

Text2Scene: Generating Compositional Scenes from Textual Descriptions

Fuwen Tan¹ Song Feng² Vicente Ordóñez¹

¹University of Virginia, ²IBM Thomas J. Watson Research Center.

fuwen.tan@virginia.edu, sfeng@us.ibm.com, vicente@virginia.edu

Abstract

In this paper, we propose Text2Scene, a model that generates various forms of compositional scene representations from natural language descriptions. Unlike recent works, our method does NOT use Generative Adversarial Networks (GANs). Text2Scene instead learns to sequentially generate objects and their attributes (location, size, appearance, etc) at every time step by attending to different parts of the input text and the current status of the generated scene. We show that under minor modifications, the proposed framework can handle the generation of different forms of scene representations, including cartoon-like scenes, object layouts corresponding to real images, and synthetic images. Our method is not only competitive when compared with state-of-the-art GAN-based methods using automatic metrics and superior based on human judgments but also has the advantage of producing interpretable results.

1. Introduction

Generating images from textual descriptions has recently become an active research topic [29, 40, 41, 14, 36, 12]. This interest has been partially fueled by the adoption of Generative Adversarial Networks (GANs) [8] which have demonstrated impressive results on a number of image synthesis tasks. Synthesizing images from text requires a level of language and visual understanding which could lead to applications in image retrieval through natural language queries, representation learning for text, and automated computer graphics and image editing applications.

In this work, we introduce Text2Scene, a model to interpret visually descriptive language in order to generate compositional scene representations. We specifically focus on generating a scene representation consisting of a list of objects, along with their attributes (e.g. location, size, aspect ratio, pose, appearance). We adapt and train models to generate three types of scenes as shown in Figure 1, (1) Cartoon-like scenes as depicted in the Abstract Scenes dataset [44], (2) Object layouts corresponding to image scenes from the COCO dataset [21], and (3) Synthetic

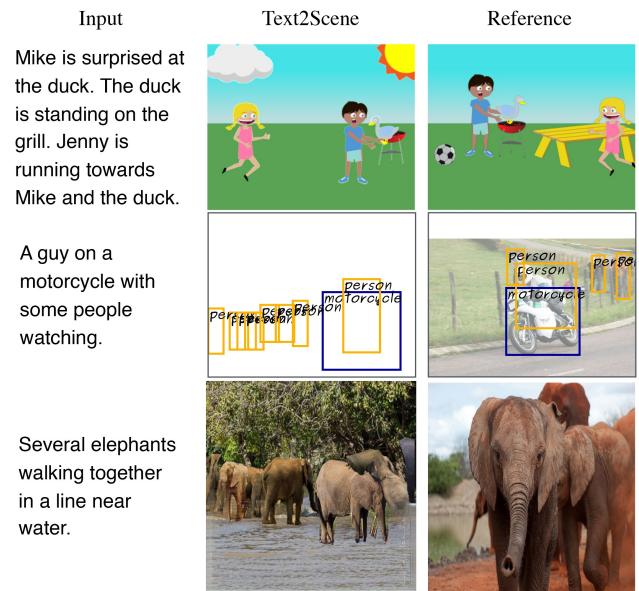


Figure 1. Sample inputs (left) and outputs of our Text2Scene model (middle), along with *ground truth* reference scenes (right) for the generation of abstract scenes (top), object layouts (middle), and synthetic image composites (bottom).

scenes corresponding to images in the COCO dataset [21]. We propose a unified framework to handle these three seemingly different tasks with unique challenges. Our method, unlike recent approaches, does not rely on Generative Adversarial Networks (GANs) [8]. Instead, we produce an interpretable model that iteratively generates a scene by predicting and adding new objects at each time step. Our method is superior to the best result reported in Abstract Scenes [44], and provides near state-of-the-art performance on COCO [21] under automatic evaluation metrics, and state-of-the-art results when evaluated by humans.

Generating rich textual representations for scene generation is a challenging task. For instance, input textual descriptions could hint only indirectly at the presence of attributes – e.g. in the first example in Fig. 1 the input text “Mike is surprised” should change the facial attribute on the generated object “Mike”. Textual descriptions often have complex information about relative spatial configura-

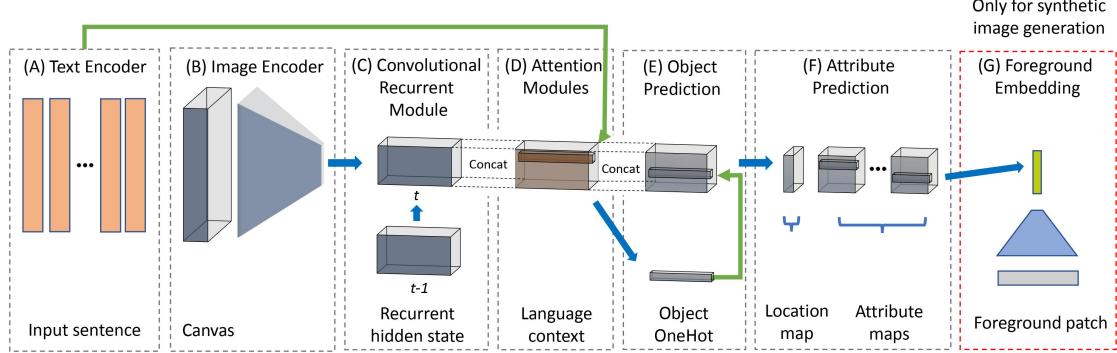


Figure 2. Overview of Text2Scene. Our general framework consists of (A) a Text Encoder that produces a sequential representation of the input, (B) an Image Encoder that encodes the current state of the generated scene, (C) a Convolutional Recurrent Module that tracks, for each spatial location, the history of what have been generated so far, (D-F) two attention-based predictors that sequentially focus on different parts of the input text, first to decide what object to place, then to decide what attributes to be assigned to the object, and (G) an optional foreground embedding step that learns an appearance vector for patch retrieval in the synthetic image generation task.

tions – e.g. in the first example in Fig. 1 the input text “Jenny is running towards Mike and the duck” makes the orientation of “Jenny” dependent on the positions of both “Mike”, and “duck”. In the last example in Fig. 1 the text “elephants walking together in a line” also implies certain overall spatial configuration of the objects in the scene. We model this text-to-scene task using a sequence-to-sequence approach where objects are placed sequentially on an initially empty canvas (see an overview in Fig 2). Generally, Text2Scene, consists of a text encoder (Fig 2 (A)) that maps the input sentence to a set of latent representations, an image encoder (Fig 2 (B)) which encodes the current generated canvas, a convolutional recurrent module (Fig 2 (C)), which passes the current state to the next step, attention modules (Fig 2 (D)) which focus on different parts of the input text, an object decoder (Fig 2 (E)) that predicts the next object conditioned on the current scene state and attended input text, and an attribute decoder (Fig 2 (F)) that assigns attributes to the predicted object. To the best of our knowledge, Text2Scene is the first model demonstrating its capacities on both abstract and real images, thus opening the possibility for future work on transfer learning across domains.

Our main contributions can be summarized as follows:

- We propose Text2Scene, a framework to generate compositional scene representations from natural language descriptions.
- We show that Text2Scene can be used to generate, under minor modifications, different forms of scene representations, including cartoon-like scenes, semantic layouts corresponding to real images, and synthetic image composites.
- We conduct extensive experiments on the tasks of abstract image generation for the Abstract Scenes dataset [44], semantic layout and synthetic image generations for the COCO dataset [21].

2. Related Work

Most research on visually descriptive language has focused on generating captions from images [5, 23, 18, 15, 34, 35, 24, 2]. Recently, there is work in the opposite direction of text-to-image synthesis [28, 29, 40, 14, 41, 36, 12]. Most of the recent approaches have leveraged conditional Generative Adversarial Networks (GANs). While these works have managed to generate results of increasing quality, there are major challenges when attempting to synthesize images for complex scenes with multiple interacting objects without explicitly defining such interactions [37]. Inspired by the principle of compositionality [42], our model does not use GANs but produces a scene by sequentially generating objects (e.g. in the forms of clip-arts, bounding boxes, or segmented object patches) containing the semantic elements that compose the scene.

