Exploiting Relationship for Complex-scene Image Generation

Tianyu Hua[1], Hongdong Zheng[1], Yalong Bai[1], Wei Zhang[1*], Xiao-Ping Zhang[2], Tao Mei[1]
[1]JD AI Research
[2]Ryerson University
patrickhua.ty@gmail.com, {hongdongzheng, ylbai}@outlook.com, wzhang.cu@gmail.com
xzhang@ee.ryerson.ca, tmei@live.com

Abstract

The significant progress on Generative Adversarial Networks (GANs) has facilitated realistic *single-object* image generation based on language input. However, *complex-scene* generation (with various interactions among multiple objects) still suffers from messy layouts and object distortions, due to diverse configurations in layouts and appearances. Prior methods are mostly object-driven and ignore their inter-relations that play a significant role in complex-scene images. This work explores relationship-aware complex-scene image generation, where multiple objects are inter-related as a scene graph. With the help of relationships, we propose three major updates in the generation framework. First, reasonable spatial layouts are inferred by jointly considering the semantics and relationships among objects. Compared to standard location regression, we show relative scales and distances serve a more reliable target. Second, since the relations between objects significantly influence an object's appearance, we design a relation-guided generator to generate objects reflecting their relationships. Third, a novel scene graph discriminator is proposed to guarantee the consistency between the generated image and the input scene graph. Our method tends to synthesize plausible layouts and objects, respecting the interplay of multiple objects in an image. Experimental results on Visual Genome and HICO-DET datasets show that our proposed method significantly outperforms prior arts in terms of IS and FID metrics. Based on our user study and visual inspection, our method is more effective in generating logical layout and appearance for complex-scenes.

Introduction

In the past few years, text-to-image generation has drawn extensive research attention for its potential applications in art generation, computer-aided design, image manipulation, etc. However, such success is only restricted to simple image generation, which only contains a single object in a small domain, such as flowers, birds, and faces (Reed et al. 2016; Bao et al. 2017). Complex-scene generation, on the other hand, targets for synthesizing realistic scene images out of complex sentences depicting multiple objects as well as their interactions. Nevertheless, generating complex-scenes on demand is still far from mature based on recent studies (Johnson, Gupta, and Fei-Fei 2018; Xu et al. 2018; Li et al. 2019; Hinz, Heinrich, and Wermter 2019).

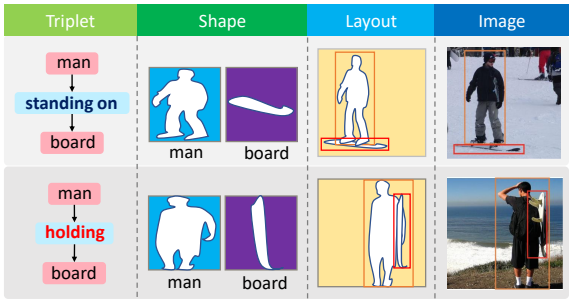

Figure 1: Relationship matters for complex-scene image generation. The same object pair (e.g., *man and board*) could show different object shapes, scene layouts and appearances under different relationships.

Scene graph, a structured language representation, captures objects and their relationships described in the sentence (Xu et al. 2017). Such representation is proven effective for image-text cross-modal tasks, such as structural image retrieval (Johnson et al. 2015; Schuster et al. 2015; Johnson, Gupta, and Fei-Fei 2018), image captioning (Yang et al. 2019; Li and Jiang 2019; Li et al. 2018) and visual question answering (Teney, Liu, and van Den Hengel 2017; Norcliffe-Brown, Vafeias, and Parisot 2018). In this work, we focus on complex-scene image generation from scene graphs. Although extensive works have been done in scene graph generation from image (Xu et al. 2017; Zellers et al. 2018; Li et al. 2017; Zhang et al. 2017a) (i.e. image→scene graph), reversely generating a complex-scene image from a scene graph remains challenging, due to the polymorphism nature of one-to-many mapping from a given scene graph to multiple reasonable images with different scene layouts.

A general pipeline for scene graph based image generation usually consists of two stages (Johnson, Gupta, and Fei-Fei 2018). The first one learns to synthesize a rough layout prediction from the scene-graph input. Usually, the object features are encoded with a graph module (Johnson, Gupta, and Fei-Fei 2018; Ashual and Wolf 2019), followed by a direct regression of bounding-box locations. At the second stage, a position-aware feature tensor, that combines object

*Corresponding author

features and layout generated in the first stage, is fed into an image decoder for generating the final image. For enhancing the object appearances in generated images, Ashual *et al.* separates appearance embedding from layout embedding.

However, previous works on complex-scene generation heavily suffer from two fundamental problems: messy layout, and object distortion. 1) *Messy layouts*. Image generation models are expected to figure out the reasonable layout from scene-graph inputs. However, there exist an infinite number of reasonable layouts for a given scene-graph. Directly fitting a specific layout introduces huge confusion, and limits the generalization ability. As a result, existing methods are still struggling with messy layouts in practice. 2) *Distortion in object appearance*. Due to the high diversity in object categories, layouts, and relationship dependencies, objects are often distorted during generation. For each object, the texture and local appearances should be inferred, respecting both its category and allocated spatial arrangement. Moreover, complex and various relations among different objects in the scene-graph can further increase the diversity of shape appearances. As shown in Fig. 1, even with the same object pairs, equipping different relationships could lead to totally different scene layouts and appearances.

