Auto-Encoding Scene Graphs with Image Reconstruction

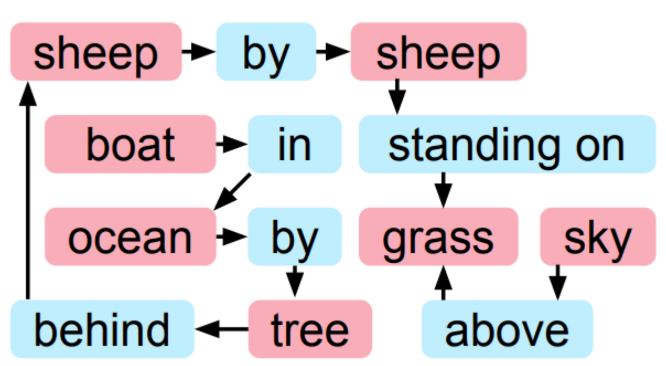
Kevin Tan

kevintan@cs.stanford.edu

Stanford
Computer Science

Motivation

- Goal: Learn to generate scene graphs from images and reconstruction by conditional image synthesis
- To fully understand the visual world requires not only recognizing individual objects, but also inferring the relationships and interactions between them
- Modern deep generative models excel in producing high-quality samples, but fall short in capturing compositional, object-centric semantic structure in visual scenes



Contributions

- We propose a method for scene graph generation in an end-to-end trainable fashion via *image reconstruction supervision*
- Auto-Encoding Scene Graphs consists of two parts:
- (1) Encoder: given an input image, proposes object regions by a region proposal network, prunes connections with relational proposal network, and aggregates contextual information via graph convolution, and outputs a scene graph
- (2) Decoder: Given a scene layout, feed noise into a cascaded refinement network to perform conditional image synthesis

Problem Statement

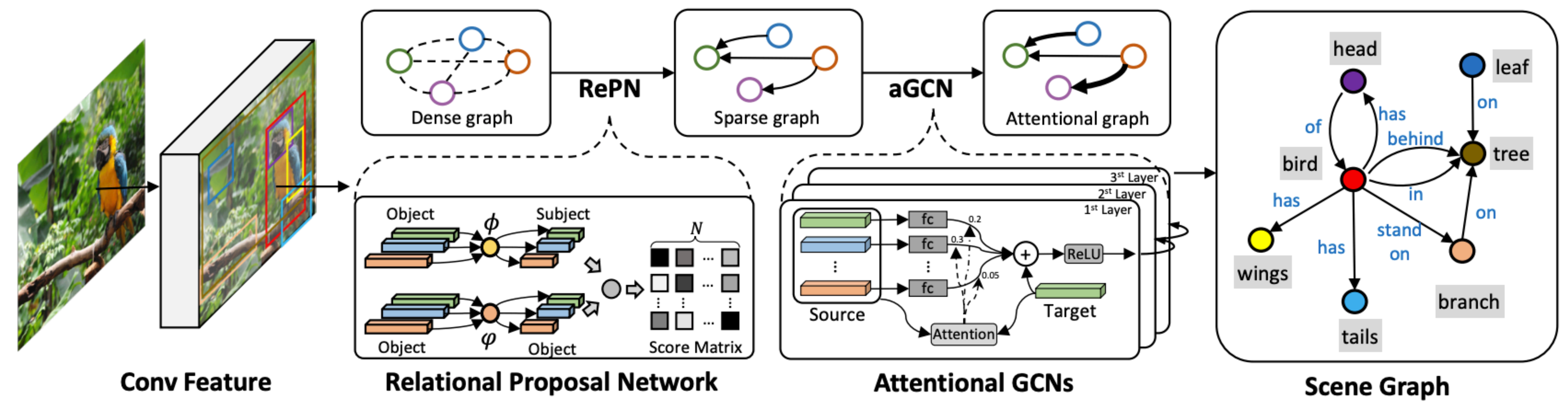
More formally, a *scene graph* can be defined by a 3-tuple set $G = \{B, O, R\}$:

- $B = \{b_1, b_2, ..., b_n\}$ is the region candidate set, with elements $b_i \in \mathbb{R}^4$ denoting the bounding box of the i^{th} region
- $O = \{o_1, o_2, ..., o_n\}$ is the object set, with element $o_i \in \mathbb{N}$ denoting the corresponding class label of region b_i
- $R = \{r_1, r_2, ..., r_m\}$ of pairwise relationships between those objects, where r_k denotes a triplet of a start node $(b_i, o_i) \in B \times O$, an end node $(b_j, o_j) \in B \times O$, and a relationship label $x_{i \to j} \in \mathcal{R}$, where $\mathcal{R}$ is the set of all possible predicate types.

