPs to estimate seven channels of the bounding box descriptor and the object category. The frequency of the object semantic category in the training set orders the node generation process. The cross-entropy loss optimizes the discretized layout and object class outputs. In the sampling process, given the user-defined category list C_g and the complexity list N_g, the CompleteNet outputs the semantic groups as graphs $\mathcal{G} = \{G_{c_g}\}_{c_g \in C_g}$.

Proposal Selection and Optimization. We recursively select the most *proper* location for each group category by scoring the candidates with several common-sense criteria [35]. Our scoring is based on four criteria: inter-group collision, group accessibility, free zone area, and compatibility of the floor shape. Additionally, we employ a discretized strategy to efficiently quantify the constraints mentioned above. Considering a group category $c_g \in C_g$, we generate a set of 2D grid checkpoints P_{c_g} at the floor plane for the output semantic group G_{c_g}, with spacing d ($d = 0.125$ in our evaluation). The floor equips with checkpoints P_{floor} covering only the supportive area.

(1) *Inter-group Collision*: Various functional regions should not overlap in a versatile room space. The inter-group collision can be measured by the intersection of two individual functional elements. We first define $f_n(P_I, P_J)$ the count of *intersected* points of two checkpoints sets P_I and P_J. The two points are considered as intersected if the spacing distance less then d. Supposing P_{sel} is the checkpoint set of the previously selected semantic groups. The normalized inter-group collided point count, N^i_{col,c_g}, for the i-th proposal of the category c_g as $f_n(P^i_{c_g}, P_{\text{sel}})/|P^i_{c_g}|$.

(2) *Accessibility*: The semantic group should be accessible by the human or the other agents to execute its functionality. Especially for the highly-oriented semantic groups (e.g., storages), the interactive surface might be blocked by other groups located in its functional area, and even all semantic groups involve no inter-group collision. We tackle this problem by extending the checkpoints in the pre-defined interactive area of each group. Thus, the potential accessibility problem can be solved along with the inter-group collision.

(3) *Free Zone Area*: In practice, the user usually places the furniture near the walls to obtain a broader activity zone and aesthetic room space. In order to effectively quantify the free zone area, we inversely count and minimize the checkpoints on the floor behind the semantic group back surface, denoted as $P^i_{\text{back},c_g} \subset P_{\text{floor}}$. The normalized checkpoint count, N^i_{free,c_g}, for quantifying the area of the free zone is $|P^i_{\text{back},c_g}|/|P^i_{c_g}|$.

(4) *Compatibility with Floor*: Our proposal selection strategy is suitable for various floor shapes because the coverage of P_{floor} varies along with the specific room shape. To measure the compatibility with the floor, we conversely count the group checkpoints out of the floor, N^i_{out,c_g}, as $1 - f_n(P^i_{c_g}, P_{\text{floor}})/|P^i_{c_g}|$.

The final score for the i-th proposal of category c_g is:

$$s^i_{c_g} = 1 - N^i_{\text{col},c_g} - N^i_{\text{free},c_g} - N^i_{\text{out},c_g}. \tag{4}$$

We recursively choose the proposal with the largest score as the best selection from the most occupying groups. This algorithmic heuristic order choice is a valid technique proven by our experiment, but might not be theoretical guaranteed. Empirically, it might happen that all proposals are not ideal. Thus we additionally optimize the group positions following

the selection process. Specifically, we sample a delta translation $(\delta x_g, \delta z_g)$ on the floor plan from a Gaussian distribution $\alpha_{\text{opt}}\mathcal{N}(0, \sigma_{\text{opt}})$, where $\alpha_{\text{opt}} = 0.25$ and $\sigma_{\text{opt}} = 0.1$ based on our empirical research. The group position is updated when the optimized group achieves a larger score defined in Eq. 4.

4 Experiments and Evaluations

Dataset Preparation and Implementation Details. We train our networks on the currently popular 3D-FRONT indoor scene dataset [8], composed with 3D-FUTURE furniture dataset [9]. We first pre-process the dataset, filtrate the rooms with reasonable size, and manually refine the objects with collisions. Additionally, to cooperate with the group-based training procedures, we re-annotate 4989 rooms by dividing the scenes into compact semantic groups. In total, we obtain 16763 semantic groups of 12 daily functionality categories. The split of training and testing the ProposeNet is 80% and 20%, respectively. The semantic groups of the training rooms are used to optimize the CompleteNet. The pre-process, the re-annotation, and the exemplars of semantic groups are detailly introduced in the supplementary material.