Our work is also related to prior research on using abstract scenes to mirror and analyze complex situations in the real world [43, 44, 7, 33]. In [44], a graphical model was introduced to generate an abstract scene from textual descriptions. Unlike this previous work, our method does not use a semantic parser but is trained end-to-end from input sentences. Our work is also related to recent research on generating images from pixel-wise semantic labels [13, 4, 27], especially [27] which proposed a retrieval-based semi-parametric method for image synthesis given the spatial semantic map. Our synthetic image generation model optionally uses the cascaded refinement module in [27] as a post-processing step. Unlike these works, our method is not given the spatial layout of the objects in the scene but learns to predict a layout indirectly from text.

Most closely related to our approach are [14, 9, 12], and [16], as these works also attempt to predict explicit 2D layout representations. Johnson et al [14] proposed a graph-convolutional model to generate images from struc-

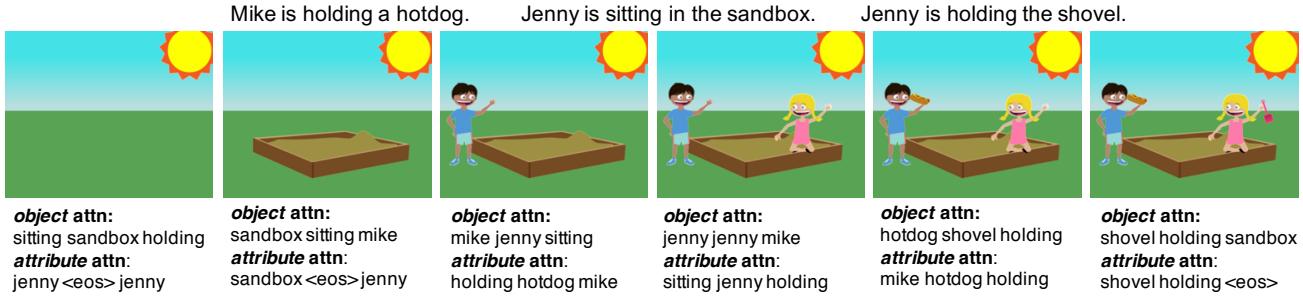


Figure 3. Step-by-step generation of an abstract scene, showing the top-3 attended words for the object prediction and attribute prediction at each time step. Notice how except for predicting the *sun* at the first time step, the top-1 attended words in the object decoder are almost one-to-one mappings with the predicted objects. The attended words by the attribute decoder also correspond semantically to useful information for predicting either pose or location, e.g. to predict the location of the *hotdog* at the fifth time step, the model attends to *mike* and *holding*.

tured scene graphs. The presented objects and their relationships were provided as inputs in the scene graphs, while in our work, the presence of objects is inferred from text. Hong et al [12] targeted image synthesis using conditional GANs but unlike prior works, it generated layouts as intermediate representations in a separably trained module. Our work also attempts to predict layouts for photographic image synthesis but unlike [12], we generate pixel-level outputs using a semi-parametric retrieval module without adversarial training and demonstrate superior results. Kim et al [16] performed pictorial generation from chat logs, while our work uses text which is considerably more underspecified. Gupta et al [9] proposed a semi-parametric method to generate cartoon-like pictures. However the presented objects were also provided as inputs to the model, and the predictions of layouts, foregrounds and backgrounds were performed by separably trained modules. Our method is trained end-to-end and goes beyond cartoon-like scenes. To the best of our knowledge, our model is the first work targeting various types of scenes (e.g. abstract scenes, semantic layouts and composite images) under a unified framework.

3. Model

Text2Scene adopts a Seq-to-Seq framework [31] and introduces key designs for spatial and sequential reasoning. Specifically, at each time step, the model modifies a background canvas in three steps: (1) the model attends to the input text to decide what is the next object to add, or decide whether the generation should end; (2) if the decision is to add a new object, the model zooms in the language context of the object to decide its attributes (e.g. pose, size) and relations with its surroundings (e.g. location, interactions with other objects); (3) the model refers back to the canvas and grounds (places) the extracted textual attributes into their corresponding visual representations.

To model this procedure, Text2Scene consists of a text encoder, which takes as input a sequence of M words w_i (section 3.1), an object decoder, which predicts sequentially T objects o_t , and an attribute decoder that predicts for each o_t their locations l_t and a set of k attributes $\{R_t^k\}$ (section 3.2). The scene generation starts from an initially empty canvas B_0 that is updated at each time step. In the synthetic image generation task, we also jointly train a foreground patch embedding network (section 3.3) and treat the embedded vector as a target attribute. Figure 3 shows a step-by-step generation of an abstract scene.

3.1. Text Encoder

Our text encoder is a bidirectional recurrent network with Gated Recurrent Units (GRUs). For a given sentence, we compute for each word w_i :

$$h_i^E = \text{BiGRU}(x_i, h_{i-1}^E, h_{i+1}^E), \quad (1)$$

Here BiGRU is a bidirectional GRU cell, x_i is a word embedding vector corresponding to the i -th word w_i , and h_i^E is a hidden vector encoding the current word and its context. We use the pairs $[h_i^E; x_i]$, the concatenation of h_i^E and x_i , as the encoded text feature.

3.2. Object and Attribute Decoders

At each step t , our model predicts the next object o_t from an object vocabulary \mathcal{V} and its k attributes $\{R_t^k\}$, using text feature $\{[h_i^E; x_i]\}$ and the current canvas B_t as input. For this part, we use a convolutional network (CNN) Ω to encode B_t into a $C \times H \times W$ feature map, representing the current scene state. We model the history of the scene states $\{h_t^D\}$ with a convolutional GRU (ConvGRU):

$$h_t^D = \text{ConvGRU}(\Omega(B_t), h_{t-1}^D), \quad (2)$$

The initial hidden state is created by spatially replicating the last hidden state of the text encoder. Here h_t^D provides

an informative representation of the temporal dynamics of each spatial (grid) location in the scene. Since this representation might fail to capture small objects, a one-hot vector of the object predicted at the previous step o_{t-1} is also provided as input to the downstream decoders. The initial object is set as a special start-of-scene token.

Attention-based Object Decoder: Our object decoder is an attention-based model that outputs the likelihood scores of all possible objects in an object vocabulary \mathcal{V} . It takes as input the recurrent scene state h_t^D , text features $\{[h_i^E; x_i]\}$ and the previously predicted object o_{t-1} :

$$u_t^o = \text{AvgPooling}(\Psi^o(h_t^D)), \quad (3)$$

$$c_t^o = \Phi^o([u_t^o; o_{t-1}], \{[h_i^E; x_i]\}), \quad (4)$$

$$p(o_t) \propto \Theta^o([u_t^o; o_{t-1}; c_t^o]), \quad (5)$$

here Ψ^o is a convolutional network with spatial attention on h_t^D , similar as in [35]. The goal of Ψ^o is to collect the spatial contexts necessary for the object prediction, e.g. what objects have already been added. The attended spatial features are then fused into a vector u_t^o by average pooling. Φ^o is the text-based attention module, similar as in [22], which uses $[u_t^o; o_{t-1}]$ to attend to the language context $\{[h_i^E; x_i]\}$ and collect the context vector c_t^o . Ideally, c_t^o encodes information about all the described objects that have not been added to the scene thus far. Θ^o is a two-layer perceptron predicting the likelihood of the next object $p(o_t)$ from the concatenation of u_t^o , o_{t-1} , and c_t^o , using a softmax function.

Attention-based Attribute Decoder The attribute set corresponding to the object o_t can be predicted similarly. We use another attention module Φ^a to “zoom in” the language context of o_t , extracting a new context vector c_t^a . Compared with c_t^o which may contain information of all the objects that have not been added yet, c_t^a focuses more specifically on contents related to the current object o_t . For each spatial location in h_t^D , the model predicts a location likelihood l_t , and a set of attribute likelihoods $\{R_t^k\}$. Here, possible locations are discretized into the same spatial resolution of h_t^D .

In summary, we have

$$\begin{aligned} &\text{context related to } o_t && \text{Current object} && \text{Context} \\ &\text{current object} && c_t^a = \Phi^a(o_t, \{[h_i^E; x_i]\}) && \uparrow \\ &\text{encoding of scene} && u_t^a = \Psi^a([h_t^D; c_t^a]) && \downarrow \\ &\text{dynamics} && p(l_t, \{R_t^k\}) = \Theta^a([u_t^a; o_t; c_t^a]), && \end{aligned} \quad (6) \quad (7) \quad (8)$$

Φ^a is a text-based attention module aligning o_t with the language context $\{[h_i^E; x_i]\}$. Ψ^a is an image-based attention module aiming to find an affordable location to add o_t . Here c_t^a is spatially replicated before concatenating with h_t^D . The final likelihood map $p(l_t, \{R_t^k\})$ is predicted by a convolutional network Θ^a , followed by softmax classifiers for l_t and discrete $\{R_t^k\}$. For continuous attributes $\{R_t^k\}$ such as the appearance vector Q_t for patch retrieval (next section), we normalize the output using an ℓ_2 -norm.

