Human-centric Indoor Scene Synthesis Using Stochastic Grammar

Siyuan Qi[1] Yixin Zhu[1] Siyuan Huang[1] Chenfanfu Jiang[2] Song-Chun Zhu[1]
[1] UCLA Center for Vision, Cognition, Learning and Autonomy
[2] UPenn Computer Graphics Group

Contribution

We present a human-centric method to sample and synthesize 3D room layouts and 2D images thereof, for the purpose of obtaining large-scale 2D/3D image data with the perfect per-pixel ground truth. An attributed spatial And-Or graph (S-AOG) is proposed to represent indoor scenes with human contexts encoded as contextual relations.

Representation

We use an attributed S-AOG to represent an indoor scene. It combines i) a probabilistic context free grammar (PCFG), and ii) contextual relations among terminal nodes defined on an Markov Random Field (MRF), *i.e.*, the horizontal links among the nodes.

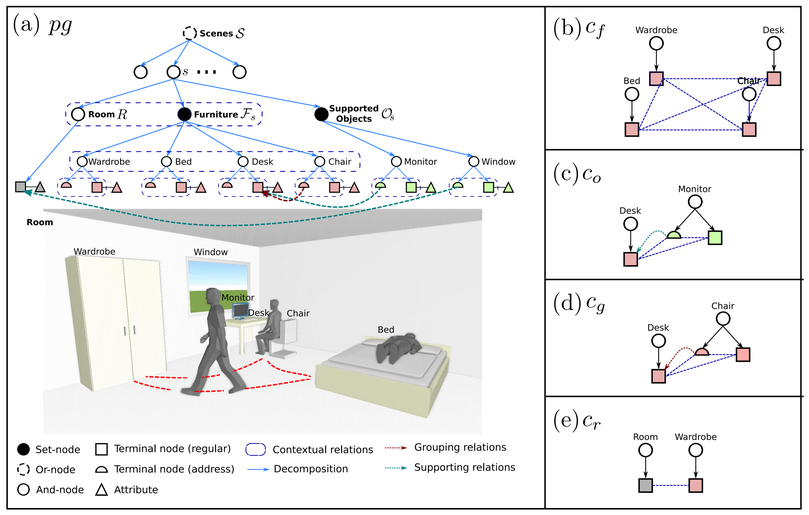


Figure 1: (a) A simplified example of a parse graph of a bedroom. The terminal nodes of the parse graph form an MRF in the terminal layer. Cliques are formed by the contextual relations projected to the terminal layer. (b)-(e): Four types of cliques representing different types of contextual relations.

Formulation

A scene configuration is represented by a parse graph pg, including objects in the scene and associated attributes. The prior probability of pg generated by an S-AOG parameterized by Θ is formulated as a Gibbs distribution:

$$p(pg|\Theta) = \frac{1}{Z}\exp\{-\mathcal{E}(pg|\Theta)\} = \frac{1}{Z}\exp\{-\mathcal{E}(pt|\Theta) - \mathcal{E}(E_{pt}|\Theta)\}, \quad (1)$$

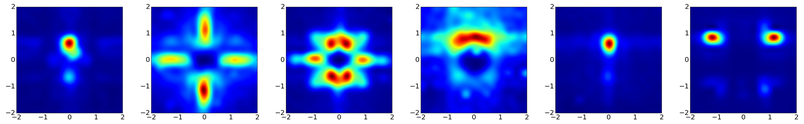
where $\mathcal{E}(pg|\Theta)$ is the energy function of a parse graph, $\mathcal{E}(pt|\Theta)$ is the energy function of a parse tree, and $\mathcal{E}(E_{pt}|\Theta)$ is the energy term of the contextual relations, given by:

$$p(E_{pt}|\Theta) = \frac{1}{Z}\exp\{-\mathcal{E}(E_{pt}|\Theta)\} = \prod_{c\in C_f}\phi_f(c)\prod_{c\in C_o}\phi_o(c)\prod_{c\in C_g}\phi_g(c)\prod_{c\in C_r}\phi_r(c), \quad (2)$$

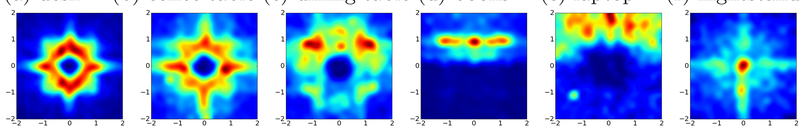
where $\phi_f(c)$, $\phi_o(c)$, and $\phi_g(c)$ are computed through human contexts. The weights for each potential term is learned by contrastive divergence.

Human Context

The contextual relations encode functional grouping relations and supporting relations modeled by object affordances. For each object, we learn the affordance distri- bution, i.e., an object-human relation, so that a human can be sampled based on that object. Besides static object affordance, we also consider dynamic human activities in a scene, constraining the layout by planning trajectories from one piece of furniture to another.



(a) desk (b) coffee table (c) dining table (d) books (e) laptop (f) nightstand



(g) fruit bowl (h) vase (i) floor lamp (j) wall lamp (k) fireplace (l) ceiling fan

Figure 2: Examples of the learned affordance maps. Given the object positioned in the center facing upwards, *i.e.*, coordinate of $(0, 0)$ facing direction $(0, 1)$, the maps show the distributions of human positions. They accurately capture the subtle differences among desks, coffee tables, and dining tables. Some objects are orientation sensitive, while some are orientation invariant.

