Object-Centric Image Generation from Layouts

Tristan Sylvain,[1,2] **Pengchuan Zhang,**[3] **Yoshua Bengio,**[1,2,4] **R Devon Hjelm,**[3,1] **Shikhar Sharma**[5]
[1]Mila, Montréal, Canada [2]Université de Montréal, Montréal, Canada [3]Microsoft Research
[4]CIFAR Senior Fellow [5]Microsoft Turing
{tristan.sylvain, yoshua.bengio}@mila.quebec, {penzhan, devon.hjelm, shikhar.sharma}@microsoft.com

Abstract

We begin with the hypothesis that a model must be able to understand individual objects and relationships between objects in order to generate complex scenes with multiple objects well. Our layout-to-image-generation method, which we call Object-Centric Generative Adversarial Network (or OC-GAN), relies on a novel Scene-Graph Similarity Module (SGSM). The SGSM learns representations of the spatial relationships between objects in the scene, which lead to our model's improved layout-fidelity. We also propose changes to the conditioning mechanism of the generator that enhance its object instance-awareness. Apart from improving image quality, our contributions mitigate two failure modes in previous approaches: (1) spurious objects being generated without corresponding bounding boxes in the layout, and (2) overlapping bounding boxes in the layout leading to merged objects in images. Extensive quantitative evaluation and ablation studies demonstrate the impact of our contributions, with our model outperforming previous state-of-the-art approaches on both the COCO-Stuff and Visual Genome datasets. Finally, we address an important limitation of evaluation metrics used in previous works by introducing Scene-FID – an object-centric adaptation of the popular Fréchet Inception Distance metric, that is better suited for multi-object images.

Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have been at the helm of significant recent advances in image generation (Goodfellow et al. 2014; Radford, Metz, and Chintala 2016; Gulrajani et al. 2017; Miyato and Koyama 2018; Brock, Donahue, and Simonyan 2019). Apart from unsupervised image generation, GAN-based image generation approaches have done well at conditional image generation from labels (Radford, Metz, and Chintala 2016; Zhang et al. 2019; Brock, Donahue, and Simonyan 2019), captions (Reed et al. 2016; Zhang et al. 2017; Xu et al. 2018b; Li et al. 2019a; Yin et al. 2019), conversations (Sharma et al. 2018; El-Nouby et al. 2019; Li et al. 2019b), scene graphs (Johnson, Gupta, and Fei-Fei 2018; Mittal et al. 2019; Ashual and Wolf 2019), layouts (Zhao et al. 2019; Sun and Wu 2019), segmentation masks (Park et al. 2019), *etc*. While the success in single-domain or single-object image generation has been remarkable, generating complex scenes with multiple objects is still challenging.

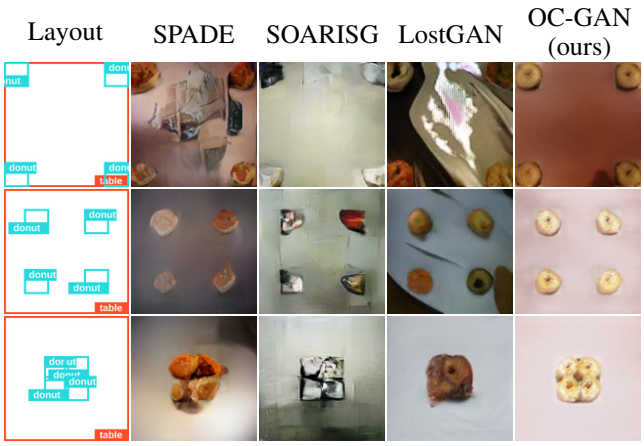

Figure 1: Each row depicts a layout and the corresponding images generated by various models. Along each column, the donuts converge to the centre. In addition to more clearly defined objects, our method is the only one that maintains distinct objects for the final layout, for which bounding boxes slightly overlap.

Generating realistic multi-object scenes is a difficult task because they have many constituent objects (e.g., the Visual Genome dataset, Krishna et al. 2017, can contain as many as 30 different objects in an image). Past methods focus on different input types, including scene graphs (Johnson, Gupta, and Fei-Fei 2018; Ashual and Wolf 2019), pixel-level semantic segmentation (Li et al. 2019a), and bounding box-level segmentation (Zhao et al. 2019; Sun and Wu 2019). In addition, some methods also consider multi-modal data, such as instance segmentation alongside pixel-wise semantic segmentation masks (Park et al. 2019; Wang et al. 2018). Orthogonal to input-related considerations, methods tend to rely on additional components to help with the complexity of scene generation, such as attention mechanisms (Xu et al. 2018b; Li et al. 2019a) and explicit disentanglement of objects from the background (Singh, Ojha, and Lee 2019).