3.3. Foreground Patch Embedding

We predict a particular attribute: an appearance vector Q_t , only for the model trained to generate synthetic image composites (i.e. images composed of patches retrieved from other images). As with other attributes, Q_t is predicted for every location in the output feature map which is used at test time to retrieve similar patches from a pre-computed collection of object segments from other images. We train a patch embedding network using a CNN which reduces the foreground patch in the target image into a 1D vector F_t . The goal is to minimize the ℓ_2 -distance between Q_t and F_t using a triplet embedding loss [6] that encourages a small distance $\|Q_t, F_t\|_2$ but a larger distance with other patches $\|Q_t, F_k\|_2$. Here F_k is the feature of a “negative” patch, which is randomly selected from the same category of F_t :

$$L_{triplet}(Q_t, F_t) = \max\{\|Q_t, F_t\|_2 - \|Q_t, F_k\|_2 + \alpha, 0\} \quad (9)$$

α is a margin hyper-parameter.

3.4. Objective encoding of foreground patch

The loss function for a given example in our model with reference values $(o_t, l_t, \{R_t^k\}, F_t)$ is:

$$\begin{aligned} L = & -w_o \sum_t \log p(o_t) - w_l \sum_t \log p(l_t) \\ & - \sum_k w_k \sum_t \log p(R_t^k) + w_e \sum_t L_{triplet}(Q_t, F_t) \\ & + w_a^O L_{attn}^O + w_a^A L_{attn}^A, \end{aligned} \quad \text{synthetic}$$

where the first three terms are negative log-likelihood losses corresponding to the object, location, and discrete attribute softmax classifiers. $L_{triplet}$ is a triplet embedding loss optionally used for the synthetic image generation task. L_{attn}^* are regularization terms inspired by the doubly stochastic attention module proposed in [35]. Here $L_{attn}^* = \sum_i [1 - \sum_t \alpha_{ti}^*]^2$ where $\{\alpha_{ti}^o\}$ and $\{\alpha_{ti}^a\}$ are the attention weights from Φ^o and Φ^a respectively. These regularization terms encourage the model to distribute the attention across all the words in the input sentence so that it will not miss any described objects. Finally, w_o , w_l , $\{w_k\}$, w_e , w_a^O , and w_a^A are hyperparameters controlling the relative contribution of each loss.

Details for different scene generation tasks In the Abstract Scenes generation task, B_t is represented directly as an RGB image. In the layout generation task, B_t is a 3D tensor with a shape of (\mathcal{V}, H, W) , where each spatial location contains a one-hot vector indicating the semantic label of the object at that location. Similarly, in the synthetic image generation task, B_t is a 3D tensor with a shape of $(3\mathcal{V}, H, W)$, where every three channels of this tensor encode the color patches of a specific category from the background canvas image. For the foreground embedding module, we adopt the canvas representation in [27] to encode

Methods	U-obj		B-obj		Pose	Expr	U-obj Coord	B-obj Coord
	Prec	Recall	Prec	Recall				
Zitnick et al. [44]	0.722	0.655	0.280	0.265	0.407	0.370	0.449	0.416
Text2Scene (w/o attention)	0.665	0.605	0.228	0.186	0.305	0.323	0.395	0.338
Text2Scene (w object attention)	0.731	0.671	0.312	0.261	0.365	0.368	0.406	0.427
Text2Scene (w both attentions)	0.749	0.685	0.327	0.272	0.408	0.374	0.402	0.467
Text2Scene (full)	0.760	0.698	0.348	0.301	0.418	0.375	0.409	0.483

Table 1. Quantitative evaluation on the Abstract Scenes dataset. Our full model performs better in all metrics except U-obj Coord which measures exact object coordinates. It also shows that our sequential attention approach is effective.

Methods	Scores	≥ 1		Obj-Single		Obj-Pair		Location	Expression
		≥ 1	≥ 2	sub-pred	sub-pred-obj	pred:loc	pred:expr		
Reference	0.919	1.0	0.97	0.905		0.88		0.933	0.875
Zitnick et al. [44]	0.555	0.92	0.49	0.53		0.44		0.667	0.625
Text2Scene (w/o attention)	0.455	0.75	0.42	0.431		0.36		0.6	0.583
Text2Scene (full)	0.644	0.94	0.62	0.628		0.48		0.667	0.708

Table 2. Human evaluation on Abstract Scenes. Each scene is generated from three textual statements. Users are asked to rate if the generated scene validates each input statement. Our method generates scenes that abide by at least one of the statements 94% of the times, and by at least two statements 64%, and is superior in all types of statements except Location.

the foreground patch for simplicity. As the composite images may exhibit gaps between patches, we also leverage the stitching network in [27] for post-processing. Note that the missing region may also be filled by any other inpainting approaches. Full details about the implementation of our model can be found in the supplementary material. Our code and data are publicly available¹.

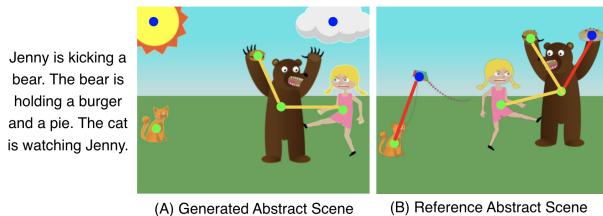


Figure 4. Evaluation metrics for the abstract scene generation task (best viewed in color): the green dots show the common U-obj between the reference (B) and the generated abstract scene (A), the blue dots show the missing and mispredicted objects. Similarly, the yellow lines show the common B-obj and the red lines show the missing and mispredicted B-obj. The U-obj precision/recall for this example is 0.667/0.667, the B-obj precision/recall is 1.0/0.5.

4. Experiments

We conduct experiments on three text-to-scene tasks: (I) constructing abstract scenes of clip-arts in the Abstract Scenes [44] dataset; (II) predicting semantic object layouts of real images in the COCO [21] dataset; and (III) generating synthetic image composites in the COCO [21] dataset.

Task (I): Clip-art Generation on Abstract Scenes We use the dataset introduced by [44], which contains over 1,000 sets of 10 semantically similar scenes of children playing outside. The scenes are composed with 58 clip-art objects. The attributes we consider for each clip-art object are the location, size ($|R^{size}| = 3$), and the direction the object is facing ($|R^{direction}| = 2$). For the person objects, we also explicitly model the pose ($|R^{pose}| = 7$) and expression ($|R^{expression}| = 5$). There are three sentences describing different aspects of a scene. After filtering empty scenes, we obtain 9997 samples. Following [44], we reserve 1000 samples as the test set and 497 samples for validation.

Task (II): Semantic Layout Generation on COCO The semantic layouts contain bounding boxes of the objects from 80 object categories defined in the COCO [21] dataset. We use the val2017 split as our test set and use 5000 samples from the train2017 split for validation. We normalize the bounding boxes and order the objects from bottom to top as the y-coordinates typically indicate the distances between the objects and the camera. We further order the objects with the same y-coordinate based on their x-coordinates (from left to right) and categorical indices. The attributes we consider are location, size ($|R^{size}| = 17$), and aspect ratio ($|R^{aspect_ratio}| = 17$). For the size attribute, we discretize the normalized size range evenly into 17 scales. We also use 17 aspect ratio scales, which are $\{\frac{1}{9}, \frac{1}{8}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{1}{1}, \frac{2}{1}, \frac{3}{1}, \frac{4}{1}, \frac{5}{1}, \frac{6}{1}, \frac{7}{1}, \frac{8}{1}, \frac{9}{1}\}$.

Task (III): Synthetic Image Generation on COCO We demonstrate our approach by generating synthetic image composites given input captions in COCO [21]. For fair comparisons with alternative approaches, we use the val2014 split as our test set and use 5000 samples from the

¹<https://github.com/uvavision/Text2Scene>

Methods	B1	B2	B3	B4	METEOR	ROUGE	CIDEr	SPICE
Captioning from True Layout [39]	0.678	0.492	0.348	0.248	0.227	0.495	0.838	0.160
Text2Scene (w/o attention)	0.591	0.391	0.254	0.169	0.179	0.430	0.531	0.110
Text2Scene (w object attention)	0.591	0.391	0.256	0.171	0.179	0.430	0.524	0.110
Text2Scene (w both attentions)	0.600	0.401	0.263	0.175	0.182	0.436	0.555	0.114
Text2Scene (full)	0.615	0.415	0.275	0.185	0.189	0.446	0.601	0.123

Table 3. Quantitative evaluation on the layout generation task. Our full model generates more accurate captions from the generated layouts than the baselines. We also include **caption generation results from ground truth layouts as an upper bound** on this task.