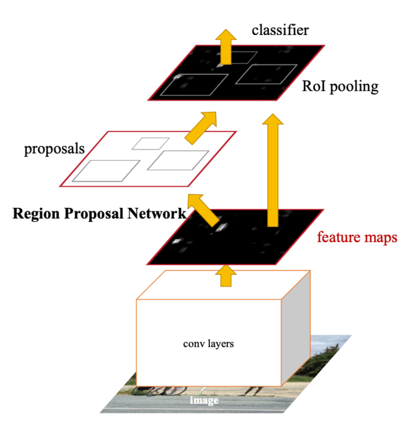
Given an image I, the goal is to decompose the probability distribution of the scene graph $P(G \mid I)$ into three components, as demonstrated previously by [17]:

$$Pr(G \mid I) = Pr(B \mid I) Pr(O \mid B, I) Pr(R \mid O, B, I) \quad (1)$$

Scene Graph Generation



Object Proposals with Faster R-CNN



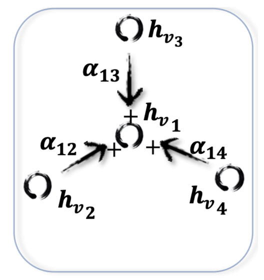
Relation Proposal Network

$$f(\boldsymbol{p}_i^o, \boldsymbol{p}_j^o) = \langle \Phi(\boldsymbol{p}_i^o), \Psi(\boldsymbol{p}_j^o) \rangle, i \neq j$$

$$IoU(\{u, v\}, \{p, q\}) = \frac{I(r_u^o, r_p^o) + I(r_v^o, r_q^o)}{U(r_u^o, r_p^o) + U(r_v^o, r_q^o)}$$

Graph Convolution Network

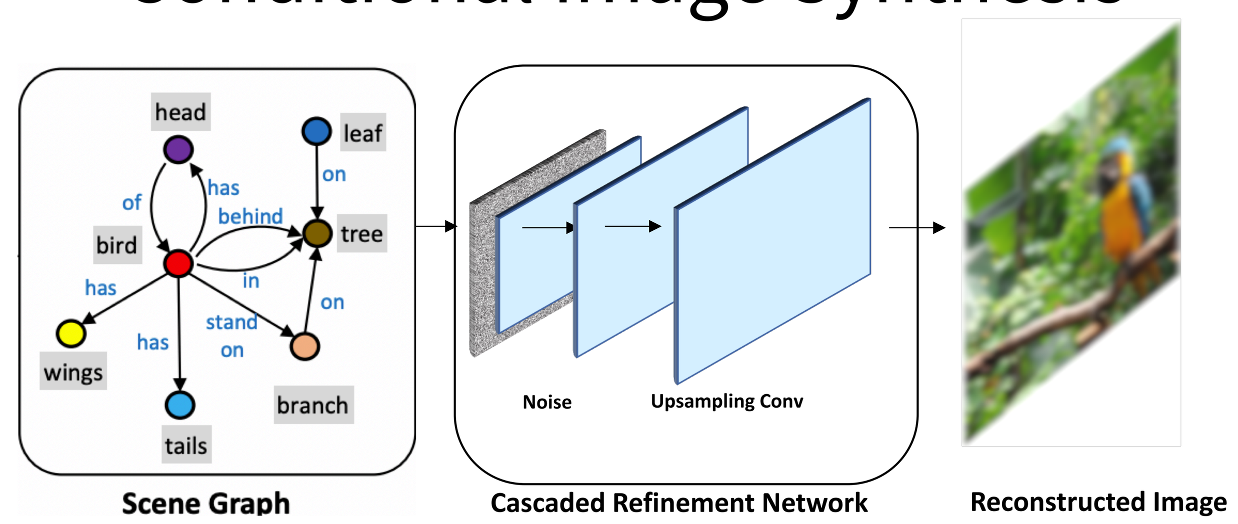
$$z_i^{(l+1)} = \sigma \left(z_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)} \right)$$