Implementation Details. We train our models on a personal computer platform with an Nvidia GTX 1080Ti GPU on the Pytorch framework. We train the ProposeNet and the CompleteNet for 1000 epochs with the Adam optimizer with 0.0001 as weight decay. The initial learning rate is 0.002, and the rate decays by 0.9 for every 50 epochs for both modules. During training the CompleteNet, empirically, the reconstruction loss weight is 1 while the KL divergence loss weight is 0.3. In our evaluation, the ProposeNet generates $N_p = 50$ for each semantic group category, and the maximum optimization iteration is 500.

Baselines. We compare our method to FastSynth [24] and ATISS [23] with their published implementations[1]. FastSynth is an image-guided approach that recursively generates scene by predicting the subsequent object arrangement based on the top-down image scene representation. The ATISS is a sequence-based SOTA method with a transformer backbone. We re-train both methods with the same dataset settings for a fair comparison.

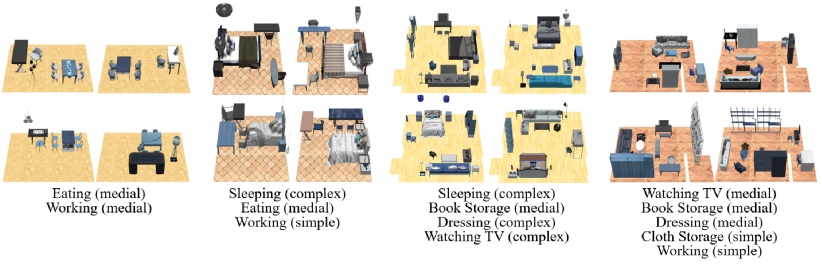

Figure 3: Qualitative evaluation of our method with various personalized group settings, which are beyond the patterns of the training data.

Qualitative Evaluation on Scene Synthesis. We execute two qualitative evaluations to validate our method's effectiveness and generative ability. In the first evaluation, we adopt personalized input settings to generate new scene layouts beyond the scene types of the

[1] We reproduce the FastSynth and ATISS methods based on their released code (https://github.com/brownvc/fast-synth and https://https://github.com/nv-tlabs/ATISS), respectively.

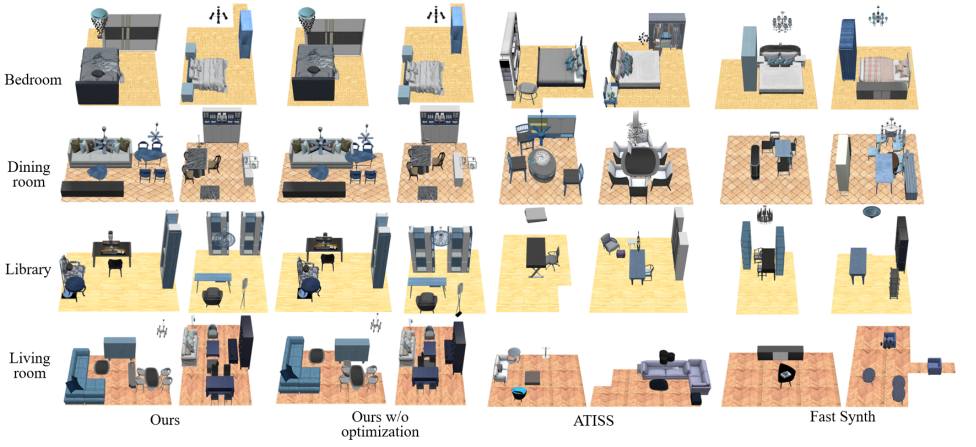

Figure 4: Qualitative evaluation of our method and the ablated version without the optimization process, comparing with the results of ATISS and FastSynth.

trained data. As shown in Figure 3, our method generates the selected semantic groups and arranges them in a collision-free and visually convincing fashion while supporting different layout sizes. Furthermore, this illustration also shows that our method can produce semantic groups of various complexity, guaranteeing the variety of the instantiation process, which recalls our main contributions. In the second evaluation, we exhibit the generated scene layouts by FastSynth, ATISS, and our method in Figure 4. The visualization includes the four scene types in the dataset. To generate the classic rooms, since our method requires additional user intent, we randomly adopt the group categories and object numbers from the test set to evaluate our model's generalization ability given unseen semantics and complexities. As we can see from this illustration, our method generates the most plausible and reasonable 3D layouts compared with other methods.