Methods	IS	B1	B2	B3	B4	METEOR	ROUGE	CIDEr	SPICE
Real image	36.00±0.7	0.730	0.563	0.428	0.327	0.262	0.545	1.012	0.188
GAN-INT-CLS [29]	7.88±0.07	0.470	0.253	0.136	0.077	0.122	–	0.160	–
SG2IM* [14]	6.7±0.1	0.504	0.294	0.178	0.116	0.141	0.373	0.289	0.070
StackGAN [40]	10.62±0.19	0.486	0.278	0.166	0.106	0.130	0.360	0.216	0.057
HDGAN [41]	11.86±0.18	0.489	0.284	0.173	0.112	0.132	0.363	0.225	0.060
Hong et al [12]	11.46±0.09	0.541	0.332	0.199	0.122	0.154	–	0.367	–
AttnGan [36]	25.89±0.47	0.640	0.455	0.324	0.235	0.213	0.474	0.693	0.141
Text2Scene (w/o inpaint.)	22.33±1.58	0.602	0.412	0.288	0.207	0.196	0.448	0.624	0.126
Text2Scene (w inpaint.)	24.77±1.59	0.614	0.426	0.300	0.218	0.201	0.457	0.656	0.130

Table 4. Quantitative evaluation on the **synthetic image generation task**. Our model is superior on automated metrics than all competing approaches except AttnGan, even without post-processing. *The result of SG2IM is evaluated on the validation set defined in [14], which is a subset of the COCO val2014 split.

	Ratio
Text2Scene > SG2IM [14]	0.7672
Text2Scene > HDGAN [41]	0.8692
Text2Scene > AttnGAN [36]	0.7588

Table 5. Two-alternative forced-choiced evaluation on the synthetic image generation task against competing methods.

train2014 split for validation. We collect **segmented object and stuff patches** from the training split. The stuff segments are extracted from the training images by taking connected components in corresponding semantic label maps from the **COCO-Stuff annotations** [11]. For object segments, we use all 80 categories defined in COCO. For stuff segments, we use the 15 super-categories defined in [11] as the class labels, which results in 95 categories in total. We order the patches as in the layout generation task but when composing the patches, we always render the object patches in front of the stuff patches. In our experiment, Q_t and F_t have a dimension of 128.

4.1. Evaluation

Automatic Metrics Our tasks pose new challenges on evaluating the models as (1) the three types of scene representations are quite different from each other; and (2) there is no absolute one-to-one correspondence between a sentence and a scene. For the abstract scene generation task, we draw inspiration from the evaluation metrics applied in machine translation [19] but we aim at aligning multimodal visual-linguistic data instead. To this end, we propose to

compute the following metrics: precision/recall on single objects ($U\text{-obj}$), “bigram” object pairs ($B\text{-obj}$); classification accuracies for poses, expressions; Euclidean distances (defined as a Gaussian function with a kernel size of 0.2) for normalized coordinates of $U\text{-obj}$ and $B\text{-obj}$. A “bigram” object pair is defined as a pair of objects with overlapping bounding boxes as illustrated in Figure 4.

In the layout generation experiment, it is harder to define evaluation metrics given the complexity of real world object layouts. Inspired by [12], we employ caption generation as an extrinsic evaluation. **We generate captions from the semantic layouts using [39] and compare them back to the original captions used to generate the scenes.** We use commonly used metrics for captioning such as BLEU [25], METEOR [19], ROUGE_L [20], CIDEr [32] and SPICE [1].

For synthetic image generation, we adopt the Inception Score (IS) metric [30] which is commonly used in recent text to image generation methods. However, as IS does not evaluate correspondence between images and captions, we also employ an extrinsic evaluation using image captioning using the Show-and-Tell caption generator [34] as in [12].

Baselines For abstract scene generation, we run comparisons with [44]. We also consider variants of our full model: (1) Text2Scene (w/o attention): a model without any attention module. In particular, we replace Eq. 3 with a pure average pooling operation on h_t^D , discard c_t^o in Eq. 5, discard c_t^a and replace u_t^a with h_t^D in Eq. 8. (2) Text2Scene (w object attention): a model with attention modules for object prediction but no dedicated attention for attribute predic-

Input	Zitnick et al. 2013	Text2Scene (w/o Attention)	Text2Scene	Reference
Jenny is wearing sunglasses. Mike is holding the red shovel. Mike is wearing a viking head.				
Mike went down the slide fast. Jenny is worried that Mike is hurt. Jenny is wearing a chef hat.				
Mike is angry at Jenny. Jenny is sad that Mike took the frisbee. The pizza is on the table.				
Jenny is holding a bucket and shovel. Mike fell off the swingset. There is rain and lightning in the sky				

Figure 5. Examples of generated abstract scenes. The first column shows the input text, and the last column shows the reference *true* scene from the dataset.

tion. Specifically, we replace (u_t^a, c_t^a) with (h_t^D, c_t^o) in Eq. 8. (3) Text2Scene (w both attentions): a model with dedicated attention modules for both object and attribute predictions but no regularization.

Human Evaluations We also conduct human evaluations using crowdsourcing on 100 groups of clip-art scenes generated for the Abstract Scene dataset using random captions from the test split. Human annotators are asked to determine whether an input text is a true statement given the generated scene (entailment). Each scene in this dataset is associated with three sentences that are used as the statements. Each sentence-scene pair is reviewed by three annotators to determine if the entailment is *true*, *false* or *uncertain*. Ignoring *uncertain* responses, we use the ratio of the sentence-scene pairs marked as *true* for evaluation.

We also perform predicate-argument semantic frame analysis [3] on our results. Using the semantic parser from [44], we subdivide the sentences as: *sub-pred* corresponding to sentences referring to only one object, *sub-pred-obj* corresponding to sentences referring to object pairs with semantic relations, *pred:loc* corresponding to sentences referring to locations, and *pred:pa* corresponding to sentences mentioning facial expressions.

For synthetic image generation we use a similar human evaluation as in [27]. We compare our method against SG2IM [14], HDGAN [41] and AttnGAN [36]. We resize our generated images to the same resolutions as in these alternative methods, 64×64 for SG2IM [14], 256×256 for HDGAN [41] and AttnGAN [36]. For each sentence randomly selected from the test set, we present images generated by our method and a competing method and allow the user to choose the one which better represents the text. We collect results for 500 sentences. For each sentence, we collect responses from 5 different annotators.

Input Caption	Predicted Layout	Reference Layout	Reference Image	Generated Caption
A happy couple is cutting a decorated cake .				A woman and a woman are cutting a cake
Four giraffes are reaching in the tree for food.				A couple of giraffes are standing in a field.
A gray cat standing on the top of a refrigerator .				A cat is sitting in a room.

Figure 6. Generated layouts from input captions and generated captions from the predicted layouts (best viewed in color). Our model successfully predicts the presence (purple text) and number of objects (blue text), and their spatial relations (red text).

4.2. Results and Discussion

Abstract Scenes and Semantic Layouts: Table 1 shows quantitative results on Abstract Scenes. Text2Scene improves over [44] and our variants on all metrics except U-obj Coord. Human evaluation results on Table 2 confirm the quality of our outputs, where Scores are the percentage of sentence-scene pairs with a *true* entailment; (≥ 1) (≥ 2) indicate if our method produces scenes that entailed at least one (or two) out of three statements. Text2Scene also shows better results on statements with specific semantic information such as Obj-single, Obj-pair, and Expression, and is comparable on Location statements. As a sanity check, we also test reference *true* scenes provided in the Abstract Scenes dataset (first row). Results show that it is more challenging to generate the semantically related object pairs. Overall, the results also suggest that our proposed metrics correlate with human judgment on the task.

Figure 5 shows qualitative examples of our models on Abstract Scenes in comparison with baseline approaches and the reference scenes. These examples illustrate that Text2Scene is able to capture semantic nuances such as the spatial relation between two objects (e.g., the bucket and the shovel are correctly placed in Jenny’s hands in the last row) and object locations (e.g., Mike is on the ground near the swing set in the last row).