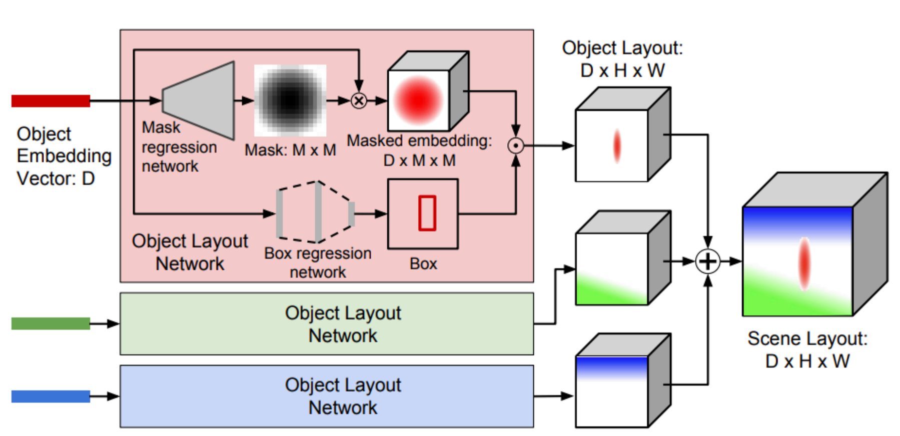
Objective Function

$$P(\mathcal{S}|\boldsymbol{I}) = \overbrace{P(\boldsymbol{V}|\boldsymbol{I})}^{\text{Object Region Proposal}} \underbrace{P(\boldsymbol{E}|\boldsymbol{V}, \boldsymbol{I})}_{\text{Relationship Proposal}} \overbrace{P(\boldsymbol{R}, \boldsymbol{O}|\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{I})}^{\text{Graph Labeling}}$$

Conditional Image Synthesis



Scene Layout



Procedure for Image Generation

Function: $\mathrm{Gen}_{\mathrm{o2i}}$
Input: Real image $\mathbf{I}$, objects (GT / predicted).
1: Object Layout Generation: $o_i^{\text{layout}} \leftarrow \{\mathbf{o}_i, r_i\}$
2: Scene Layout Generation: $S^{\text{layout}} = \sum_i o_i^{\text{layout}}$
3: Image Reconstruction: $\hat{\mathbf{I}} = G_i(z, S^{\text{layout}})$
4: Update image generator G_i parameters using (17).
5: Update image discriminator D_i parameters using (16).

Image-level Supervision

$$\mathcal{L}_{D_i} = \mathbb{E}_{I \sim p_{\text{real}}}[\log D_i(\mathbf{I})]$$

$$\mathcal{L}_{G_i} = \mathbb{E}_{\hat{I} \sim p_G}[\log(1 - D_i(\hat{\mathbf{I}})] + \lambda_p \mathcal{L}_{\text{pixel}}$$