Some works (Ashual and Wolf 2019) simplify the task by only taking a few simple spatial relationships among objects (such as "left", "right" or "above") but ignoring other complex relationships (such as verbs). Meanwhile, to reduce the complexity, some works only consider one specific stage of this task, such as layout generation from scene-graph (Jyothi et al. 2019), image generation from layout (Zhao et al. 2018; Sun and Wu 2019). All these works did not take account of the semantics and complex relationships among objects, which limits their application prospects.

In this work, we explore relationships to mitigate the above issues. We consider both simple spatial relationships and complex semantic relationships into consideration. We observed that, in different realistic images, relative scale and distance ratios between two interrelated objects from the same "subject-relation-object" triplet usually conform to a norm distribution with low variance, as in Fig. 2. Even though the "human" have various poses, and the skateboard can be oriented to different directions, the scale ratio between the two bounding boxes is naturally clustered with very low variance. Thus, we first introduce relative scale ratios and distance for measuring the rationality of layouts generated from the scene graph. It means that all *various reasonable layouts* relevance to one specific scene graph can be measured under a common standard and result in very similar results. After that, we proposed a *Pair-wise Spatial Constraint Module* for assisting layout generation. Our Spatial Constraint Module is influenced by object pairs and their corresponding relation jointly. Meanwhile, the objective of this novel module is to correct the layout by fitting the relative scale ratio and relative distance ratio between interrelated object pair beside the absolute position coordinates of each object. In this way, the spatial commonsense of complex scene with multiple objects can be modeled.

Moreover, for enhancing the influence of relation for object appearance generation, we proposed a *Relation-guided*

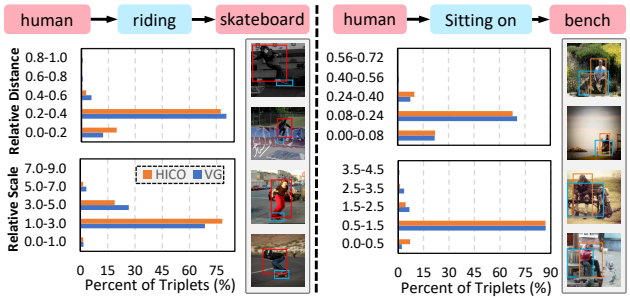

Figure 2: Distributions of relative scale and distance for "man riding skateboard" and "man sitting on bench".

Appearance Generator and a novel *Scene Graph Discriminator* for guiding image generation. Unlike the traditional discriminator that only judges whether the image is fake or not, our proposed new discriminator has two main functions. One is to determine whether the objects in the generated image are relevant to the objects described in the text scene graph or not, and the other is to discriminate the relation predictions among objects from the generated image are correspondence with the relationship described in the input scene graph. By feeding back these strong discriminant signals, our Scene Graph Discriminator guarantees the generated object patches align with not only single object fine-grained information but also the relation discrepancy among objects.

The main contributions can be summarized as follows:

- A novel pair-wise spatial constraint module with supervisions of relative scale and distance between objects for learning relationship-aware spatial perceptions.
- A relation-guided appearance generator module followed by a scene graph discriminator for generating reasonable object patches respecting object fine-grained information and relation discrepancy.
- A general framework for synthesizing scene layout and images from scene graphs. The experimental results on Visual Genome (Krishna et al. 2017) and human-objects interactions dataset HICO-DET (Chao et al. 2018) demonstrate the complex-scene images generated by our proposed method follow the common sense.

Related Work

Image Synthesis from Sentence is a conditional image generation task whose conditional signal is natural language. Textual descriptions are traditionally fed directly to a recurrent model for semantic information extractions. After that, a generative model will produce the results conditioned on this vectorized sentence representation. Most of these tasks specialize in single object image generation (Reed et al. 2016; Zhang et al. 2017b; Xu et al. 2018), whose layout is simple and the object usually centered with a large area in the image. However, generating realistic multi-object images conditioned on text descriptions is still a difficult task, since it addresses very complex sense layout generation and various object appearances mapping, and both of scene layout and object appearances are heavily influenced by the

spatial and semantic relationships cross objects. Furthermore, encoding all information, including multiple object categories and the interactions among them into one vector, usually leads to critical details lost. Meanwhile, directly decoding images from such an encoded vector hurts the interpretability of the model.

Scene Graph (Xu et al. 2017) is a directed graph that represents the structured relationships among objects in an explicit manner. Scene graphs have been widely used in many tasks such as image retrieval (Johnson et al. 2015), image captioning (Anderson et al. 2016), which serves as a medium that bridges the gap between language and vision.

Image Synthesis from Scene Graph (Johnson, Gupta, and Fei-Fei 2018) is a derivative of sentence based multiple-object image generation. Since the conditional signals are scene graphs, graphic models are usually applied for extracting essential information from the textual scene graph. After that, these extracted features are directly used for regressing the scene layouts and then treated as input to an image decoder for generating the final image (Ashual and Wolf 2019). Such a framework is applicable to generation image contains multiple objects with simple spatial interactions. However, it is still suffering from modeling the reasonable scene layouts and appearances following commonsense due to the implication of semantic relationships in the scene graph.

In this paper, we focus on image generation from the textual scene graph. Different from previous methods, we highlight the impact of relationships among objects for dealing with the messy layout and various object appearance.