Table 3 shows an extrinsic evaluation on the layout generation task. We perform this evaluation by generating captions from our predicted layouts. Results show our full method generates the captions that are closest to the captions generated from true layouts. Qualitative results in Figure 6 also show that our model learns important visual concepts such as presence and number of object instances, and their spatial relations.

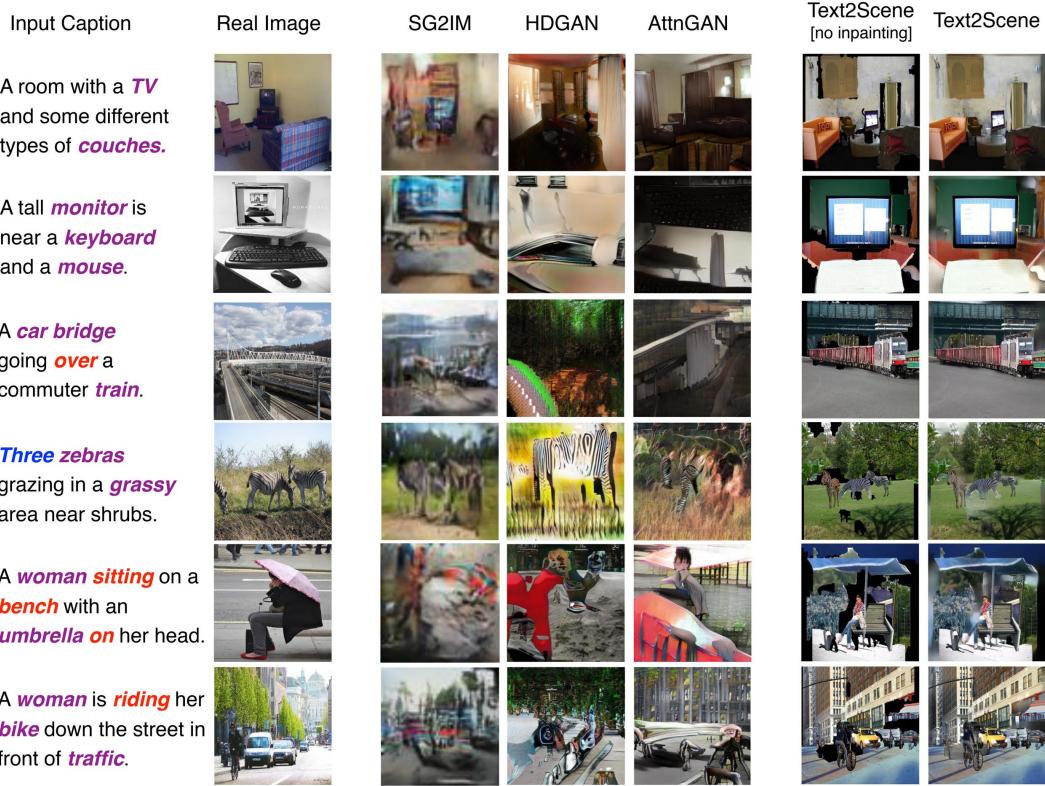


Figure 7. Qualitative examples of synthetic image generation (best viewed in color). The first column shows input captions with manually highlighted objects (purple), counts (blue) and relations (red). The second columns shows the *true* images. Columns in the middle show competing approaches. The last two columns show the outputs of our model before and after pre-processing.

Synthetic Image Composites: Table 4 shows evaluation of synthetic scenes using automatic metrics. Text2Scene without any post-processing already outperforms all previous methods by large margins except AttnGAN [36]. As our model adopts a composite image generation framework without adversarial training, gaps between adjacent patches may result in unnaturally shaded areas. We observe that, after performing a regression-based inpainting [27], the composite outputs achieve consistent improvements on all automatic metrics. We posit that our model can be further improved by incorporating more robust post-processing or in combination with GAN-based methods. On the other hand, human evaluations show that our method significantly outperforms alternative approaches including AttnGAN [36], demonstrating the potential of leveraging realistic image patches for text-to-image generation. It is important to note that SG2IM [14] and Hong et al [12] also use segment and bounding box supervision – as does our method, and AttnGan [36] uses an Imagenet (ILSVRC) pretrained Inceptionv3 network. In addition, as our model contains a patch retrieval module, it is important that the model does not generate a synthetic image by simply retrieving patches from a single training image. On average, each composite image generated from our model contains 8.15 patches from 7.38

different source images, demonstrating that the model does not simply learn a global image retrieval. Fig. 7 shows qualitative examples of the synthetic image composites. We include examples of generated images along with their corresponding source images from which patch segments are retrieved, and more extensive qualitative results in the supplemental material. Since our model learns about objects and relations separately, we also observed that it is often able to generalize to uncommon situations (as defined in [38]).

5. Conclusions

This work presents a novel sequence-to-sequence model for generating compositional scene representations from visually descriptive language. We provide extensive quantitative and qualitative analysis of our model for different scene generation tasks on datasets from two different domains: Abstract Scenes [44] and COCO [21]. Experimental results demonstrate the capacity of our model to capture finer semantic concepts from visually descriptive text and generate complex scenes.

Acknowledgements: This work was partially supported by an IBM Faculty Award to V.O, and gift funding from SAP Research.

Supplementary Material

A. Network Architecture

Here we describe the network architectures for the components of our model in different tasks.

A.1. Text Encoder

We use the same network architecture for the text encoders in all our experiments, which consists of a single layer bidirectional recurrent network with Gated Recurrent Units (GRUs). It takes a linear embedding of each word as input and has a hidden dimension of 256 for each direction. We initialize the word embedding network with the pre-trained parameters from GloVe [26]. The word embedding vectors are kept fixed for abstract scene and semantic layout generations but finetuned for synthetic image generation.

A.2. Scene Encoder

The scene encoder Ω for abstract scene generation is an Imagenet (ILSVRC) pre-trained ResNet-34 [10]. Its parameters are fixed in all the experiments on Abstract Scene [44]. For layout and synthetic image generations, we develop our own scene encoders as the inputs for these tasks are not RGB images.

Table 6 and 7 show the architecture details. Here $|\mathcal{V}|$ is the size of the categorical vocabulary. In the layout generation task, $|\mathcal{V}|$ is 83, including 80 object categories in COCO [21] and three special categorical tokens: *sos*, *eos*, *pad*, representing the start and end points for sequence generation and the padding token. For synthetic image generation, $|\mathcal{V}|$ is 98, including 80 object categories in COCO [21], 15 supercategories for stuffs in COCO-stuff [11] and the special categorical tokens: *sos*, *eos*, *pad*.

As described in the main paper, the input for synthetic image generation has a layer-wise structure where every three channels contain the color patches of a specific category from the background canvas image. In this case, the categorical information of the color patches can be easily learned. On the other hand, since the input is a large but sparse volume with very few non-zero values, to reduce the number of parameters and memory usage, we use a depthwise separable convolution as the first layer of Ω (index (2)), where each group of three channels (g3) is convolved to one single channel in the output feature map.

A.3. Convolutional Recurrent Module

The scene recurrent module for all our experiments is a convolutional GRU network [45] with one ConvGRU cell. Each convolutional layer in this module have a 3×3 kernel with a stride of 1 and a hidden dimension of 512. We pad the input of each convolution so that the output feature map has

Index	Input	Operation	Output Shape
(1)	-	Input	$ \mathcal{V} \times 64 \times 64$
(2)	(1)	Conv(7×7 , $ \mathcal{V} \rightarrow 128$, s2)	$128 \times 32 \times 32$
(3)	(2)	Residual($128 \rightarrow 128$, s1)	$128 \times 32 \times 32$
(4)	(3)	Residual($128 \rightarrow 256$, s2)	$256 \times 16 \times 16$
(5)	(4)	Bilateral upsampling	$256 \times 28 \times 28$

Table 6. Architecture of our scene encoder Ω for layout generation. We follow the notation format used in [14]. Here $|\mathcal{V}|$ is the size of the categorical vocabulary. The input and output of each layer have a shape of $C \times H \times W$, where C is the number of channels and H and W are the height and width. The notation $Conv(K \times K, C_{in} \rightarrow C_{out})$ represents a convolutional layer with $K \times K$ kernels, C_{in} input channels and C_{out} output channels. The notation s2 means the convolutional layer has a stride of 2. The notation $Residual(C_{in} \rightarrow C_{out})$ is a residual module consisting of two 3×3 convolutions and a skip-connection layer. In the first residual block (index (3)), the skip-connection is an identity function and the first convolution has a stride of 1 (s1). In the second residual block (index (4)), the skip-connection is a 1×1 convolution with a stride of 2 (s2) and the first convolution also has a stride of 2 to downsample the feature map. Here all the convolutional layers are followed by a ReLU activation.