Quantitative Evaluation on Scene Synthesis. In order to measure the quality of the generated scenes, we calculate the Fréchet Inception Distance (FID) [15] between the orthographic projections of the synthetic rooms and the test set rooms as in [23]. We also compute the KL divergence of the instance categories between the test set and the synthesized scenes. It is shown in Table 1 that our method achieves a comparable score with the ATISS for bedrooms and the minor scores compared with other generative models for the other room types. Additionally, our method provides the closest semantic distributions except for the bedroom types compared with other methods. Besides, we also provide the FIDs and KL divergence between the test and the training samples to show the intrinsic bias due to the data split. The quantitative experiment does not emphasize the performance gain since we do not focus on improving the existing works. The intent is to validate our framework's property that can achieve SOTA performance with a different group-based approach.

Perceptual Study. We conduct a perceptual study with 60 two-alternative forced-choice questions and 40 realistic ranking questions, similar to GRAINS [19]. Twenty-seven anonymous voluntary participants with different technical backgrounds were involved in this perceptual study: sixteen graphics/vision researchers and eleven non-technical subjects. The comparison results are shown in Fig. 5 (a). The layouts sampled from our method are 61.12% more plausible than the FastSynth and 16.30% more realistic than the ATISS. Our method

	FID (↓)				KL (↓)			
	Bedroom	Dining	Library	Living	Bedroom	Dining	Library	Living
Training Set	*56.147*	*67.318*	*64.755*	*49.546*	*0.0095*	*0.0086*	*0.0259*	*0.0066*
FastSynth	68.871	80.956	76.469	76.629	0.0968	0.0679	0.0784	0.0732
ATISS	**60.304**	75.838	69.074	57.580	**0.0052**	0.0083	0.0249	0.0087
Ours (w/o opt.)	61.769	75.683	69.022	57.025	-	-	-	-
Ours ($N_p = 50$)	60.810	**74.408**	**68.340**	**55.344**	0.0145	**0.0078**	**0.0201**	**0.0075**
Ours ($N_p = 30$)	63.890	74.643	72.270	56.135	-	-	-	-
Ours ($N_p = 10$)	65.424	74.821	73.741	57.157	-	-	-	-

Table 1: The FID scores between the real data and the synthesized results on 256^2 pixels. KL divergences between the real and the predicted groups. We provide the score of the training set as baseline. Lower is better.

can synthesize the layouts most similar to the ground truth among the three evaluated methods. Furthermore, we show the scene's realistic scoring results in the box-whisker chart in Fig. 5 (b). Our method averages the best realistic score among the three evaluated methods. Our new approach is collectively validated as an effective scene-generation method.

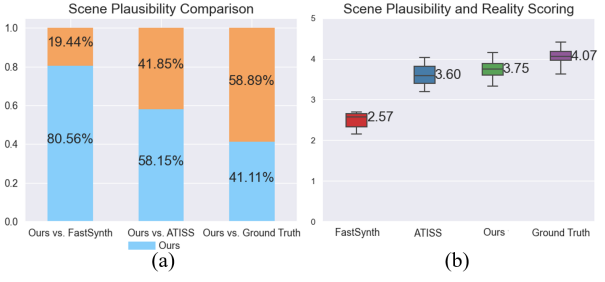

Figure 5: Perceptual study results.