Sampling

Synthesizing scene configurations is accomplished by sampling a parse graph pg from the prior probability $p(pg|\Theta)$ defined by the S-AOG. We use Markov chain Monte Carlo (MCMC) to draw a typical state in the distribution.

Figure 3: MCMC sampling of scene configurations with simulated annealing.

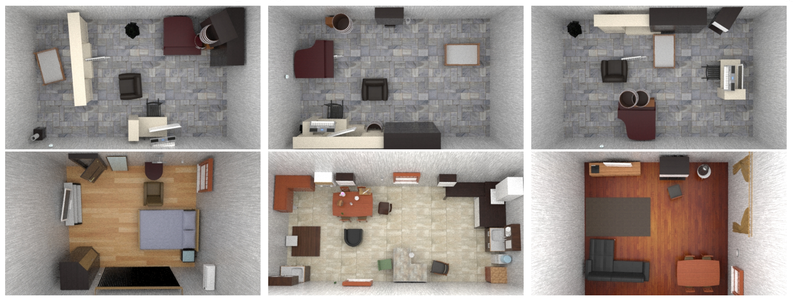


Figure 4: **Top**: previous methods only re-arranges a given scene with a fixed room size and a predefined set of objects. **Bottom**: our method samples a large variety of scenes.

Qualitative Results

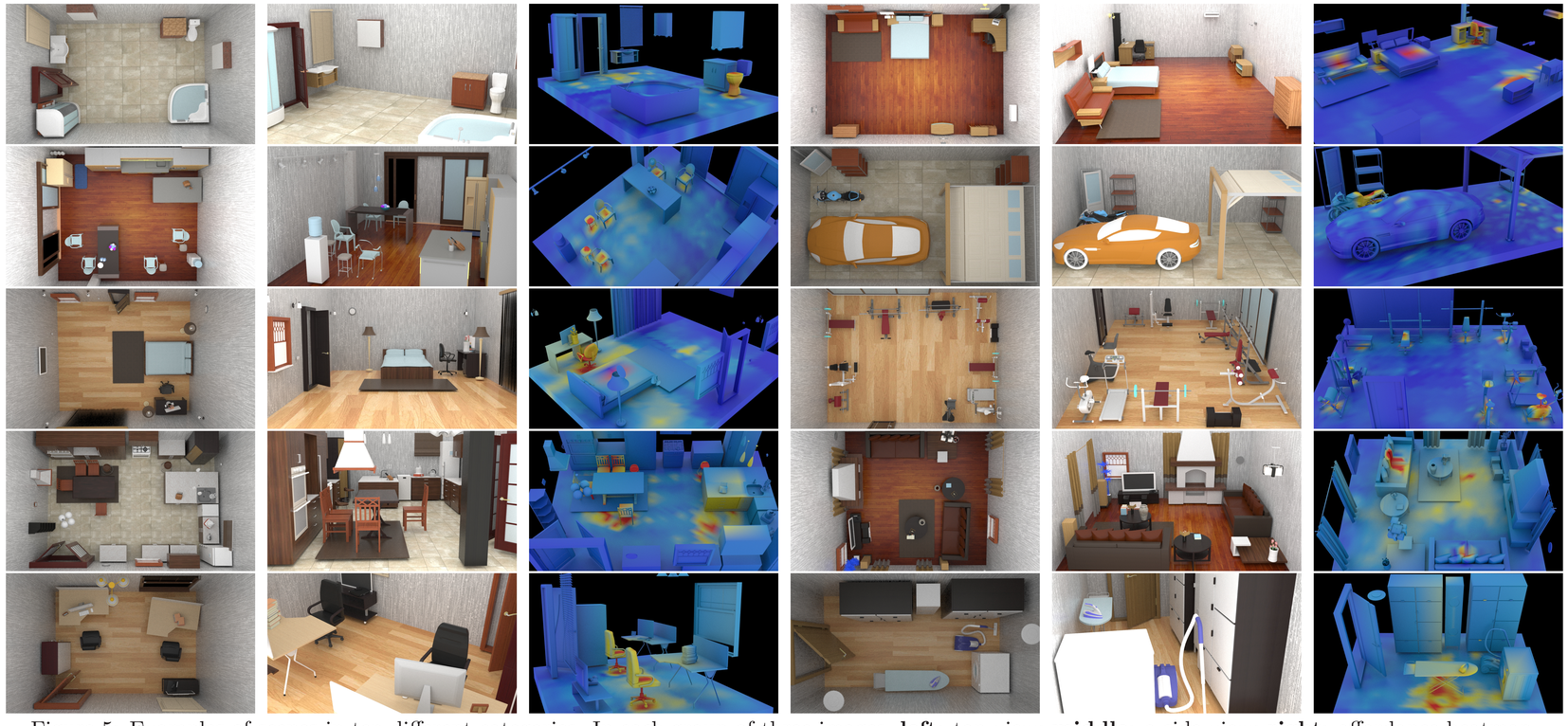


Figure 5: Examples of scenes in ten different categories. In each group of three images, **left**: top-view; **middle**: a side-view; **right**: affordance heatmap.