Index	Input	Operation	Output Shape
(1)	-	Input	$3 \mathcal{V} \times 128 \times 128$
(2)	(1)	Conv(7×7 , $3 \mathcal{V} \rightarrow \mathcal{V} $, s2, g3)	$ \mathcal{V} \times 64 \times 64$
(3)	(2)	Residual($ \mathcal{V} \rightarrow \mathcal{V} $, s1)	$ \mathcal{V} \times 64 \times 64$
(4)	(3)	Residual($ \mathcal{V} \rightarrow 2 \mathcal{V} $, s1)	$2 \mathcal{V} \times 64 \times 64$
(5)	(4)	Residual($2 \mathcal{V} \rightarrow 2 \mathcal{V} $, s1)	$2 \mathcal{V} \times 64 \times 64$
(6)	(5)	Residual($2 \mathcal{V} \rightarrow 3 \mathcal{V} $, s2)	$3 \mathcal{V} \times 32 \times 32$
(7)	(6)	Residual($3 \mathcal{V} \rightarrow 3 \mathcal{V} $, s1)	$3 \mathcal{V} \times 32 \times 32$
(8)	(7)	Residual($3 \mathcal{V} \rightarrow 4 \mathcal{V} $, s1)	$4 \mathcal{V} \times 32 \times 32$

Table 7. Architecture of our scene encoder Ω for synthetic image generation. The notations are in the same format of Table 6. The first convolution (index (2)) is a depthwise separable convolution where each group of three channels (g3) is convolved to one single channel in the output feature map. All the convolutional layers are followed by a LeakyReLU activation with a negative slope of 0.2.

the same spatial resolution as the input. The hidden state is initialized by spatially replicating the last hidden state from the text encoder.

A.4. Object and Attribute Decoders

Table 8 shows the architectures for our object and attribute decoders. Ψ^o and Ψ^a are the spatial attention modules consisting of two convolutional layers. Θ^o is a two-layer perceptron predicting the likelihood of the next object using a softmax function. Θ^a is a four-layer CNN predicting the likelihoods of the location and attributes of the object. As explained in the main paper, the output of Θ^a has $1 + \sum_k |R^k|$ channels, where $|R^k|$ denotes the discretized range of the k-th attribute, or the dimension of the appearance vector Q_t used as the query for patch retrieval for synthetic image generation. The first channel of the output from Θ^a predicts the location likelihoods which are nor-

Module	Index	Input	Operation	Output Shape
Ψ^o	(1)	-	Conv(3×3, 512→256)	256 × H × W
	(2)	(1)	Conv(3×3, 256→1)	1 × H × W
Ψ^a	(1)	-	Conv(3×3, 1324→256)	256 × H × W
	(2)	(1)	Conv(3×3, 256→1)	1 × H × W
Θ^o	(1)	-	Linear((1324 + V)→512)	512
	(2)	(1)	Linear(512→ V)	V
Θ^a	(1)	-	Conv(3×3, (1324+ V)→512)	512 × H × W
	(2)	(1)	Conv(3×3, 512→256)	256 × H × W
	(3)	(2)	Conv(3×3, 256→256)	256 × H × W
	(4)	(3)	Conv(3×3, 256→(1 + $\sum_k R^k $))	(1 + $\sum_k R^k $) × H × W

Table 8. Architectures for the object and attribute decoders. The notation $Linear(C_{in} \rightarrow C_{out})$ represents a fully connected layer with C_{in} input channels and C_{out} output channels. All layers, except the last layer of each module, are followed by a ReLU activation.

malized over the spatial domain using a softmax function. The rest channels predict the attributes for every grid location. During training, the likelihoods from the ground-truth locations are used to compute the loss. At each step of the test time, the top-1 location is first sampled from the model. The attributes are then collected from this sampled location. The text-based attention modules are defined similarly as in [22]. When denoting $d_i = [h_i^E; x_i]$, $s_t^o = [u_t^o; o_{t-1}]$, and $s_t^a = o_t$, Φ^o and Φ^a are defined as:

$$c_t^* = \Phi^*(s_t^*, \{d_i\}) = \sum_i \frac{\exp(\text{score}(s_t^*, d_i))}{\sum_j \exp(\text{score}(s_t^*, d_j))} \cdot d_i$$

$$\text{score}(s_t^*, d_k) = (s_t^*)^\top W_\Phi^* d_k, * \in o, a$$

Here, W_Φ^o and W_Φ^a are trainable matrices which learn to compute the attention scores for collecting the context vectors c_t^o and c_t^a .

These architecture designs are used for all the three generation tasks. The only difference is the grid resolution (H, W). For abstract scene and layout generations, (H, W) = (28, 28). For synthetic image generation, (H, W) = (32, 32). Note that, although our model uses a fixed grid resolution, the composition can be performed on canvases of different sizes.

A.5. Foreground Patch Embedding

The foreground segment representation we use is similar with the one in [27], where each segment P is represented by a tuple $(P^{color}, P^{mask}, P^{context})$. Here $P^{color} \in \mathbb{R}^{3 \times H \times W}$ is a color patch containing the segment, $P^{mask} \in \{0, 1\}^{1 \times H \times W}$ is a binary mask indicating the foreground region of P^{color} , $P^{context} \in \{0, 1\}^{|V| \times H \times W}$ is a semantic map representing the semantic context around P . The context region of P is obtained by computing the bounding box of the segment and enlarging it by 50% in each direction.

Table 9 shows the architecture of our foreground patch embedding network. Here, the concatenation of $(P^{color}, P^{mask}, P^{context})$ is fed into a five-layer convolutional network which reduces the input into a 1D feature vector F_s (index (7)). As this convolutional backbone is relatively

Index	Input	Operation	Output Shape
(1)	-	Input layout	$(V + 4) \times 64 \times 64$
(2)	(1)	Conv(2 × 2, $(V + 4) \rightarrow 256$, s2)	256 × 32 × 32
(3)	(2)	Conv(2 × 2, 256 → 256, s2)	256 × 16 × 16
(4)	(3)	Conv(2 × 2, 256 → 256, s2)	256 × 8 × 8
(5)	(4)	Conv(2 × 2, 256 → 256, s2)	256 × 4 × 4
(6)	(5)	Conv(2 × 2, 256 → 128, s2)	256 × 2 × 2
(7)	(6)	Global average pooling	256
(8)	-	Input patch feature	2048
(9)	(7)(8)	Linear((256 + 2048) → 128)	128

Table 9. Architecture of our foreground patch embedding network for synthetic image generation. All the convolutional layers are followed by a LeakyReLU activation with a negative slope of 0.2.

shallow, F_s is expected to encode the shape, appearance, and context, but may not capture the fine-grained semantic attributes of P . In our experiments, we find that incorporating the knowledge from the pre-trained deep features of P^{color} can help retrieve segments associated with strong semantics, such as the "person" segments. Therefore, we also use the pre-trained features F_d (index (8)) of P^{color} from the mean pooling layer of ResNet152 [10], which has 2048 features. The final vector F_t is predicted from the concatenation of (F_s, F_d) by a linear regression.

A.6. Inpainting Network

Our inpainting network has the same architecture as the image synthesis module proposed in [27], except that we exclude all the layer-normalization layers. To generate the simulated canvases on COCO, we follow the procedures proposed in [27], but make minor modifications: (1) we use the trained embedding patch features to retrieve alternative segments to stencil the canvas, instead of the intersection-over-union based criterion used in [27]. (2) we do not perform boundary elision for the segments as it may remove fine grained details of the segments such as human faces.

B. Optimization

For optimization we use Adam [17] with an initial learning rate of $5e - 5$. The learning rate is decayed by 0.8 every 3 epochs. We clip the gradients in the back-propagation such that the norm of the gradients is not larger than 10. Models are trained until validation errors stop decreasing. For abstract scene generation, we set the hyperparameters $(w_o, w_l, w_{pose}, w_{expression}, w_{size}, w_{direction}, w_a^O, w_a^A)$ to $(8, 2, 2, 2, 1, 1, 1)$. For semantic layout generation, we set the hyperparameters $(w_o, w_l, w_{size}, w_{aratio}, w_a^O, w_a^A)$ to $(5, 2, 2, 2, 1, 0)$. For synthetic image generation, we set the hyperparameters $(w_o, w_l, w_{size}, w_{aratio}, w_a^O, w_a^A, w_e, \alpha)$ to $(5, 2, 2, 2, 1, 0, 10, 0.5)$. The hyperparameters are chosen to make the losses of different components comparable. Exploration of the best hyperparameters is left for future work.