Despite these advances, models still struggle in creating realistic scenes. As shown in Figs. 1 and 2, even sim-

Figure 2: Existing models introduce spurious objects not specified in the layout, a failure mode over which our model improves significantly.

ple layouts can result in merged objects, spurious objects, and images that do not match the given layout (low layout-fidelity). To counter this, we propose Object-Centric GAN (OC-GAN), an architecture to generate realistic images with *high layout-fidelity* and *sharp objects*. Our primary contributions are:

- We introduce a set of novel components that are well-motivated and improve performance for complex scene generation. Our proposed scene-graph-based retrieval module (SGSM) improves layout-fidelity. We also introduce other improvements, such as conditioning on instance boundaries, that help generating sharp objects and realistic scenes.
- Our model improves significantly on the previous state of the art in terms of a set of classical metrics. In addition to standard metrics, we also perform a detailed ablation study to highlight the effect of each component, and a human evaluation study to further validate our findings.
- We discuss the validity of the metrics currently used to evaluate layout-to-image methods, and building on our findings, motivate the use of SceneFID, a new evaluation setting which is more adapted to multi-object datasets.

Related Work

Conditional scene generation For some time, the image generation community has focused on scenes that contain multiple objects in the foreground (Reed et al. 2016; Zhang et al. 2017; Johnson, Gupta, and Fei-Fei 2018). Such scenes, which can contain large amount of objects of very different scales, are very complex relative to single-object images. Several conditional image generation tasks have been formulated using different subsets of annotations. Text-based image generation using captions (Reed et al. 2016; Zhang et al. 2017; Xu et al. 2018b; Li et al. 2019a; Yin et al. 2019) or even multi-turn conversations (Sharma et al. 2018; El-Nouby et al. 2019; Li et al. 2019b) have gained significant interest. However, with increasing numbers of objects and their relationships in the image, understanding long textual captions becomes difficult (Johnson, Gupta, and Fei-Fei 2018; Sharma et al. 2018). Text-based image generation approaches are also not immune to small perturbations in text leading to quite different images (Yin et al. 2019).

Layout-based synthesis Generating images from a given layout makes the analysis more interpretable by decoupling the language understanding problem from the image generation task. Another advantage of generating from layouts is more controllable generation: it is easy to design interfaces to manipulate layouts. In this work we will focus on coarse layouts, where the scene to be generated is specified by bounding-box-level annotations. Layout-based approaches fall into 2 broad categories. Some methods take scene-graphs as inputs, and learn to generate layouts as intermediate representations (Johnson, Gupta, and Fei-Fei 2018; Ashual and Wolf 2019). In parallel, other approaches have focused on generating directly from coarse layouts (Sun and Wu 2019; Zhao et al. 2019). Models that perform well on fine-grained pixel-level semantic maps also can be easily applied to this setting (Park et al. 2019; Isola et al. 2017; Wang et al. 2018). Almost all recent approaches have in common the use of *patch* and *object discriminators* (to ensure whole image and object quality). In addition to this, image quality has been improved by the addition of *perceptual losses* (Park et al. 2019; Ashual and Wolf 2019; Wang et al. 2018), *multi-scale patch-discriminators* (Park et al. 2019), which motivate some of our architecture choices. Finally, modulating the parameters of batch- or instance-normalization layers (Ioffe and Szegedy 2015; Ulyanov, Vedaldi, and Lempitsky 2016) with a function of the input condition can provide significant gains, and this is done per-channel in (Odena, Olah, and Shlens 2017) or per pixel (Park et al. 2019; Sun and Wu 2019). As bounding box layouts are coarse for this task, it is common to introduce *unsupervised mask generators* (Sun and Wu 2019; Ma et al. 2018) to provide estimated shapes for this conditioning.

Finally, there is a growing body of literature involving semi-parametric (Qi et al. 2018; Li et al. 2019c) models that use ground-truth training images to aid generation. We consider the case of such models in the Appendix.

Scene-graphs and image matching Scene graphs are an object-centric representation that can provide an additional useful learning signal when dealing with complex scenes. Scene-graphs are often used as intermediate representations in image captioning (Yang et al. 2019; Anderson et al. 2016), reconstruction (Gu et al. 2019) and retrieval (Johnson et al. 2015), as well as in sentence to scene graph (Schuster et al. 2015) and image to scene graph prediction (Lu et al. 2016; Newell and Deng 2017).

By virtue of being a simpler abstraction of the scene than a layout, they emphasize *instance awareness* more than layouts which focus on pixel-level class labels. Secondly, for scenarios that might require generating multiple diverse images, they provide more variability in reconstruction and matching tasks as the mapping from a scene graph to an image is one to many usually. These points explain their use in higher-level visual reasoning tasks such as visual question answering (Teney, Liu, and van Den Hengel 2017) and zero-shot learning (Sylvain, Petrini, and Hjelm 2020a,b), and also motivate the use of scene graph-based retrieval in our model. In our work, we generate scene graphs depicting positional relationships (such as “to the left of”, “above”, “inside”,

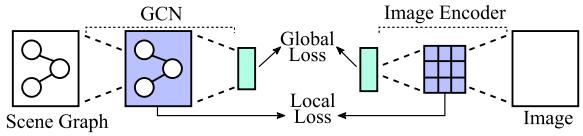

Figure 3: The SGSM module. The SGSM module computes similarity between the scene-graph and the generated image, providing fine-grained matching-based supervision between the positional scene-graph and the generated image.

etc.) from given spatial layouts and leverage them to learn the relationships between objects, which would be more difficult for a model to distill from pixel-level layouts.