Method

A scene graph is denoted as $\mathcal{G} = \{\mathcal{C}, \mathcal{R}, \mathcal{E}\}$, where $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ indicate the nodes in the graph, each $c_i \in \mathcal{C}$ denotes the category embedding of an object or instance. Note that we use words like "object" or "instance" in reference to a broad range of categories from "human", "tree" to "sky", "water" etc. The edges of the graph are extracted as a relationship embedding set $\mathcal{R}$. Two related objects c_j and c_k associate with one relationship $r_{jk} \in \mathcal{R}$, which leads to a triplet $\langle c_j, r_{jk}, c_k \rangle$ in the directed edge set $\mathcal{E}$.

Given a scene graph $\mathcal{G}$ and its corresponding image I, scene graph-based image generation model aims to generate an image $\widehat{I}$ according to $\mathcal{G}$ by minimizing $D(I, \widehat{I})$, where $D(I, \widehat{I})$ measures differences between I and $\widehat{I}$. A standard scene graph to image generation task can be formulated as two separate tasks: a scene graph to layout generation task which extracts object features with spatial constraints from scene graphs, and an image generation task, which generates images conditioned on the predicted object features and learned layout constraints, as shown in Fig. 3 (left).

In this paper, we extend the traditional framework by emphasizing the influence of relationship $\mathcal{R}$ for both object layouts and object appearances generation. As shown in Fig. 3 (right), three novel modules are proposed:

- **Pair-wise Spatial Constraint Module**: a module for constraining layout generation according to the semantic information extracted from $\mathcal{E}$.
- **Relation-guided Appearance Generator**: for each object c_i, we introduce the semantic information of its connected relationships $\{r_j | \langle c_i, r_j, * \rangle \in \mathcal{E}\}$ to influence the shape and appearance of the generated image of c_i.
- **Scene Graph Discriminator** (D_{sg}): a novel discriminator for strengthening the generated image $\widehat{I}$ to be relevant to the appearances of object $\mathcal{C}$, and the relationships $\mathcal{R}$ in the edge set $\mathcal{E}$.

Layout Generator

Layout generator aims to predict bounding boxes $b_i = (x_i, y_i, w_i, h_i)$ for each object o_i in given scene graph $\mathcal{G}$, where x_i, y_i, w_i, h_i specifies normalized coordinates of the center, width and height in ground-truth image I.

In previous work, the object representations are usually extracted from scene graph inputs, and then be passed to a box regression network to get the bounding box predictions $\widehat{b}_i = (\widehat{x}_i, \widehat{y}_i, \widehat{w}_i, \widehat{h}_i)$. The box regression network is trained by maximizing the objective:

$$\mathcal{L}_{box} = -\sum_{i=1}^{n} \parallel b_i - \widehat{b}_i \parallel_2, \tag{1}$$

which penalize the L_2 difference between ground-truth and predicted boxes. n indicates the amount of objects.

Since there are various reasonable layouts existing, as previously stated, a scene graph to layout task requires addressing challenging one-to-many mapping. Directly regressing layout to offsets of one specific bounding box would hurt the generalization ability of the layout generator, and make the layout generator to be difficult to convergence. In order to generate reasonable layouts, we relax the constraint of bounding box offsets regression and proposed a novel spatial constraint module for ensuring the rationality of layout.

Our **Pair-wise Spatial Constraint Module** introduces two novel metrics for measuring the realistic of layouts.
1. Relative Scale Invariance. The scale of an object is represented by the diagonal length of its bounding box. For any given $\langle c_j, r_{jk}, c_k \rangle$ triplet, the ratio between the scale of the subject and the scale of the object in different images are often roughly the same, as shown in Fig 4 (Left). We formulate the relative scale between the layout b_j and b_k as

$$s_{jk} = \sqrt{w_j^2 + h_j^2} \Big/ \sqrt{w_k^2 + h_k^2}. \tag{2}$$

2. Relative Distance Invariance. Similar to relative scale, relative distance target at calculating the distance between two objects in triplet normalized by the scales of two objects. The relative distance of related object pair c_j and c_k in realistic images is also naturally clustered to one specific value, and the distributions of relative distance for different triplets are usually with low variance, as shown in Fig 4 (Right). Normally, horizontal flips of images rarely alter spatial relationship distributions, we relax this constraint by using the absolute value of the horizontal coordinate difference. Most importantly, we normalize distance by the summed scales of object pairs so that the zooming effect of object depth can

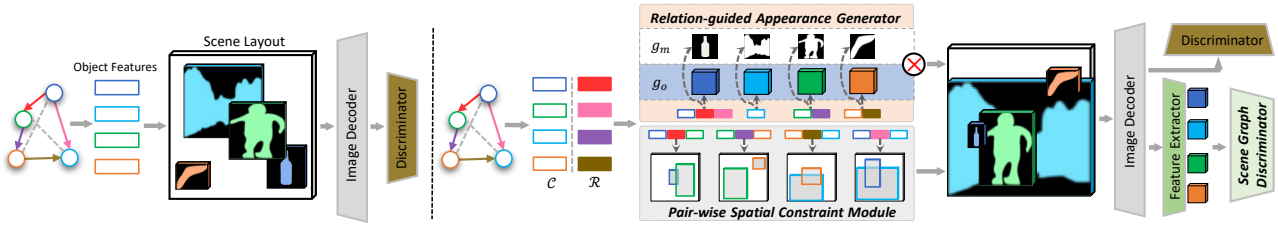

Figure 3: Illustrations of standard (left) and our (right) framework for scene graph to image generation. Left: Directly generating layout and image based on object features extracted from scene graph. Right: Our proposed framework with object pair-wise spatial constraints and appearance supervision respecting relationships among objects.

