IEEE 2015 Conference on Computer Vision and Pattern Recognition

Object-Level Generative Models for 3D Scene Understanding

Ehsan Jahangiri, René Vidal, Laurent Younes, Donald Geman
Center for Imaging Science, Johns Hopkins University.

1. Introduction:

- A core challenge in computer vision is to develop generative models of the world that capture rich contextual relationships among scene entities. Such models can serve different applications:
  - **Scene Understanding:** regularizing the output of image descriptors in a Bayesian framework, generating sequences of unpredictable queries for testing computer vision systems (see the visual Turing test by Geman et al. [1]).
  - **Robotics:** Simultaneous Localization and Mapping (SLAM), path planning, grasping and manipulating objects.
  - **Computer Graphics:** creating synthetic content.
- Many man-made scenes are composed of multiple parallel supporting surfaces upon which instances from different object categories are placed [2].
- Designing and learning (from purely object-annotated images) 3D models which encode favored relationships but still accommodate real-world variability is not straightforward.
- We propose a new probabilistic, generative model of 3D scenes consisting of multiple objects lying on a plane. Our distribution is over random **"Generative Attributed Graphs (GAG)"** that encode favored layouts while accounting for variations in the number and relative poses of objects.

2. Proposed Model:

- A scene is described as a collection of object instances from different categories at different poses. Each object instance is associated with a vertex $v \in V$ of a base graph $g_0 \in G_0$ which captures contextual relationships among object instances.
- An attributed graph is a triple $g = (g_0, c_V, \theta_V)$, where $c_V = \{c_v,\ v \in V\}$ and $\theta_V = \{\theta_v,\ v \in V\}$ denote the set of category labels and 3D poses of objects, respectively.
- The model is a probability distribution on the space of attributed graphs conditioned on the environment's geometric properties T, specified by four sets of distributions:

1. $p^{(0)}(n_{(0,1)}, \ldots, n_{(0,K)}|T)$: conditional joint distribution for the number of root nodes from each object category.
2. $\{p^{(c)}(n_1, \ldots, n_K),\ c \in C\}$: the joint distribution of the number of children from each object category (Multi-type branching process). Restricted by a "Master Graph".

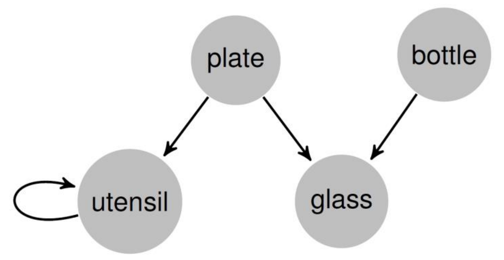
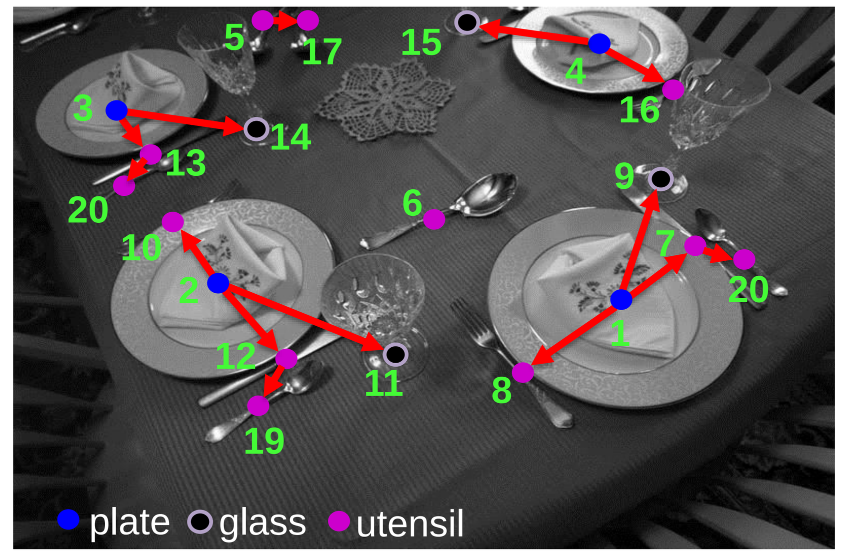


Fig 1. An Example "Master Graph"