There has been strong interest in image and caption similarity modules for retrieval (Fang et al. 2015; Huang et al. 2013) and for text-to-image generation, most recently with the DAMSM model proposed in (Xu et al. 2018b). Despite similar interest in scene graph to image retrieval (Johnson et al. 2015; Quinn et al. 2018), and the large improvements in text-to-image synthesis resulting from the DAMSM (Xu et al. 2018b; Li et al. 2019a), our approach is the first to use a scene graph to image retrieval module when training a generative model.

Proposed Method

Scene-Graph Similarity Module

We introduce the Scene Graph Similarity Module (SGSM) as a means of increasing the *layout-fidelity* of our generated images. This multi-modal module, described summarily in Fig. 3, takes as input an image and a scene-graph (nodes corresponding to objects, and edges corresponding to spatial relations). We extract *local visual features* $\boldsymbol{v}_i$ from the *mixed_6e* layer in an Inception-V3 network (Szegedy et al. 2016) pre-trained on the ImageNet dataset. We extract *global visual features* $\boldsymbol{v}^G$ from the final pooling layer. We encode the graph using a Graph Convolutional Network (GCN, Goller and Kuchler 1996) to obtain *local graph features* $\boldsymbol{g}_j$ and apply a set of graph convolutions followed by a graph pooling operation to obtain *global graph features* $\boldsymbol{g}^G$. Note that each local and global feature is extracted and linearly projected to a common semantic space. In what follows, cos is the cosine similarity, and the γ_ks are normalization constants. We use L/G when the local and global terms are interchangeable. We use the modified dot-product attention mechanism of Xu et al. (2018b) to compute the *visually attended local graph embeddings* $\tilde{\boldsymbol{g}}_j$:

$$\boldsymbol{s}_{ij} = \gamma_1 \frac{\exp\left(\boldsymbol{g}_j{}^T \boldsymbol{v}_i\right)}{\sum_{i'} \exp\left(\boldsymbol{g}_j{}^T \boldsymbol{v}_{i'}\right)}, \quad \tilde{\boldsymbol{g}}_j = \frac{\sum_i \exp(\boldsymbol{s}_{ij}) \boldsymbol{v}_i}{\sum_i \exp(\boldsymbol{s}_{ij})} \quad (1)$$

Then we can define a *local similarity metric* between the source graph embedding $\boldsymbol{g}_j$ and the visually aware local embedding $\tilde{\boldsymbol{g}}_j$ similar to Xu et al. (2018b). Intuitively, the similarity will be strong when the source graph embedding is close to the visually aware embedding. This local similarity will encourage different patches of the image to match the objects expected from the scene graph. The *global similarity metric* is classically the cosine distance between embeddings:

Figure 4: Blue indicates 0 and black indicates 1. (Left) The per-class mask constructed from the layout by many previous methods makes it impossible to distinguish unique object instances in several cases. (Right) Our mask consists of instance boundaries making it easier for the model to distinguish unique object instances using no extra information than already contained in the layout.

$$\begin{cases} \text{Sim}^L(S, I') = \log\left(\sum_j \exp\left(\gamma_2 \cdot \cos(\tilde{\boldsymbol{g}}_j, \boldsymbol{g}_j)\right)\right)^{\frac{1}{\gamma_2}} & (2) \\ \text{Sim}^G(S, I') = \cos\left(\boldsymbol{v}^G, \boldsymbol{g}^G\right) & (3) \end{cases}$$

Finally we can define a global and local probability model in a similar way to e.g. Huang et al. (2013):

$$\mathbb{P}^{L/G}(S, I') \propto \exp\left(\gamma_3 \cdot \text{Sim}^{L/G}(S, I')\right) \quad (4)$$

Normalizing over the images or scenes in the batch B (negative examples are selected by mis-matching the image and scene-graph pairs in the batch) leads to e.g.: $\mathbb{P}^{L/G}(S|I) = \frac{\mathbb{P}^{L/G}(S,I)}{\sum_{I' \in B} \mathbb{P}^{L/G}(S,I')}$. We define the loss terms as the log posterior probability of matching an image I and *the corresponding* scene graph (and vice-versa):

$$\begin{cases} \mathcal{L}_{L/G} = -\log \mathbb{P}_{L/G}(S|I) - \log \mathbb{P}_{L/G}(I|S) & (5) \\ \mathcal{L}_{\text{SGSM}} = \mathcal{L}_L + \mathcal{L}_G & (6) \end{cases}$$

Empirically, the SGSM resulted in large gains in performance as shown in Table 4. Our hypothesis is that the scene graph, in a similar way to a caption, provides easier, simpler to distil relational information contained in the layout, which results in stronger performance compared to generation using just the layout. Architectural details of the SGSM and related data processing are described in the Appendix.