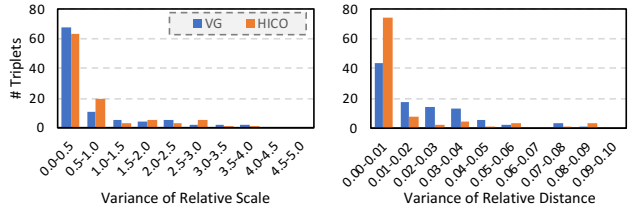

Figure 4: Distributions of relative scale and distance variance among top-100 triplets in VG and HICO-DET datasets. Low diversity of relative scale and distance is observed, following the property of commonsense knowledge.

be canceled out. Therefore, the relative distance between the layout b_j and b_k can be formulated as

$$\vec{d_{jk}} = [|x_j - x_k|, y_j - y_k]^T \Big/ \left(\sqrt{w_j^2 + h_j^2} + \sqrt{w_k^2 + h_k^2}\right). \tag{3}$$

We have keenly observed that relationship in a semantic form comes with it an inherent spatial constraint that has not been fully explored by others. For example, the relationship "holding" implies that the object should be within arm's reach of the subject instead of miles away. The relationship "walking" indicates the relative vertical arrangement between subject and object heavily, whether it's "man walking-on street" or "dog walking-on floor". It means the relative scale and relative distance between two objects heavily depend on the relationship or interaction between these two objects. Therefore, we devise a training scheme that explicitly leverages this constraint.

In this work, the scene graph $\mathcal{G}$ is first converted to object feature vectors $\mathcal{C}$ and relation embeddings $\mathcal{R}$, and then fed into a Graph Convolutional Network (GCN). The GCN outputs updated object level feature vector $o_i = T(c_i, \mathcal{C}_i, \mathcal{R}_i)$ aggregated with relation information, where T is the graph convolutional operation, $\mathcal{C}_i$ is the set of object embeddings relevant to c_i, $\mathcal{R}_i$ is the set of embeddings for relations among c_i and $\mathcal{C}_i$. It means the output vector o_i for an object c_i should depend on representations of relationships and all objects connected via graph edged jointly. After that, we apply the updated object representations for generating the layout for object c_i by $\widehat{b}_i = B(o_i)$, where B is an bounding box offset regression network. We construct a scale-distance objective for our proposed spatial constraint module to assist the training progress of B:

$$\mathcal{L}_{scm} = - \sum_{\substack{0<j,k<n \\ \langle c_j, r_{jk}, c_k\rangle \in \mathcal{E}}} \| s_{jk} - \widehat{s}_{jk} \|_2 + \| \vec{d}_{jk} - \vec{\widehat{d}}_{jk} \|_2, \tag{4}$$

where $\widehat{s}_{jk}$ and $\vec{\widehat{d}}_{jk}$ is the relative scale and relative distance between generated layouts for related object pair c_j and c_k respectively. $\mathcal{L}_{scm}$ is only computed on the connected object pairs in scene graph, since the relative scale and distance of two objects depend on the relationship between them, as we shown in Fig. 1.

With the supervision of relative scale and distance, the box regression network learns to arrange object boxes properly for reasonable layout generation.

Image Generator

Starting from the original object representations $\mathcal{C} \in \mathbb{R}^{n \times d_1}$ and initial relation embeddings $\mathcal{R} \in \mathbb{R}^{m \times d_2}$, we can compute a combined "object-relation" vector v_i for each object c_i in scene graph:

$$v_i = \left(c_i +\!\!+ \frac{1}{|\mathcal{E}_i|} \sum_{\langle c_i, r_j, * \rangle \in \mathcal{E}_i} r_j\right) + z_i, \tag{5}$$

where $+\!\!+$ indicates a vector concatenation operation, $\mathcal{E}_i \in \mathcal{E}$ is the collection of all triplet relevant to object c_i, z_i is a $d_1 + d_2$ dimensional noise vector randomly sampled from a Gaussian distribution, which aims to generate non-deterministic object features. The object and averaged relation embeddings are eventually be concatenated as inputs of our **Relation-guided Appearance Generator**, which consists of an object mask predictor g_m, an object appearnce feature predictor g_o and a full image generator.

The combined vector $\{v_i\}_{i=1}^n$ will be sent simultaneously to g_m and g_o, both of which are four-layer conv nets normalized with spectral normalization techniques (Miyato et al. 2018). Through an STN block (Jaderberg et al. 2015), the two outputs for different objects will first be filled into their respective bounding box layouts. Then we obtain a set of object shape tensor and appearance tensor. By multiplying these two tensor, we can generate the final *relation-guided* appearance feature tensor for all objects in scene graph as

$$a(\mathcal{G}) = \{S(\widehat{b}_i, g_m(v_i)) \circ S(\widehat{b}_i, g_o(v_i))\}_{i=1}^n, \tag{6}$$

where S is the STN block.