Model Complexity and Time Consumption. Our method is the most lightweight one among all the evaluated methods. Our deep model has only 1.727 million trainable parameters with a relatively small latent dimension and avoids computationally expensive convolution and multi-head attention operations. Furthermore, our networks are also the most efficient. We randomly sampled 100 scenes of bedroom type with the tested methods. Our networks reduce 32.16% of inference time compared with ATISS. However, the post-optimization phase dramatically costs a large part of the time, limiting our efficiency. The detailed network parameters and inference time comparison are reported in Table 2. We also report the time consumption with 100 and 500 as the maximum optimization iteration.

Methods	FastSynth	ATISS	Ours w/o opt	Ours (opt ×100)	Ours (opt ×500)
Params	38.180 (0.00%)	36.053 (5.57%↓)	**1.727** (**95.48%**↓)	-	-
Time	1.309 (0.00%)	0.171 (86.94%↓)	**0.116** (**91.14%**↓)	0.889 (32.09%↓)	4.013 (206.56%↑)

Table 2: Network parameters (in millions) and time consumption (in seconds). The percentages are the relative improvement comparing with the baseline method FastSynth.

Ablation Study. We evaluate the ablated versions of our novel framework for a better understanding. In the first experiment, we remove each criterion in the proposal selection and optimization phase. Fig. 6 illustrates the final scene layouts ignoring collision, free zone area, and floor compatibility scores, respectively. It illustrates that the criteria can explicitly

reduce the inter-group collision, maximize the free zone, and guarantee the scene objects placed inside the floor plan. In the second experiment, we conduct scene generation under various proposal numbers (i.e., N_p). Table 1 shows that our model requires a relatively large proposal number to provide enough samples for the selection and optimization phase.

Figure 6: Ablation study on propose selection and optimization criteria.

Limitation. Our novel method has several limitations. Firstly, our model has a minor performance boost compared with the ATISS model. Nonetheless, our main contribution to the community focuses on realizing a flexible and individualized scene synthesis process instead of only improving the typical room synthesis. Another potential limitation is that the output scene layouts could be overloaded since we do not limit the number of semantic groups. The users could set a large number of the desired groups, which might cause collisions. As a result, the final results would be out of control. Additionally, a post-optimization process could be a time-cost operation that needs improvement with our future efforts.

5 Conclusion, Discussion, and Future Work

This paper has articulated a novel propose-and-complete framework supporting custom-made scene layout generation with high versatility. The key innovation is founded upon a flexible combination of indoor functional semantic groups, with which we propose potential group-level locations by the ProposeNet and complete the detailed intra-group objects by the CompleteNet in a divide-and-conquer fashion. Given sampled potential group locations and generated group objects, we selected the *best* group location proposal using four plausibility criteria. We then optimized the chosen locations toward a realistic and reasonable indoor room. Extensive experiments validate all the perceived advantages of our new approach. Our near-term research efforts are geared towards immediate improvement with better spatial solutions, optimization efficiency, and a user-friendly graphical interface. Furthermore, our work assumes an equivalence relation between the functionality and semantic group with user intent bridging the above two entities. Although proven valid in practice, our assumption might cause information loss without further room functionality analysis, which is still an open question.

Acknowledgment

This research is partially supported by supported by National Natural Science Foundation of China (No.62272021), and National Science Foundation of USA:IIS-1715985 and USA:IIS-1812606.

References

[1] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

[2] Carlo Bretti and Pascal Mettes. Zero-shot action recognition from diverse object-scene compositions. In *BMVA Press BMVC*, page 281, 2021.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *ACL NAACL-HLT*, pages 4171–4186, 2019.

[4] Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *IEEE/CVF ICCV*, pages 16332–16341, 2021.

[5] Otto Fabius, Joost R. van Amersfoort, and Diederik P. Kingma. Variational recurrent auto-encoders. In *ICLR Workshop*, 2015.

[6] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas A. Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. *ACM TOG*, 31(6):135:1–135:11, 2012.

[7] Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *ACM TOG*, 34(6): 179:1–179:13, 2015.

[8] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. In *IEEE/CVF ICCV*, pages 10913–10922, 2021.

[9] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Stephen J. Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *IJCV*, 129(12):3313–3337, 2021.

[10] Qiang Fu, Xiaowu Chen, Xiaotian Wang, Sijia Wen, Bin Zhou, and Hongbo Fu. Adaptive synthesis of indoor scenes via activity-associated object relation graphs. *ACM TOG*, 36(6):201:1–201:13, 2017.