Instance-Aware Conditioning

As in Park et al. (2019); Sun and Wu (2019), the parameters γ, β of our batch-normalization layers are *conditional* and determined on a per-pixel level (as opposed to classical conditional batch-normalization, De Vries et al. 2017). In our case, these parameters are determined by three concatenated inputs: *masked object embeddings, bounding-box layouts and bounding-box instance boundaries*. Masked object embeddings (Ma et al. 2018; Sun and Wu 2019) and bounding-box layouts (using 1-hot embeddings) have been previously used in the layout to image setting. A shortcoming of these conditioning inputs is that they do not provide any way to distinguish between objects of the same class if

their bounding boxes overlap. We use the layout's bounding-box boundaries, shown in Figure 4, as additional conditioning information. The addition of the bounding-box instance boundaries helps the model in mapping overlapping conditioning semantic masks to separate object instances, the absence of which led previous state-of-the-art methods to generate merged outputs as shown in the donut example in Fig. 1. Importantly, the instance boundaries do not add any additional information compared to the baselines: (1) they are bounding-box rather than fine-grained boundaries, and (2) instance information is already available to other models (Layout2Im and LostGAN have object-specific codes as an example). Rather, adding these boundaries acts like a prior encouraging our model to focus on generating distinct objects.

Architecture

Our OC-GAN model is based on the GAN framework. The generator module generates the images conditioned on the ground-truth layout. The discriminator predicts whether the input image is generated or real. The discriminator has an additional component which has to discriminate objects present in the input image patches corresponding to the ground-truth layout object bounding boxes. We present an overview of the model in Fig. 5 and describe the components below. Additional details are in the Appendix.

Generator As a means of disentangling our model's performance from a specific choice of generator architecture, we used a classical residual (He et al. 2016) architecture consisting of 4 layers for 64×64 inputs, and 5 layers for 128×128 and 256×256 inputs, as used recently in Park et al. (2019); Sun and Wu (2019); Wang et al. (2018). The residual decoder G takes as input image-level noise. As described in the previous section, we further condition the generation by making the normalization parameters of the batch-norm layers of the decoder dependent on the layout and instance boundaries.

Discriminator We use two different types of discriminators, an object discriminator, and a set of patch-wise discriminators. The object discriminator D_{obj} takes as input crops of the objects (as identified by their input bounding boxes) in real and fake images resized to size 32×32 and is trained using the Auxiliary-Classifier (AC, Odena, Olah, and Shlens 2017) framework, resulting in a classification and an adversarial loss. Next, two *patch-wise discriminators* D_1^p, D_2^p output estimates of whether a given patch is consistent with the input layout. We apply them to the original image and the same image down-sampled by a factor of 2 (no weight sharing) in a similar fashion to Park et al. (2019); Wang et al. (2018).

Loss Functions

In the following, x denotes a real image, l a layout, and z noise. We also denote objects with o and their labels y_o.

Perceptual loss Adding a perceptual loss (Dosovitskiy and Brox 2016; Gatys, Ecker, and Bethge 2016; Johnson, Alahi, and Fei-Fei 2016) to our model improved results slightly. We extract features using a VGG19 network (Simonyan and Zisserman 2015). The loss has expression: $\mathcal{L}_P = \mathbb{E}_{x,l,z} \sum_{i=1}^{N} \frac{1}{D_i} ||F^{(i)}(x) - F^{(i)}(G(l,z))||_1$ where $F^{(i)}$ extracts the output at the i-th layer of the VGG and D_i is the dimension of the flattened output at the i-th layer.

Generator and Discriminator losses We train the generator and patch discriminators using the adversarial hinge loss (Lim and Ye 2017):

$$\mathcal{L}_G^{\text{GAN}} = -\mathbb{E}_{l,z}\Big[D_1^p(G(l,z),l) + D_2^p(G(l,z),l)\Big] \tag{7}$$

$$\begin{aligned}\mathcal{L}_{D^p} = \sum_{i=1}^{2} \Big\{ &- \mathbb{E}_{x,l}\Big[\min(0, -1 + D_i^p(x,l))\Big] \\ &- \mathbb{E}_{l,z}\Big[\min(0, -1 - D_i^p(G(l,z),l)\Big]\Big\}\end{aligned} \tag{8}$$

The object discriminator follows the AC-GAN framework, leading to $\mathcal{L}_G^{AC}$ and $\mathcal{L}_{D_{obj}}^{AC}$. The final expression is:

$$\mathcal{L}_G = \mathcal{L}_G^{\text{GAN}} + \lambda_P \mathcal{L}_{\text{P}} + \lambda_{\text{SGSM}} \mathcal{L}_{\text{SGSM}} + \lambda_{\text{AC}} \mathcal{L}_G^{AC} \tag{9}$$

$$\mathcal{L}_D = \mathcal{L}_{D^p} + \lambda_o \mathcal{L}_{D_{obj}}^{AC} \tag{10}$$

We fix $\lambda_P = 2, \lambda_o = 1, \lambda_{\text{SGSM}} = 1, \lambda_{\text{AC}} = 1$ in our experiments.

Experiments

Datasets

We run experiments on the COCO-Stuff (Caesar, Uijlings, and Ferrari 2018) and Visual Genome (VG) (Krishna et al. 2017) datasets which have been the popular choice for layout- and scene-to-image tasks as they provide diverse and high-quality annotations. The former is an expansion of the Microsoft Common Objects in Context (MS-COCO) dataset (Lin et al. 2014). We apply the same pre-processing and use the same splits as Johnson, Gupta, and Fei-Fei (2018); Zhao et al. (2019). The summary statistics of the two datasets are presented in the appendix, Table 2.