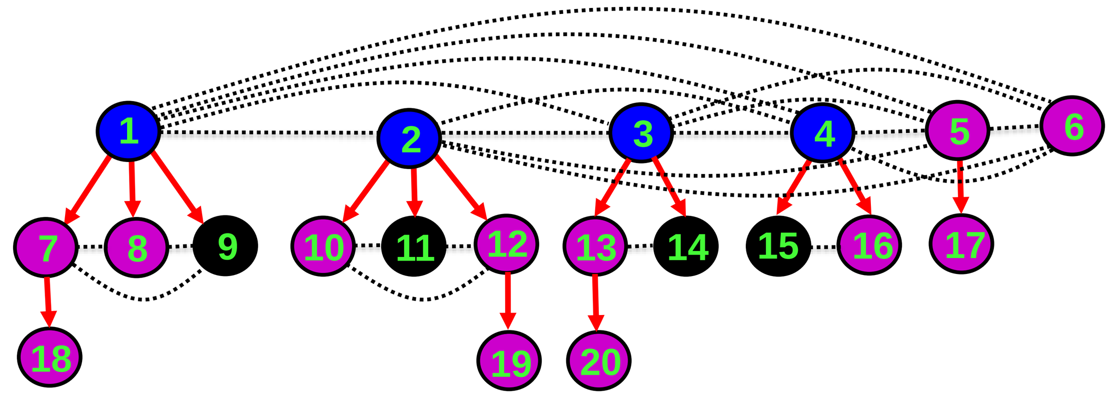


Fig 2. A table-setting scene (left) and its corresponding category-labeled base graph (right).

3. $p(\theta_{V_0}|c_{V_0}, T)$: the joint distribution of the poses of the root nodes given T.
4. $\{p\big(\theta_{ch(v)}|c_{ch(v)}, c_v, \theta_v, T\big),\ v \in V\backslash V_T\}$: the joint distribution of the poses of the children of v given their parent's pose and the corresponding category labels and T.

The full distribution on attributed graphs $g \in \mathcal{G}$:

$$\boxed{\begin{aligned} p(g|T) &= p(g_0, c_V|T) \times p(\theta_V|g_0, c_V, T) \\ &= p^{(0)}(n_{(0,1)}, \cdots, n_{(0,K)}|T)\ p(\theta_{V_0}|c_{V_0}, T) \times \\ &\quad \textstyle\prod_{v\in V\backslash V_T} p^{(c_v)}(n_{(v,1)}, \cdots, n_{(v,K)})\ p(\theta_{ch(v)}|c_{ch(v)}, c_v, \theta_v, T). \end{aligned}}$$

3. Model Learning:

- From annotated scenes:
  - *Observable:* $\mathcal{D} = \{c_V[j], \theta_V[j]\}_{j=1}^{J}$, and *Hidden:* $\mathcal{M} = \{g_0[j]\}_{j=1}^{J}$
- Parameter Estimation Using Expectation-Maximization (EM):

$$\Phi^{t+1} = \mathrm{argmax}_{\Phi}\ \ \textstyle\sum_{j=1}^{J} \sum_{g_0[j]} p(g_0[j] \mid c_V[j], \theta_V[j], \Phi^t) \times \log p(g[j] \mid \Phi)$$

- Stochastic Expectation-Maximization using MCMC:

$$\Phi^{t+1} \approx \mathrm{argmax}_{\Phi}\ \ \textstyle\sum_{j=1}^{J} \sum_{l=1}^{N} \log p(g^{(l)}[j] = (g_0^{(l)}[j], c_V[j], \theta_V[j]) \mid \Phi)$$

Gibbs Sampling $p(g_0[j] \mid c_V[j], \theta_V[j], \Phi^t)$:

- Conditional base graph distribution:

$$p(g_0[j] \mid c_V[j], \theta_V[j], \Phi^t) = \frac{1}{Z(c_V[j], \theta_V[j], \Phi^t)} \times p(g_0[j], c_V[j], \theta_V[j] \mid \Phi^t)$$

- For a scene with $|V|$ annotated objects we can encode the base graph with:

$$\mathbf{z} = \big(z_1 = pa(v_1), z_2 = pa(v_2), \ldots, z_{|V|} = pa(v_{|V|})\big), \qquad g_0 = f(\mathbf{z})$$

Step-1. Begin with initial configuration i ← 1, l ← 1, and:

$$\mathbf{z}^{(0)} = (z_1^{(0)} = \emptyset, \ldots, z_{|V|}^{(0)} = \emptyset)$$

Step-2. Sweep **z** by sampling one element at the time according to:

$$p(z_i) = \frac{p(g_0^{(l)}(z_i), c_V, \theta_V)}{\sum_{z_i \in S_{pa(v_i)}} p(g_0^{(l)}(z_i), c_V, \theta_V)},\ \ z_i \in S_{pa(v_i)}$$

Step-3. Generate the corresponding base graph sample $g_0^{(l)}$.
Set $l \leftarrow l + 1$ and $i \leftarrow (l \bmod |V|)$ and go back to Step-2.