After that, our full image generator generate the image conditioned on all object appearance feature tensors $a(\mathcal{G})$ and an additional noise vector z_I. In detail, our image generator utilizes the ResNet architecture (He et al. 2016) consists of six ResBlocks as backbone. Consider generating a 256×256 image for scene graph, a randomly generated global latent vector z_I is a vector sampled from normal distribution. The vector is then mapped and reshaped to a 1024×4×4 (channels, width, and height) tensor through a fully-connected layer. Then, the tensor will be sent to the first ResBlock. Each of the six ResBlocks will upsample it's inputs bilinearly with a ratio of two. In the meantime, the channel number drops by a factor of 2 except for the third block. Block by block, we fuse object appearance tensor $\{f_i = G_{obj}(v_i)\}_{i=1}^{n}$ with the outputs of each ResBlock (global appearance tensor) using the ISLA-Norm method proposed by (Sun and Wu 2019). The final generated image $\widehat{I} = \widehat{I}_t$ comes from the outputs of the last ResBlock,

$$\begin{aligned} \widehat{I}_t &= R_t(\widehat{I}_{t-1}, a(\mathcal{G})) \\ \widehat{I}_1 &= R_1(z_I, a(\mathcal{G})) \end{aligned} \tag{7}$$

where R indicate a ResBlock equipped with ISLA-Norm module, t is the amount of ResBlocks in our image generator, $\widehat{I}_i$ is output of the i-th ResBlock.

Our image generator takes the object appearance features with relational information and global random noise as condition, adapts scene composition, and finally generates the realistic image $\widehat{I}$ for scene graph S.

Image Discriminator

Similar to the image generator, we adapt ResNet with downsampling blocks for image discriminator. The ResNet backbone consists of a different number of downsampling ResBlocks with respect to the input image sizes. The image downsampled by ResBlocks goes through a linear layer, and the outputs of the linear layer are further summed channel-wise to form a scalar output as the global discriminator score D_{img} to measure whether the input image is real or not, which is similar to traditional GAN based methods.

Since different relationships result in diverse appearance in the same object, we argue that the learned object feature representation reflects not just class-related object styles but also the relationship-aware appearances. Thus, we proposed a novel **Scene Graph Discriminator** D_{sg} to measure whether the scene graph extracted from the generated image is associated with the given textual scene graph or not. In detail, we first extract object-level feature patches $\{p_i\}_{i=1}^{n}$ rerouted from the second layer of ResNet backbone, then resize these feature patches to the same size by an RoI align layer (He et al. 2017). Then we introduce an object classifier F_{obj}, which attempts to classify the feature patches into categories. By pairing object feature tensors according to the edges of the scene graph, we send the paired object feature p_j and p_k to the relationship classifier F_{rel}, which aims to predict the type of relationship of given object feature pair. Our proposed D_{sg} aims to encourage the image generator to be aware of the object categories and relationships exists in the scene graph:

$$D_{sg}(I) = \frac{1}{n}\sum_{i=0}^{n} F_{obj}(c_i|p_i) + \frac{1}{|\mathcal{E}|}\sum_{\substack{0<j,k<n \\ \langle c_j, r_{jk}, c_k\rangle \in \mathcal{E}}} F_{rel}(r_{jk}|p_j, p_k). \tag{8}$$

Moreover, we introduce an object discriminator D_{obj} to measure whether each object in image appears realistic based on $\{p_i\}_{i=1}^{n}$.

The overall objective function for training layout generator, image generator and discriminators is defined as:

$$\mathcal{L} = \lambda_1\mathcal{L}_{box} + \lambda_2\mathcal{L}_{scm} + \lambda_3\mathcal{L}_{obj} + \lambda_4\mathcal{L}_{sg} + \lambda_5\mathcal{L}_{img}, \tag{9}$$

where $\mathcal{L}_{img}$ is image adversarial loss from D_{img}, $\mathcal{L}_{obj}$ is object adversarial loss from D_{obj}, $\mathcal{L}_{sg}$ is scene graph relevant loss from D_{sg}. In our experiments, we set the loss weight parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4 = 1$, $\lambda_5 = 0.1$.

Experimental Results

We evaluate our proposed method for generating images at three different resolutions 64×64, 128×128, and 256×256 in below two datasets:

Visual Genome (Krishna et al. 2017) was constructed with cognitive tasks that provide crowd-sourced dense annotations of both scene graphs and images. Following the settings of (Johnson, Gupta, and Fei-Fei 2018), we experiment on Visual Genome version 1.4. We keep 178 objects and 45 relations in the dataset by removing images with object and relationship categories less than 2000 and 500, respectively.

HICO-DET (Chao et al. 2018) was built for modeling humans-object interactions. Compared with Visual Genome, the scene graphs provided in the dataset are human-centered. We keep images that have object categories higher than 1000 and discard images with interaction types that repeat less than 250 times, leaving 19 objects and 22 relationship types in total. Images with Object size below 32×32 and images with objects less than 2 or more than 10 are ignored. Finally, we get 15963 train images and 4034 test images.

The COCO dataset (Caesar, Uijlings, and Ferrari 2018) is not used in this paper because the relationship types in COCO are too simple, consisting mainly of naive spatial arrangement relations. We trained models using Adam with an initial *lr*=10^{-4} and batch size of 32 for 200 epochs.