Our OC-GAN model takes three different inputs:

- The spatial layout *i.e.* object bounding boxes and object class annotations.
- Instance boundary maps computed directly from the layout. While they appear redundant once the bounding boxes are provided, they aid the model in better differentiating different objects especially different instances of the same object class.
- Scene-graphs. These are constructed from the objects and spatial relations inferred from the bounding box positions following the setup in (Johnson, Gupta, and Fei-Fei 2018). While VG provides more complex scene graphs, we restricted ourselves to spatial relations only for compatibility between the two datasets.

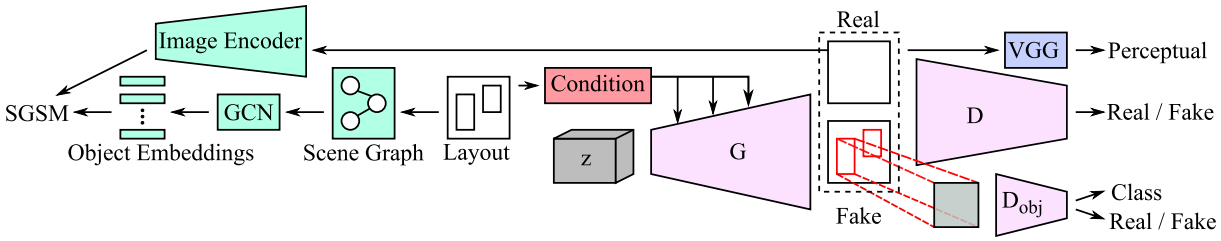

Figure 5: Overview of our OC-GAN model. The GCN and Image Encoder modules are trained separately and then frozen. The condition for the Generator's normalization and the Scene Graph encoding the spatial relationships between objects are both derived from the input layout. The SGSM and the instance-aware normalization lead our model to generate images with higher layout-fidelity and sharper, distinct objects. The 'Condition' box corresponds to the three inputs listed in the subsection on the instance-aware conditioning.

Implementation and Training Details

Our code is written in PyTorch (Paszke et al. 2019). We apply Spectral Normalization (Miyato et al. 2018) to all the layers in both the generator and discriminator networks. Each experiment ran on 4 V100 GPUs in parallel. We use synchronized BatchNorm (all summary statistics are shared across GPUs).

We used the Adam (Kingma and Ba 2015) solver, with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The global learning rate for both generator and discriminators is 0.0001. 128×128 models and above were trained for up to 300 000 iterations, 64×64 models were trained for up to 200 000 iterations (early stopping on a validation set). The SGSM module is trained separately for 200 epochs. It is then fixed, and the rest of the model is trained.

Baselines

We consider all recent methods that allow layout-to-image generation (Layout2Im (Zhao et al. 2019), LostGAN (Sun and Wu 2019), LostGAN-v2 (Sun and Wu 2020)). We report results for scene-graph-to-image methods (SG2Im (Johnson, Gupta, and Fei-Fei 2018), SOARISG (Ashual and Wolf 2019)) evaluated with *ground-truth layouts* for a fair comparison. Finally, methods originally designed for generation from pixel-level semantic segmentation maps (SPADE (Park et al. 2019) and Pix2PixHD (Wang et al. 2018)) are also considered as they can be readily adapted to this new context.

Evaluation

Evaluation of GANs is a complex issue, and the subject of a vast body of literature. In this paper, we focus on three existing evaluation metrics: Inception Score (IS) (Salimans et al. 2016), Fréchet Inception Distance (FID) (Heusel et al. 2017) and Classification Accuracy (CA). For the CA score, a ResNet-101 (He et al. 2016) network is trained on object crops obtained from the real images of the train set of the corresponding dataset, as suggested by (Ashual and Wolf 2019). The FID metric computes the 2-Wasserstein distance between the real and generated distributions, and therefore serves as an efficient proxy for the diversity and visual quality of the generated samples. While the FID metric focuses on the whole image, the CA metric allows us to demonstrate the ability of our model to generate realistic-looking objects within a scene. Finally, we include the Inception Score as a legacy metric.

Our proposed metric: SceneFID We note that there exist many concerns in the literature regarding the use of metrics that are not designed or adapted to the task at hand. The Inception Score has been criticised (Barratt and Sharma 2018), notably due to issues caused by the mismatch between the domain it was trained on (the ImageNet dataset comprising single objects of interest) and the domain of VG and COCO-Stuff images (comprising multiple objects in complex scenes), making it a potentially poor metric to evaluate generative ability of models in our setting. While the FID metric was introduced in response to Inception Score's criticisms, and was shown empirically to alleviate some of the concerns with it (Im et al. 2018; Xu et al. 2018a; Lucic et al. 2018), it still suffers from problems in the layout-to-image setting. In particular, the single manifold assumption behind FID was found in Liu et al. (2018) to be problematic in a multi-class setting. This is *a fortiori* the case in a multi-object setting as in VG and COCO. While (Liu et al. 2018) introduce a class-aware version of FID, this is not applicable to our setting. We introduce the *SceneFID* metric, where we compute the FID on the crops of all objects, resized to same size (224×224), instead of on the whole image. Thus, the SceneFID metric measures FID in the single manifold assumption it was designed for and extends it to the multi-object setting.