4. Table-Setting Scenes:

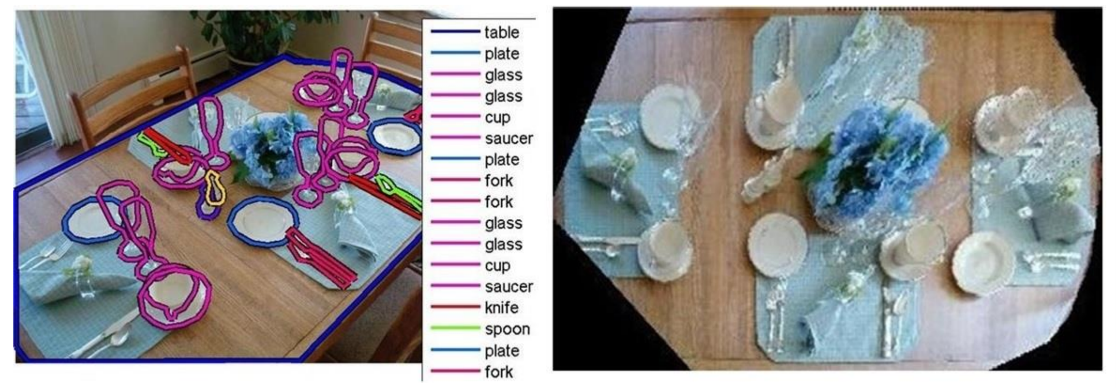
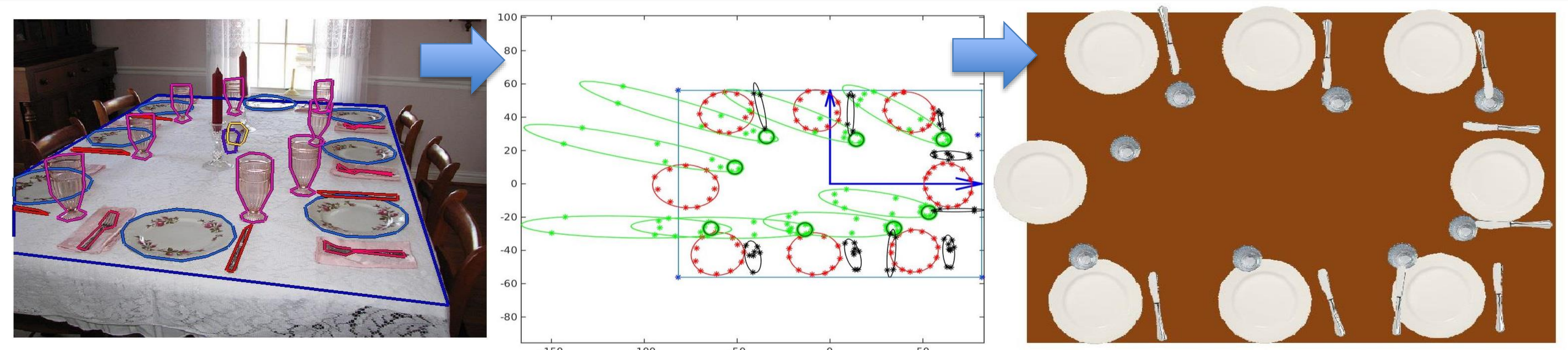


Fig 3. JHU table-setting dataset including > 3000 fully annotated images and the corresponding manually estimated homographies.

Fig 4.