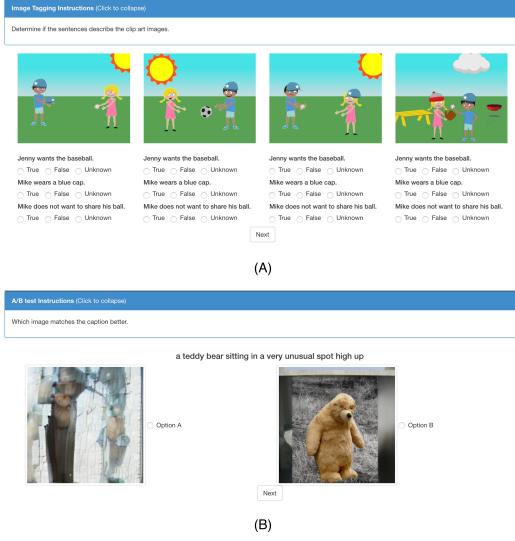


Figure 8. Screen shots of the user interfaces for our human subject studies on Amazon Mechanical Turk. (A) User interface for the evaluation study of the abstract scene generation experiment; (B) User interface for the evaluation study of the synthetic image generation experiment.

C. User Study

We conduct two user studies on Amazon Mechanical Turk (AMT).

The first user study is to evaluate if the generated clip-art scenes match the input sentences. To this end, we randomly select 100 groups of images generated from the sentences in the test set. Each group consists of three images generated by different models, and the ground truth reference image. During the study, these images and the corresponding sentences are presented in random orders. The human annotators are asked to determine if the entailment between the generated scene and the sentence is true, false or uncertain. Each group of images is seen by three annotators. We ignore the uncertain responses and report the results using majority opinions. Figure 8 (A) shows the user interface of this study.

The second user study is on the synthetic image generation task, where we compare the generated images from our model and three state-of-the-art approaches: SG2IM [14], HDGAN [41], and AttnGAN [36]. In each round of the study, the human annotator is presented with one sentence and two generated images: one from our model, the other from an alternative approach. The orders of the images are randomized. We ask the human annotator to select the image which matches the sentence better. In total, we collect results for 500 sentences randomly selected from the test set, using five annotators for each. Figure 8 (B) shows the user interface of this study.

D. More qualitative examples

D.1. Abstract Scene

We present more qualitative examples on Abstract Scene [44] in Fig. 9. The examples show that our model does not simply replicate the ground truth reference scenes, but generates dramatically different clip-arts which still match the input textual descriptions.

D.2. Layout Generation

We present more qualitative examples for layout generation in Fig. 10. The examples include various scenes containing different object categories. Our model manages to learn important semantic concepts from the language, such as the presence and count of the objects, and their spatial relations.

D.3. Synthetic Image Generation

To demonstrate our model does not learn an image-level retrieval on the training set, we present in Fig. 11 the generated images and the corresponding source images from which the patch segments are retrieved for compositing. For each generated image, we show three source images for clarity. The examples illustrate that our model learns not only the presence and spatial layout of objects, but also the semantic knowledge that helps retrieve segments in similar contexts. Fig. 12 shows more qualitative examples of our model for synthetic image generation.

Caption	Text2Scene	Reference	Caption	Text2Scene	Reference
Mike talks to the dog. Jenny kicks the soccer ball. The duck wants to play.			Jenny is wearing the catcher's mitt. Mike is going to throw the tennis ball. There is a plane in the sky.		
The snake wants the cherry pie. The owl wants to eat the snake. The snake likes to play football.			Jenny is on the swing. Jenny is wearing glasses. Mike is sitting alone in the grass.		
Mike is wearing a gold crown. The dog is wearing sunglasses. Rain is coming out of the gray cloud.			Jenny is kicking a football. A cat is sitting on a table. Mike is about to eat a hotdog.		
Jenny offers to share her pie with the bear. Mike found a bunch of balloons. The bear cannot scare Mike and Jenny.			Mike and Jenny are on the swings. Jenny is mad the Frisbee almost hit the cat. The cat is watching Mike and Jenny swing.		
Jenny is kicking her foot. Mike is happy that it is raining. The pie is cooking.			Mike let the balloons fly away. Jenny wants the cold drink. Red apples grow on the tree.		
Mike is kicking the soccer ball. The sandbox is empty. Nobody is playing on the swings.			Mike is wearing a silly hat. Jenny is wearing a viking hat. Jenny and Mike are happy to see the bear.		
Mike is holding a bottle of mustard. Jenny is holding a bottle of ketchup. Jenny is crying because she hates rain.			Jenny is throwing a Frisbee. Mike is wearing a catcher's mitt. Jenny is standing at the tree.		
It is getting stormy in the park. The lightning started a fire in the park. Jenny and Mike are worried about the stormy weather.			Mike and Jenny are mad. The hotdog and drink are on the table. The basketball is in the grass.		
Jenny is wearing a hat. Jenny is walking a dog. There is an owl in the tree.			Mike and Jenny are scared of the snake. Mike and Jenny are holding hands. The cat is sitting next to Jenny.		
The balloon landed in the park. Jenny wants to go see the balloon. Mike is wearing a viking hat.			Mike has a pirate hat on. Some fruit is in the tree. Jenny has some balloons.		

Figure 9. More qualitative examples of the abstract scene generation experiment.

Input Caption	Predicted Layout	Reference Layout	Reference Image	Input Caption	Predicted Layout	Reference Layout	Reference Image
An attractive young woman leads a grey horse through a paddock.				A couple of women ride horses through some water.			
A gray cat standing on the top of a refrigerator .				A cat standing next to of an open refrigerator door.			
A person holding a surf board in a body of water.				This is a man riding a board in the water.			
A laptop computer a keyboard and two monitors.				A woman is riding her bike down the street in front of traffic .			
A man and a woman stand under an umbrella at a street crossing on a rainy day.				Two women walk outside, both holding up umbrellas.			
A bowl full of fresh green apples are kept.				Cat sleeping in front of a powered on laptop.			
A woman holds a phone next to the laptop a child is working on.				The woman sits at the table with the two children doing crafts.			
A man helping a boy on a paddle board in the water.				A man holding a horse, so a little boy can take a ride.			
A small boat in the water beside a sea airplane.				Children are sitting on the side of the boat in the water.			
This is a small kitchen with white cabinets and appliances.				A room with a fireplace and television inside of it.			

Figure 10. More qualitative examples of the layout generation experiment (best viewed in color). The presences (purple), counts (blue), and spatial relations (red) of the objects are highlighted in the captions. The last row shows the cases when the layouts are underspecified in the input captions.