In addition to the above quantitative metrics, we also perform qualitative assessment of the model, notably by considering the effect of modifying the input layout on the output image.

Quantitative Results

We report comparisons of our model's performance to the set of all recent state-of-the-art methods. Where applicable and possible, we use metric values reported by the authors of the papers. SOARISG (Ashual and Wolf 2019) depends on semantic segmentation maps being available, and therefore it was not feasible to include results on VG for this method. Some papers introduced additional data-augmentation, such as LostGAN (Sun and Wu 2019) which introduced flips of the real images during training. Where applicable, we report results using the same experimental setup as the authors, and

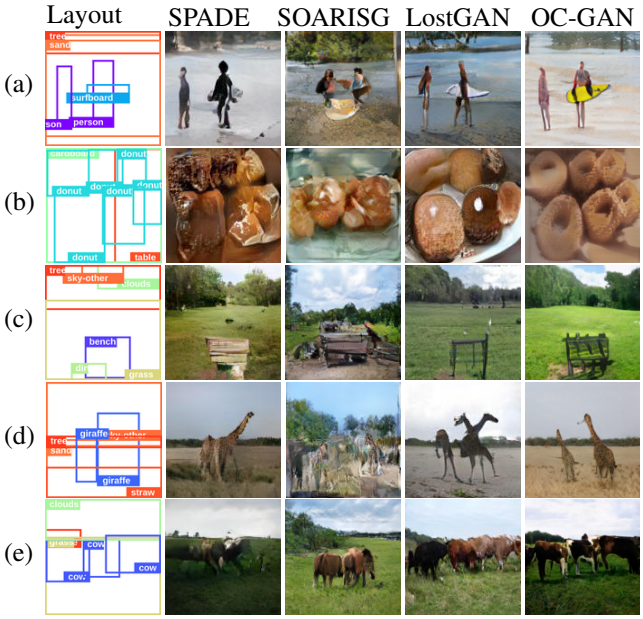

Figure 6: 128×128 COCO-Stuff test set images, taken from our method (OC-GAN), and multiple competitive baselines. Note the overall improved visual quality of our samples. In addition, for (d, e) many baselines introduce spurious objects, and for (b, d, e) spatially close objects are poorly defined and sometimes fused for the baselines.

highlight it in the results table. For all models that do not report CA scores, we evaluate them using images generated with the pre-trained models provided by their authors.

Table 1 shows that our model consistently outperforms the baselines in terms of IS, FID and CAS, often significantly. We note that for some models, the CAS score is above that reported for ground-truth images. This is due to the fact that a sufficiently capable generator will start to generate objects that are both realistic, and of the same distribution as the training distribution, rather than the test one.

On the proposed SceneFID metric, Table 2 shows that our method outperforms the others significantly. Thus, our model is significantly better at generating realistic objects compared to the baselines. Note that the LostGAN model obtains better FID compared to our model exceptionally on 128×128 COCO-Stuff images but our OC-GAN model outperforms it on the SceneFID metric which is more appropriate in this multi-class setting.

Qualitative Results

We compare and analyse image samples generated by our method and competitive baselines in Fig. 6. In addition to generating higher quality images, our OC-GAN model does not introduce spurious objects (objects not specified in the layout but present in the generated image). This can be attributed to the SGSM module which, by virtue of the retrieval task and the scene-graph being a higher-level abstraction than pixels, aids the model in learning a better mapping from the spatial layout to the generated image. Our model also keeps object instances identifiable even when bounding boxes of objects of the same class overlap slightly or are in close proximity.

To further validate the previous observations, in Fig. 1, we consider the effect of generating from artificial layouts of gradually converging donuts, to tease out the model's ability to correctly generate separable object instances. Our model generates distinct donuts even when occluded, whereas the other models generate realistic donuts when the bounding boxes are far apart, but fail to do so when they overlap.

We also conducted a user study to evaluate the model's layout-fidelity. 10 users were shown 100 layouts from the test sets of both datasets, with the corresponding images generated by our OC-GAN, LostGAN, and for COCO-Stuff, SOARISG, shuffled in a random order. For each layout, users were asked to select the model which generates the best corresponding image. The results from this study are in Table 3 and demonstrate that our model has higher layout-fidelity than previous SOTA methods.

In Table 4, we present an ablation study performed by removing certain components of our model. The effect of adding another patch discriminator is measurable, both in terms of FID and CA. Removing the patch discriminator significantly lowers FID (the model has no more supervision in terms of matching the distribution of the real full images. This actually improves the CA, as the generator will use more capacity to focus on generating realistic objects.

We also find that removing either the object discriminator or the SGSM results in a significant drop in performance. This does not however prevent the model from generating realistic objects (the CA score remains above some of the baselines), meaning that the roles of the two components are to some extent complementary. As soon as both are removed, the CA score drops sharply.

Removing the perceptual loss has little effect in itself, but it greatly helps the SGSM when present. Removing the SGSM altogether strongly impairs results, highlighting its importance. Finally, removing the bounding-box instance boundaries has a modest impact on both metrics, but a large qualitative impact with more clearly defined objects.