$$||\mathbf{I} - \hat{\mathbf{I}}||_1$$

Experiments

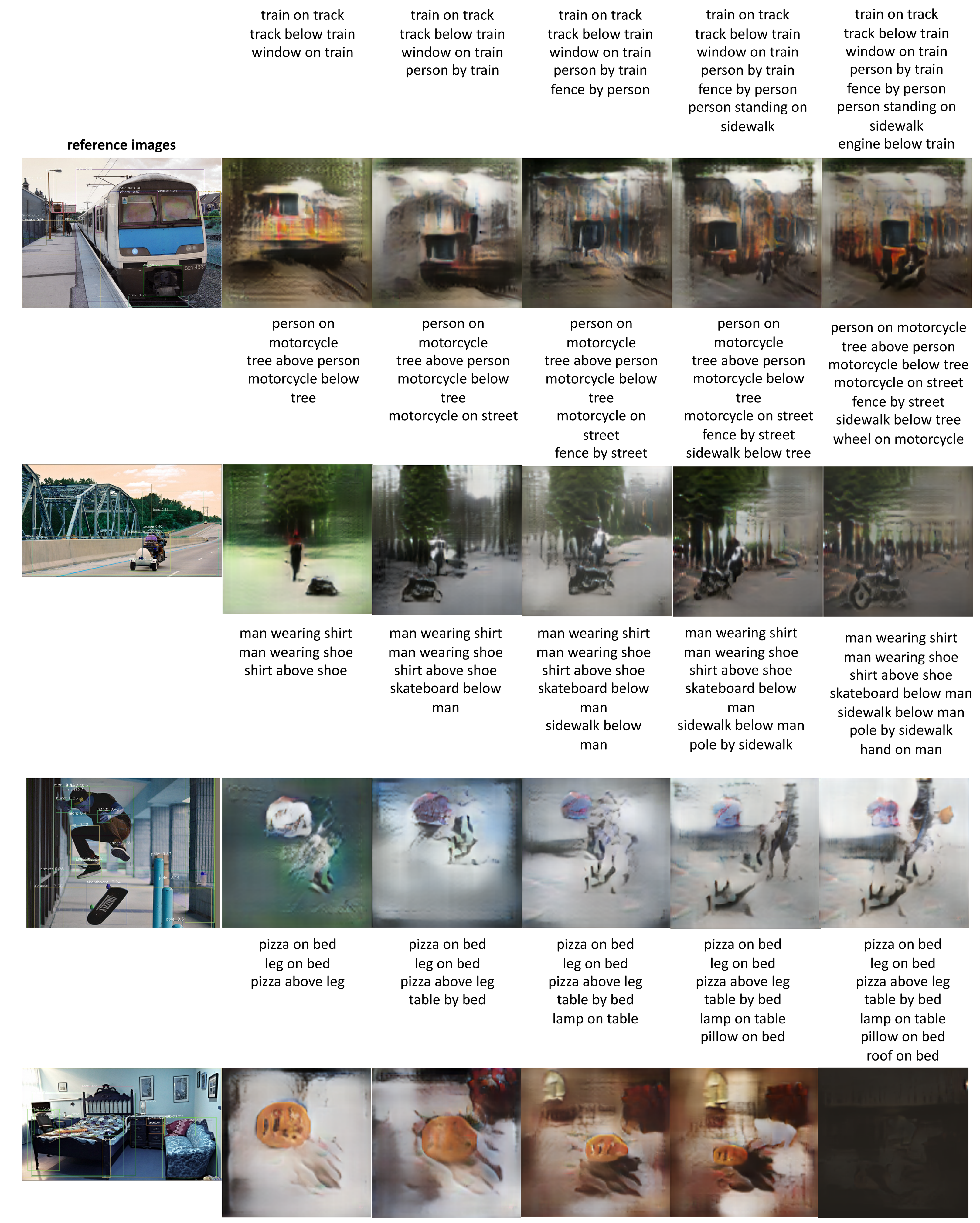


Figure: Qualitative results for conditional image synthesis on Visual Genome (VG). The first column depicts a reference image sampled from VG with the predicted bounding boxes. Each row begins with a simple scene graph and gradually adds more objects and relationships to gain complexity. Generated images demonstrate contextual knowledge and respect object relationships (e.g. pizza on bed).

	SGGen+		SGGen		PhrCls		PredCls	
Method	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
IMP	25.6	27.7	6.4	8.0	20.6	22.4	40.8	45.2
MSDN	25.8	28.2	7.0	9.1	27.6	29.9	53.2	57.9
NM-Freq	26.4	27.8	6.9	9.1	23.8	27.2	41.8	48.8
Graph R-CNN	**28.5**	**35.9**	**11.4**	**13.7**	**29.6**	**31.6**	**54.2**	**59.1**

Table: Comparisons on Visual Genome test set. We use Graph-RCNN for the scene graph generator outperforming the baseline of Iterative Message Passing (IMP). MSDN and NM-Freq refer to Multi-level Scene Description Network and Neural Motifs Frequency Prior, respectively.