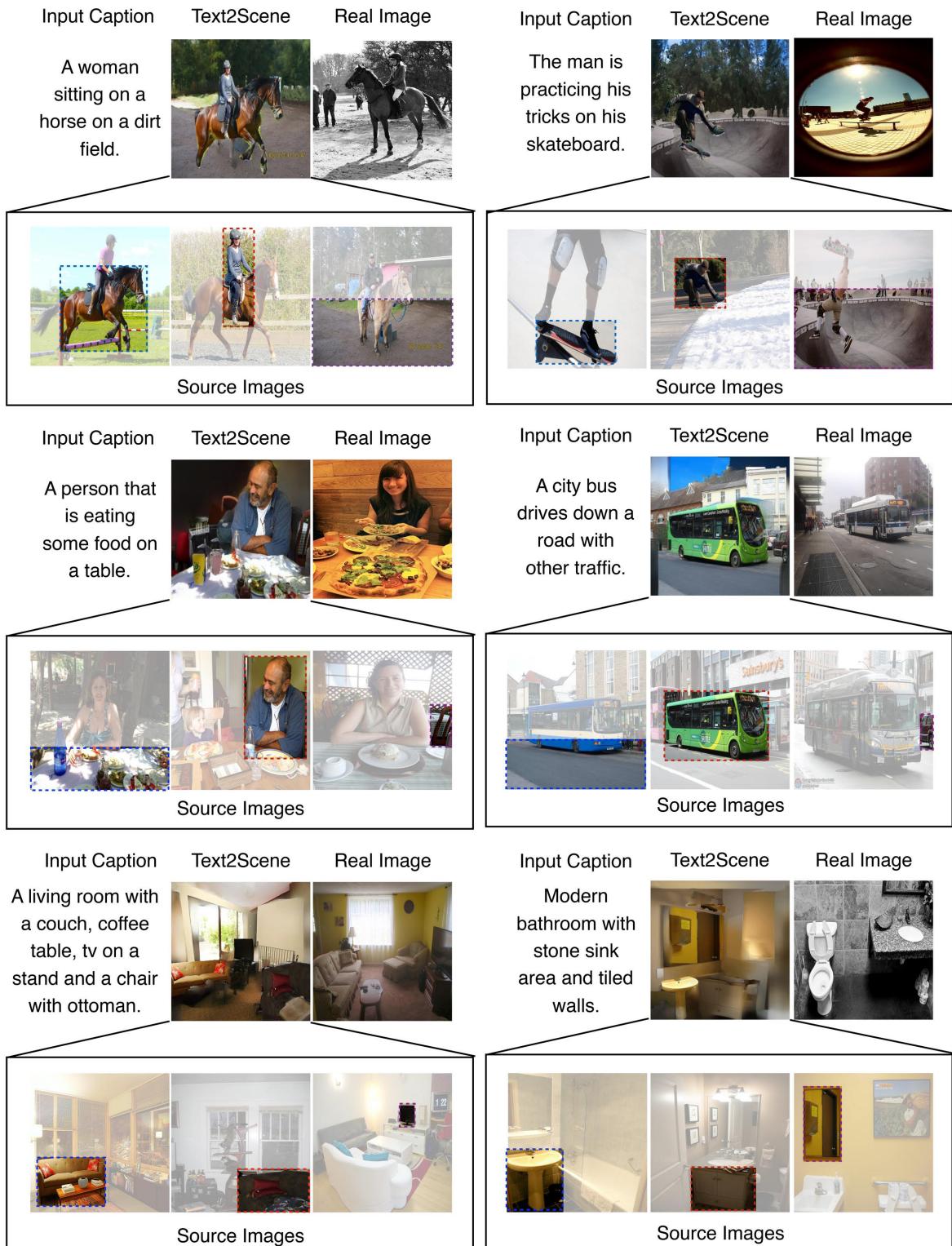


Figure 11. Example synthetic images and the source images from which the patch segments are retrieved for compositing. For each synthetic image, we show three source images for clarity.

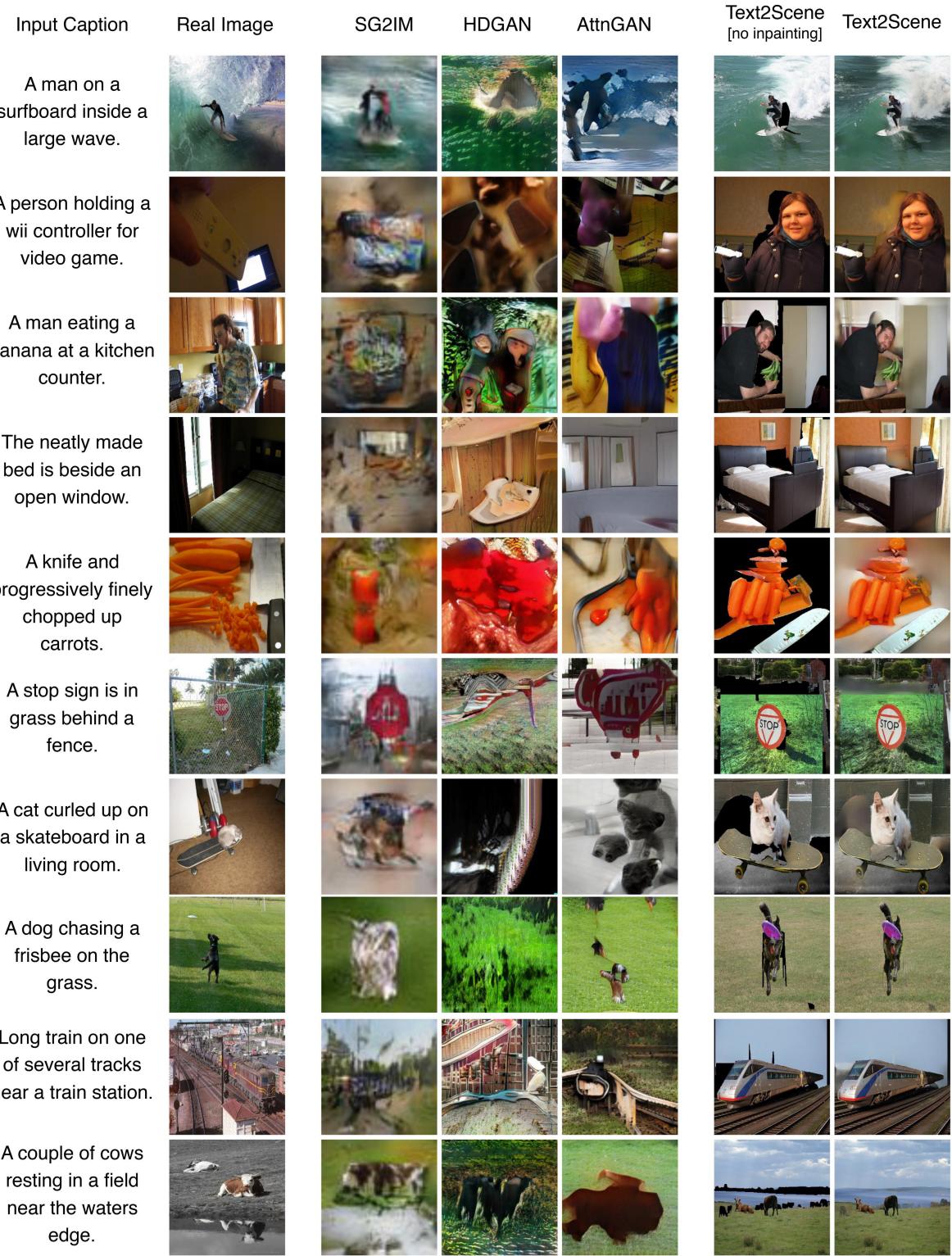


Figure 12. More qualitative examples of the synthetic image generation experiment.

References

- [1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision (ECCV)*, pages 382–398, 2016.
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.
- [3] Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Conference on Computational Natural Language Learning (CoNLL)*, pages 152–164, 2005.
- [4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [5] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision (ECCV)*, pages 15–29, 2010.
- [6] James Philbin, Florian Schroff, Dmitry Kalenichenko. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [7] David F Fouhey and C Lawrence Zitnick. Predicting object dynamics in scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
- [9] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *European Conference on Computer Vision (ECCV)*, 2018.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [11] Jasper Uijlings, Holger Caesar, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [12] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [14] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.
- [16] Jin-Hwa Kim, Devi Parikh, Dhruv Batra, Byoung-Tak Zhang, and Yuandong Tian. Codraw: Visual dialog for collaborative drawing. *arXiv preprint arXiv:1712.05558*, 2017.
- [17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [18] Polina Kuznetsova, Vicente Ordonez, Tamara Berg, and Yejin Choi. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association of Computational Linguistics*, 2(1):351–362, 2014.
- [19] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, 2007.
- [20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.
- [21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV)*, 2014.
- [22] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015.
- [23] Rebecca Mason and Eugene Charniak. Nonparametric method for data-driven image captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 592–598, 2014.
- [24] Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Girish Kulkarni, Margaret Mitchell, Kota Yamaguchi, Karl Stratos, Amit Goyal, Jesse Dodge, Alyssa Mensch, et al. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision (IJCV)*, 119(1):46–59, 2016.
- [25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.
- [26] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [27] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [28] Scott Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [29] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Learning Representations (ICLR)*, 2016.
- [30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques

- for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [31] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
- [32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.
- [33] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. Learning common sense through visual abstraction. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.
- [34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, volume 37, pages 2048–2057, 2015.
- [36] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [37] Mark Yatskar, Vicente Ordonez, and Ali Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, 2016.
- [38] Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi. Commonly uncommon: Semantic sparsity in situation recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7196–7205, 2017.
- [39] Xuwang Yin and Vicente Ordonez. Obj2text: Generating visually descriptive language from object layouts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.
- [40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [41] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [42] Xiaodan Zhu and Edward Grefenstette. Deep learning for semantic composition. In *ACL tutorial*, 2017.
- [43] C. Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [44] C. Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the visual interpretation of sentences. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

Dataset