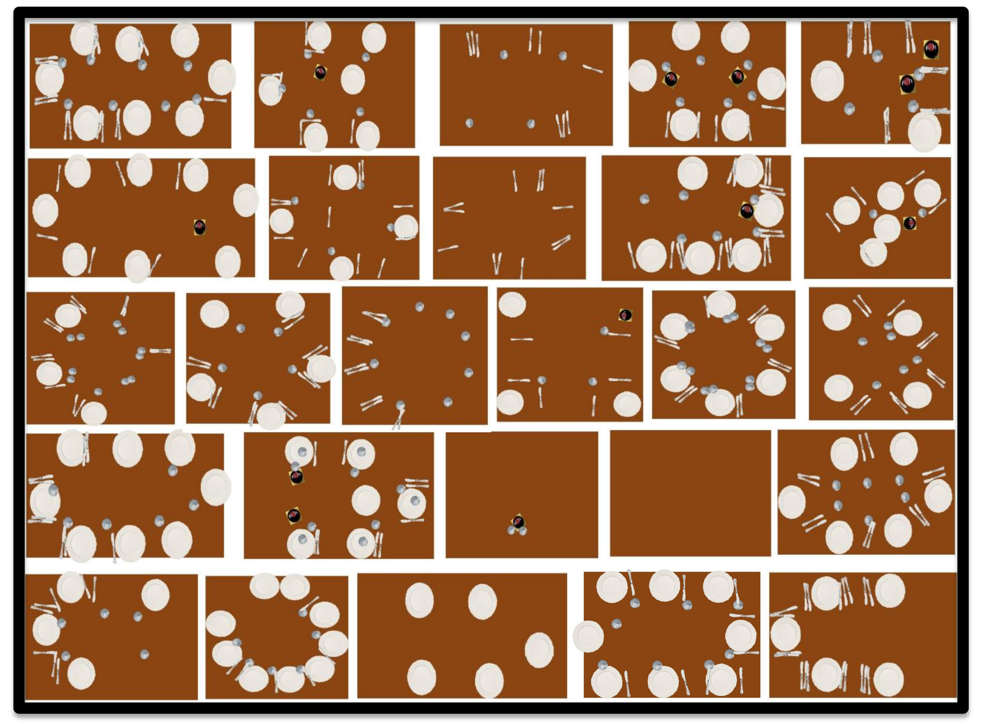


Fig 5. Top-view visualization of some annotated images.

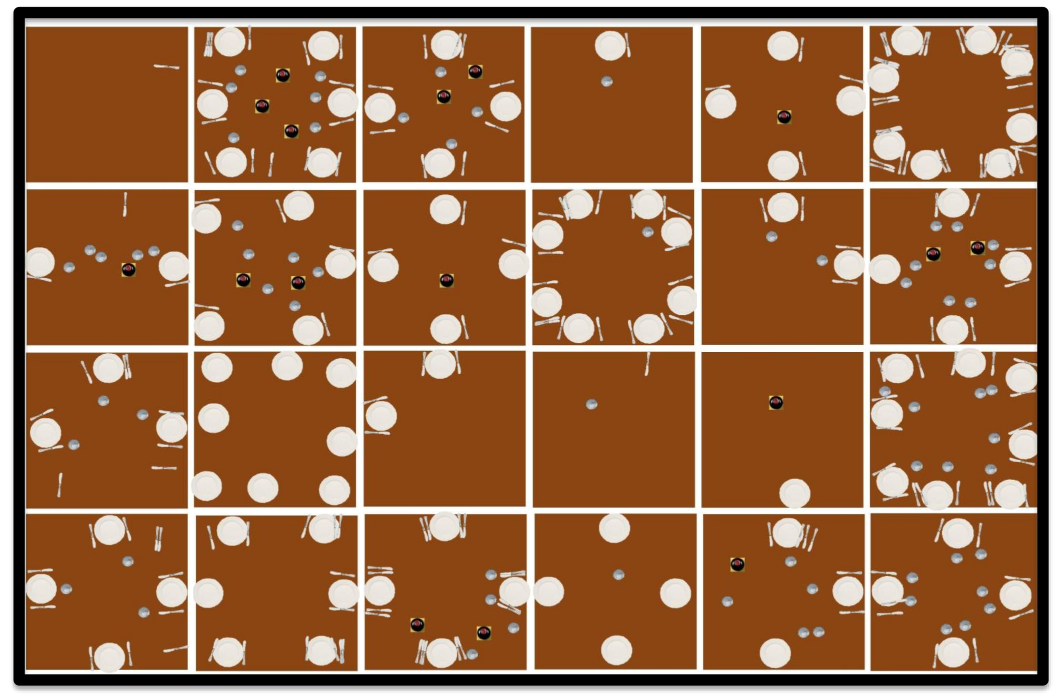


Fig 6. Top-view visualization of model samples.

2D model from 3D model: $p(\xi_V, g, T, \mathcal{W}) = p(\xi_V|g, \mathcal{W})\, p(g|T)\, p(\mathcal{W})$

References:
[1] D. Geman, et al. "A visual turing test for computer vision systems". In PNAS, 2014.
[2] S. Y. Bao, et al. "Toward coherent object detection and scene layout understanding". In CVPR, 2010.