Several previous works target at multi-object image generation. Most of these works are about image synthesis from the ground-truth layout or pixel-level instance segmentation annotation (Sun and Wu 2019; Hong et al. 2018; Li et al. 2019). The work of Ashual and Wolf aims to generated images from an input scene graph. However, the scene graph used in their work is simplified by only equipping six spatial relationships (right-of, left-of, above, below, surrounding and inside). Moreover, location attributes are assisted by additional information for each node in scene graph. Moreover, location attributes are assisted by additional information for each node in scene graph. Luo et al. only focus on spatial relationships instead of semantic relationships. Besides, objects used in their paper are mostly rigid bodies. Our paper learns from not just spatial relationships, but semantic relationships (e.g. looking at) as well. We use datasets

Resolution	Method	Visual Genome		HICO-DET	
		IS	FID	IS	FID
64×64	*I*	13.9 ± 1	0.0	9.8 ± 0.5	0.0
	sg2im†	6.3 ± 0.2	47.6	4.4 ± 0.1	99.9
	LostGAN†	6.9±0.1	38.7	4.5±0.3	86.4
	Ours†	**7.5 ± 0.4**	**29.0**	**5.5 ± 0.1**	**41.7**
	sg2im	5.5 ± 0.1	47.5	4.4 ± 0.1	94.3
	PasteGAN	6.9±0.2	58.5	-	-
	Ours	**7.0 ± 0.2**	**37.7**	**5.3 ± 0.7**	**47.4**
128×128	*I*	22.5 ± 1.9	0.0	13.7 ± 0.7	0.0
	sg2im†	6.3 ± 0.2	83.9	4.6 ± 0.1	83.7
	LostGAN†	7.4±0.3	53.4	4.8±0.1	79.9
	Ours†	**9.4 ± 0.4**	**41.0**	**6.5 ± 0.1**	**60.6**
	sg2im	6.2 ± 0.2	83.8	4.6 ± 0.1	123.0
	Ours	**9.2 ± 0.8**	**53.0**	**5.0 ± 0.3**	**61.4**
256×256	*I*	30.1 ± 2.3	0.0	16.3 ± 0.5	0.0
	Ours†	12.6 ± 0.5	68.3	7.5 ± 0.1	78.3
	Ours	10.8 ± 0.9	85.7	6.9 ± 0.3	80.5

Table 1: The comparison of IS and FID among different methods. On each dataset, the test set samples are randomly split into 5 groups. The mean and standard deviation across splits are reported in the above table. † indicates that the images are generated based on the ground-truth layouts instead of the generated layouts. *I* denotes the real image.

that involve a large number of non-rigid objects that have various shapes and appearances and be sensitive to their relevant semantic relationships, which drastically increase the difficulty of our task. PasteGAN (Yikang et al. 2019) applies both of the scene graph and ground-truth image crops as the inputs for complex-scence image generation. According to our best knowledge, sg2im (Johnson, Gupta, and Fei-Fei 2018) is the only related work about complex-scene image generation images from scene graphs that contain semantic and complex relationships among objects.

Compared Methods In this paper, we compare our proposed method with **sg2im** and **PasteGAN** for complex-scene image generation. Moreover, to demonstrate the effectiveness of our relation-guided appearance generator and scene graph discriminator, we also compare our method with **LostGAN** (Sun and Wu 2019) which is designed for generating images by given ground-truth layout.

Quantitative Results

We adopt two metrics to evaluate the generated images.

Inception Score (*IS*) (Salimans et al. 2016) measures the diversity of generated images and their quality. A pre-trained InceptionV3 model is adapted to predict the class probabilities for given image. Larger inception scores are better.

Fréchet Inception Distance (*FID*) (Heusel et al. 2017)[1] measures the Fréchet distance between the multivariate Gaussian distribution of real images and generated ones. Lower *FID* scores are better.

These two metrics are widely used evaluation metrics for generative models. *IS* aims to evaluate the reality of a single object, while *FID* is more suitable to reflect the quality of the generated image contains multiple objects.

[1]https://github.com/mseitzer/pytorch-fid

Method	IS	FID
Ours	9.2 ± 0.8	53.0
w/o Pair-wise Spatial Constrain Module ($\mathcal{L}_{scm}$)	8.6 ± 1.2	59.8
w/o Relation-guided Appearance Generator	8.7 ± 0.9	57.4
w/o Scene Graph Discriminator ($\mathcal{L}_{sg}$)	7.4 ± 0.2	73.3

Table 2: Ablation studies conducted on our proposed method. The experimental results are reported on 128×128 resolution image generation task in Visual Genome.

User study	sg2im	Same	Ours
Image is more realistic	9%	26%	65%
Image has reasonable object arrangement	12%	27%	61%
Image reflects relationships in scene graph	9%	19%	72%
Layout is more reasonable	11%	30%	59%

Table 3: We performed a user study to compare the quality of generated layouts and images of our method against sg2im.

Table 1 summarizes the performances on the two aforementioned datasets in terms of Inception Score and *FID* score. Our model outperforms sg2im on both VG and HICO-DET datasets. Moreover, even without the external information like image crops, our method can still achieve better results reported in PasteGAN. In addition, we conduct experiments of GT Layout versions using ground-truth bounding boxes during both training and testing. This method gives an upper bound to the model's performance in the case of perfect layout construction. As shown in Table 1, our method has more potential than sg2im and LostGAN.

We also conducted ablation studies in Table 2. The relative importance of the *Pair-wise Spatial Constraint Module*, and *Scene Graph Discriminator* are measured by removing $\mathcal{L}_{scm}$ and $\mathcal{L}_{sg}$ from the overall objective function respectively. The ablation studies of *Relation-guided Appearance Generator* is measured by erasing relation embeddings during computing object shape and texture features. It can be found that the layout constraint module predicts reasonable spatial layout arrangements that improve the generated image qualities. The relation-guided generator introduces more reasonable appearance information. The scene graph discriminator can advance the correspondence between textual scene graph inputs and generated images.