Conclusion

We observed that current state-of-the-art layout-to-image generation methods exhibit low layout-fidelity and tend to generate low quality objects especially in cases of occlusion. We proposed a novel Scene-Graph Similarity Module that mitigated the layout-fidelity issues aided by an improved understanding of spatial relationships derived from the layout. We also proposed to condition the generator's normalization layers on instance boundaries which led to sharper, more distinct objects compared to other approaches. The addition of the proposed components to the image generation pipeline led to our model outperforming previous state-of-the-art approaches on a variety of quantitative metrics. A comprehensive ablation study was performed to analyse the contribution of the proposed and existing components of the model. Human users also rated our approach higher on generating better-suited images for the layout over existing methods.

Evaluation metrics for GAN popularized in the single-object-class setting have been criticized as inappropriate in

	Methods	Inception Score ↑		FID ↓		CA ↑	
		COCO	VG	COCO	VG	COCO	VG
Real Images	64×64	16.3 ± 0.4	13.9 ± 0.5	0	0	54.48	49.57
	128×128	22.3 ± 0.5	20.5 ± 1.5	0	0	60.71	56.25
	256×256	28.10 ± 0.5	28.6 ± 1.2	0	0	63.04	60.40
64×64	SG2Im (Johnson, Gupta, and Fei-Fei 2018)†	7.3 ± 0.1	6.3 ± 0.2	67.96	74.61	30.04	40.29
	Pix2PixHD (Wang et al. 2018)	7.2 ± 0.2	6.6 ± 0.3	59.95	47.71	20.82	16.98
	SPADE (Park et al. 2019)	8.5 ± 0.3	7.3 ± 0.1	43.31	35.74	31.61	23.81
	Layout2Im (Zhao et al. 2019)†	9.1 ± 0.1	8.1 ± 0.1	38.14	31.25	50.84	48.09
	SOARISG (Ashual and Wolf 2019)∗ †	10.3 ± 0.1	N/A	48.7	N/A	46.1	N/A
	OC-GAN (ours)	$\mathbf{10.5 \pm 0.3}$	$\mathbf{8.9 \pm 0.3}$	**33.1**	**22.61**	**56.88**	**57.73**
64×64 with flips	LostGAN (Sun and Wu 2019) (flips) †	9.8 ± 0.2	8.7 ± 0.4	34.31	34.75	37.15	27.1
	OC-GAN (ours)	$\mathbf{10.8 \pm 0.5}$	$\mathbf{9.3 \pm 0.2}$	**29.57**	**20.27**	**60.39**	**60.79**
128×128	Pix2PixHD (Wang et al. 2018)	10.4 ± 0.3	9.8 ± 0.3	62	46.55	26.67	25.03
	SPADE (Park et al. 2019)	13.1 ± 0.5	11.3 ± 0.4	40.04	33.29	41.74	34.11
	Layout2Im (Zhao et al. 2019) ⋄	12.0 ± 0.4	10.1 ± 0.3	43.21	38.21	49.06	51.13
	SOARISG (Ashual and Wolf 2019) †∗	12.5 ± 0.3	N/A	59.5	N/A	44.6	N/A
	OC-GAN (ours)	$\mathbf{14.0 \pm 0.2}$	$\mathbf{11.9 \pm 0.5}$	**36.04**	**28.91**	**60.32**	**58.03**
128×128 with flips	LostGAN (Sun and Wu 2019) †	13.8 ± 0.4	11.1 ± 0.6	29.65	29.36	41.38	28.76
	LostGAN-V2 (Sun and Wu 2020) †	14.2 ± 0.4	10.71 ± 0.27	**24.76**	29.00	43.27	35.17
	OC-GAN (ours)	$\mathbf{14.6 \pm 0.4}$	$\mathbf{12.3 \pm 0.4}$	36.31	**28.26**	**59.44**	**59.40**
256×256	SOARISG (Ashual and Wolf 2019) †∗	15.2 ± 0.1	N/A	65.95	N/A	45.3	N/A
	OC-GAN (ours)	$\mathbf{17.0 \pm 0.1}$	14.4 ± 0.6	**45.96**	39.07	**53.47**	57.89
256×256 with flips	LostGAN-V2 (Sun and Wu 2020) †	$\mathbf{18.0 \pm 0.5}$	14.1 ± 0.4	42.55	47.62	54.40	53.02
	OC-GAN (ours)	17.8 ± 0.2	$\mathbf{14.7 \pm 0.2}$	**41.65**	**40.85**	**57.16**	**53.28**

Table 1: Performance on 64, 128 and 256 dimension images. All models use ground-truth layouts. We use † to denote results taken from the original paper. ∗ denotes a model that uses pixel-level semantic segmentation during training. ⋄ denotes models for which the openly available source code was not adapted to generation at a specific image size. We altered the code to allow this and ran a hyperparameter search on the new models.