[11] Qiang Fu, Hongbo Fu, Hai Yan, Bin Zhou, Xiaowu Chen, and Xueming Li. Human-centric metrics for indoor scene assessment and synthesis. *Graphical Models*, 110: 101073, 2020.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[13] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pages 1024–1034, 2017.

[14] Pat Hanrahan. Scenegrok: inferring action maps in 3d environments. *ACM TOG*, 33 (6):212:1–212:10, 2014.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6626–6637, 2017.

[16] Christian Horvat and Jean-Pascal Pfister. Denoising normalizing flow. In *NeurIPS*, pages 9099–9111, 2021.

[17] Salman Khan and Fabio Cuzzolin. Spatiotemporal deformable scene graphs for complex activity detection. In *BMVA Press BMVC*, page 278, 2021.

[18] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, pages 10236–10245, 2018.

[19] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao (Richard) Zhang. GRAINS: generative recursive autoencoders for indoor scenes. *ACM TOG*, 38(2):12:1–12:16, 2019.

[20] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. In *ICLR*, 2016.

[21] Andrew Luo, Zhoutong Zhang, Jiajun Wu, and Joshua B. Tenenbaum. End-to-end optimization of scene layout. In *IEEE/CVF CVPR*, pages 3753–3762, 2020.

[22] George Papamakarios, Iain Murray, and Theo Pavlakou. Masked autoregressive flow for density estimation. In *NIPS*, pages 2338–2347, 2017.

[23] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: autoregressive transformers for indoor scene synthesis. In *NeurIPS*, pages 12013–12026, 2021.

[24] Daniel Ritchie, Kai Wang, and Yu-An Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *IEEE/CVF CVPR*, pages 6182–6190, 2019.

[25] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM TOG*, 35(4):139:1–139:12, 2016.

[26] Xi Tian, Yongliang Yang, and Qi Wu. Enhancing person synthesis in complex scenes via intrinsic and contextual structure modeling. In *BMVA Press BMVC*, page 491, 2022.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[28] Hai Wan, Yonghao Luo, Bo Peng, and Wei-Shi Zheng. Representation learning for scene graph completion via jointly structural and visual embedding. In *IJCAI*, pages 949–956, 2018.

[29] Kai Wang, Manolis Savva, Angel X. Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM TOG*, 37(4):70:1–70:14, 2018.

[30] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X. Chang, and Daniel Ritchie. Planit: planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM TOG*, 38(4):132:1–132:15, 2019.

[31] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *IEEE 3DV*, pages 106–115, 2021.

[32] Kai Xu, Rui Ma, Hao Zhang, Chenyang Zhu, Ariel Shamir, Daniel Cohen-Or, and Hui Huang. Organizing heterogeneous scene collections through contextual focal points. *ACM TOG*, 33(4):35:1–35:12, 2014.

[33] Kun Xu, Kang Chen, Hongbo Fu, Wei-Lun Sun, and Shi-Min Hu. Sketch2scene: sketch-based co-retrieval and co-placement of 3d models. *ACM TOG*, 32(4):123:1–123:15, 2013.

[34] Mingjia Yang, Yu-Xiao Guo, Bin Zhou, and Xin Tong. Indoor scene generation from a collection of semantic-segmented depth images. In *IEEE/CVF ICCV*, pages 15183–15192, 2021.

[35] Lap-Fai Yu, Sai Kit Yeung, Chi-Keung Tang, Demetri Terzopoulos, Tony F. Chan, and Stanley J. Osher. Make it home: automatic optimization of furniture arrangement. *ACM TOG*, 30(4):86, 2011.

[36] Suiyun Zhang, Zhizhong Han, Yu-Kun Lai, Matthias Zwicker, and Hui Zhang. Active arrangement of small objects in 3d indoor scenes. *IEEE TVCG*, 27(4):2250–2264, 2021.

[37] Zaiwei Zhang, Zhenpei Yang, Chongyang Ma, Linjie Luo, Alexander Huth, Etienne Vouga, and Qixing Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM TOG*, 39(2):17:1–17:21, 2020.

[38] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *IEEE/CVF ICCV*, pages 7383–7391, 2019.