Qualitative Results

Fig. 5 shows the capability of our method compared with that of sg2im on VG and HICO-DET. In the 1st column, sg2im predicts the human layout is above the motorcycle, which is not a normal position arrangement. Similarly, in the 5th column, sg2im predicts that "pant" is not vertically in line with the "shirt". Moreover, in the last column sg2im predicts that the scale of "sky" is too small. It leads to the chaotic color fill in the generated image. Similarly, in the 7th column sg2im predicts the scale of "snow" is much bigger than "mountain", which conflicts with the triplet "snow on mountain". These displacements occur when the training process is not enhanced with relative distance and scales.

Fig. 6 shows the generated images conditioned on ground

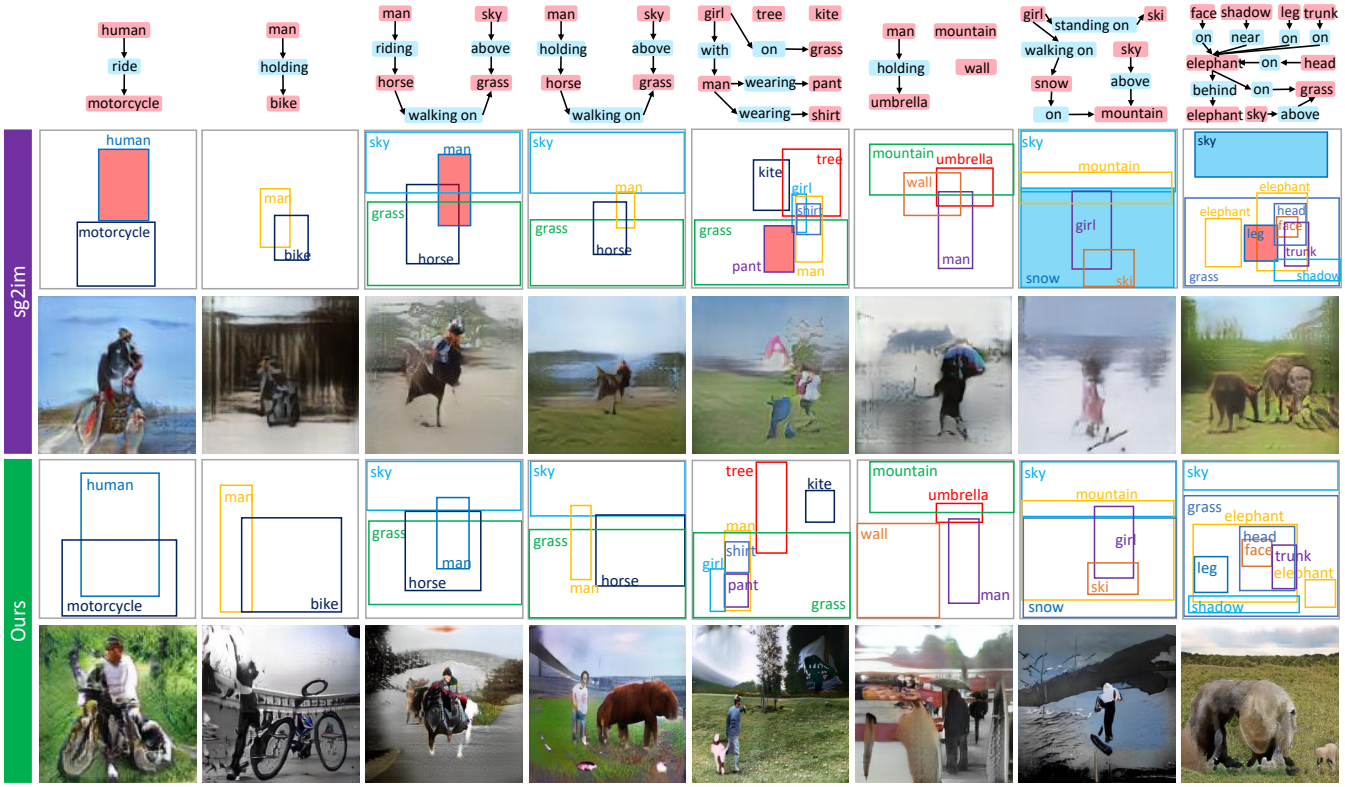

Figure 5: Examples of layouts and images generated from scene graphs in Visual Genome and HICO-DET for our method and sg2im. In the layout examples, we use red color patches to denote bounding boxes that fail to reflect the distance between object pairs. The bounding boxes with blue background have an unnatural scale configuration. Best viewed in color version.

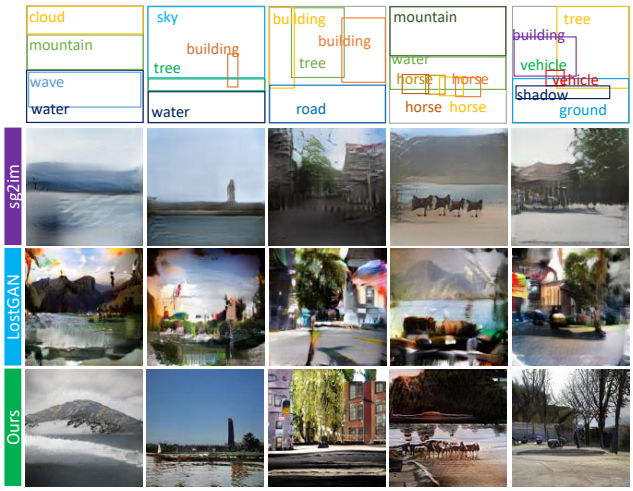

Figure 6: Generated samples from ground truth layouts on Visual Genome by sg2im, LostGAN and our method.

truth layouts. We compared our model grounded on the same position layouts compared with sg2im and LostGAN. It can be found that our method is more likely to generate realistic images from rich layouts and with natural objects.

User Study We also conduct a user study to measure human preference between images generated by out method and sg2im in Table 3. We choose the 128×128 resolution models for both cases. Our user study involves 40 students having a background in computer science. We generate 500 test cases from the VG test set for user study. A majority of users preferred the generated layouts and images from our method in 65% of image pairs.

Conclusion

The relationship between objects significantly rectifies the localization of objects and even their appearances. Prior literature mainly focuses on fitting the single object appearance. Semantic interactions among objects were overlooked, which may result in inconsistent and chaotic results. In this paper, we proposed a new framework to generate complex-scene images by exploring the importance of relationships among multiple objects in complex-scene image generation. Quantitative results, qualitative results, and user studies show our method's ability for reasonable layout generation and object interactions alignment.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600, and 2020AAA0103800.

Ethical Impact

Images are tiny visual samples of the grand physical world. The elementary particles that comprise our world first evolve and cluster into objects. Then appearance of objects are gradually shaped by the interactions between their counterparts. To model the natural clustering and interaction, we construct a generative model that strictly and structurally mimic the graph-like natural arrangement of our world. Our model builds a projection from symbolic graph space to the pixel space and we demonstrate how a small alternation in the object relationship can greatly affect the appearance of surrounding objects. In the future, we should research more on the neural design that can easily fit the data structure of pixels and encode more in the model visually commonsensical (relational) patterns. Overall, this paper has a positive impact on both industry and academia and enhance people's understanding of visual thinking.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 382–398. Springer.

Ashual, O.; and Wolf, L. 2019. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 4561–4569.

Bao, J.; Chen, D.; Wen, F.; Li, H.; and Hua, G. 2017. CVAE-GAN: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, 2745–2754.

Caesar, H.; Uijlings, J. R. R.; and Ferrari, V. 2018. COCO-Stuff: Thing and Stuff Classes in Context. *computer vision and pattern recognition* 1209–1218.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, 381–389. IEEE.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 6626–6637.

Hinz, T.; Heinrich, S.; and Wermter, S. 2019. Generating multiple objects at spatially distinct locations. *arXiv preprint arXiv:1901.00686* .

Hong, S.; Yang, D.; Choi, J.; and Lee, H. 2018. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. *computer vision and pattern recognition* 7986–7994.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; and Kavukcuoglu, K. 2015. Spatial Transformer Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, 2017–2025. Cambridge, MA, USA: MIT Press.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1219–1228.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Jyothi, A. A.; Durand, T.; He, J.; Sigal, L.; and Mori, G. 2019. Layoutvae: Stochastic scene layout generation from a label set. In *Proceedings of the IEEE International Conference on Computer Vision*, 9895–9904.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123(1): 32–73.

Li, W.; Zhang, P.; Zhang, L.; Huang, Q.; He, X.; Lyu, S.; and Gao, J. 2019. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12174–12182.

Li, X.; and Jiang, S. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia* 21(8): 2117–2130.

Li, Y.; Ouyang, W.; Zhou, B.; Wang, K.; and Wang, X. 2017. Scene Graph Generation from Objects, Phrases and Region Captions. In *IEEE International Conference on Computer Vision*, 1270–1279.

Li, Y.; Yao, T.; Pan, Y.; Chao, H.; and Mei, T. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7492–7500.

Luo, A.; Zhang, Z.; Wu, J.; and Tenenbaum, J. B. 2020. End-to-End Optimization of Scene Layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3754–3763.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* .

Norcliffe-Brown, W.; Vafeias, S.; and Parisot, S. 2018. Learning conditioned graph structures for interpretable visual question answering. In *Advances in Neural Information Processing Systems*, 8334–8343.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396* .

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*, 2234–2242.

Schuster, S.; Krishna, R.; Chang, A.; Fei-Fei, L.; and Manning, C. D. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, 70–80.

Sun, W.; and Wu, T. 2019. Image synthesis from reconfigurable layout and style. In *Proceedings of the IEEE International Conference on Computer Vision*, 10531–10540.

Teney, D.; Liu, L.; and van Den Hengel, A. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5410–5419.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10685–10694.

Yikang, L.; Ma, T.; Bai, Y.; Duan, N.; Wei, S.; and Wang, X. 2019. Pastegan: A semi-parametric method to generate image from scene graph. In *Advances in Neural Information Processing Systems*, 3948–3958.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5831–5840.

Zhang, H.; Kyaw, Z.; Chang, S.; and Chua, T. 2017a. Visual Translation Embedding Network for Visual Relation Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3107–3115.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017b. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915.

Zhao, B.; Meng, L.; Yin, W.; and Sigal, L. 2018. Image Generation from Layout. *arXiv: Computer Vision and Pattern Recognition* .