	SceneFID ↓	
Methods	COCO	VG
Pix2PixHD (Wang et al. 2018)	42.92	42.98
SPADE (Park et al. 2019)	23.44	16.72
Layout2Im (Zhao et al. 2019)	22.76	12.56
SOARISG (Ashual and Wolf 2019)∗	33.46	N/A
LostGAN (Sun and Wu 2019) (flips)	20.03	13.17
OC-GAN (ours w/ flips)	**16.76**	**9.63**

Table 2: SceneFID scores on object crops resized to size 224 × 224, extracted from the 128 × 128 outputs of the different models, for both datasets. All models use ground-truth layouts. ∗ denotes a model that uses pixel-level semantic segmentation during training. SOARISG cannot be trained on VG due to the absence of pixel-level semantic segmentations.

Dataset	SOARISG	LostGAN	Ours
COCO-Stuff	16.8%	36.8%	**46.4%**
VG	N/R	31.4%	**68.6%**

Table 3: User study results. 10 computer-science professionals were shown 100 COCO-Stuff and 100 VG test set layouts and corresponding images generated by various models, shuffled randomly. Users were asked to select the highest layout-fidelity image for each layout at 128×128 resolution. SOARISG is marked marked non-rated (N/R), as it cannot be trained on VG.

	FID ↓	CA ↑
Full	**29.57**	60.27
Single patchD	30.54	59.86
No patchD	33.85	**62.48**
No objectD	31.62	48.03
No bounding-box instance boundaries	30.12	59.54
No SGSM	34.32	52.57
No objectD, no SGSM	33.15	41.50
No perceptual loss	31.14	57.22
No perceptual loss, no SGSM	36.54	47.94

Table 4: Quantitative comparison of different ablated versions of our model on the COCO-Stuff dataset (64×64 images). These results highlight the importance of the SGSM (and its positive interaction with the perceptual loss) in the bottom row block, as well as the impact of removing some of the discriminators (middle row block).

the multi-class setting in literature. Our proposed SceneFID metric addresses those concerns and presents a useful metric for the image generation community which will increasingly deal with multi-class settings in the future. Our proposed OC-GAN model also showed a large improvement over existing approaches on the SceneFID evaluation criteria which further highlights the impact of our contributions.

Acknowledgments

We acknowledge Emery Fine, Adam Ferguson, Philip Bachman and Hannes Schulz for their insightful suggestions and valuable assistance. We also thank the many researchers who contributed to the human evaluation study.

Appendix

The Appendix can be found in the arXiv version of this paper located at https://arxiv.org/abs/2003.07449.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*, 382–398.

Ashual, O.; and Wolf, L. 2019. Specifying Object Attributes and Relations in Interactive Scene Generation. In *ICCV*.

Barratt, S.; and Sharma, R. 2018. A note on the inception score. *arXiv preprint arXiv:1801.01973* .

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. COCO-Stuff: Thing and Stuff Classes in Context. In *CVPR*.

De Vries, H.; Strub, F.; Mary, J.; Larochelle, H.; Pietquin, O.; and Courville, A. C. 2017. Modulating early visual processing by language. In *NIPS*, 6594–6604.

Dosovitskiy, A.; and Brox, T. 2016. Generating images with perceptual similarity metrics based on deep networks. In *NIPS*, 658–666.

El-Nouby, A.; Sharma, S.; Schulz, H.; Hjelm, D.; El Asri, L.; Ebrahimi Kahou, S.; Bengio, Y.; and Taylor, G. W. 2019. Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. In *ICCV*.

Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R. K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J. C.; et al. 2015. From captions to visual concepts and back. In *CVPR*, 1473–1482.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *CVPR*, 2414–2423.

Goller, C.; and Kuchler, A. 1996. Learning task-dependent distributed representations by backpropagation through structure. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, 347–352.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*, 2672–2680. Curran Associates, Inc.

Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; and Ling, M. 2019. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 1969–1978.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved Training of Wasserstein GANs. In *NIPS*, 5767–5777. Curran Associates, Inc.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 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 *NIPS*, 6626–6637.

Huang, P.-S.; He, X.; Gao, J.; Deng, L.; Acero, A.; and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2333–2338.

Im, D. J.; Ma, A. H.; Taylor, G. W.; and Branson, K. 2018. Quantitatively Evaluating GANs With Divergences Proposed for Training. In *ICLR*.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-To-Image Translation With Conditional Adversarial Networks. In *CVPR*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 694–711.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image Generation From Scene Graphs. In *CVPR*.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *CVPR*, 3668–3678.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. *IJCV* .

Li, W.; Zhang, P.; Zhang, L.; Huang, Q.; He, X.; Lyu, S.; and Gao, J. 2019a. Object-Driven Text-To-Image Synthesis via Adversarial Training. In *CVPR*.

Li, Y.; Gan, Z.; Shen, Y.; Liu, J.; Cheng, Y.; Wu, Y.; Carin, L.; Carlson, D.; and Gao, J. 2019b. StoryGAN: A Sequential Conditional GAN for Story Visualization. In *CVPR*.

Li, Y.; Ma, T.; Bai, Y.; Duan, N.; Wei, S.; and Wang, X. 2019c. PasteGAN: A Semi-Parametric Method to Generate Image from Scene Graph. In *NeurIPS*, 3948–3958. Curran Associates, Inc.

Lim, J. H.; and Ye, J. C. 2017. Geometric gan. *arXiv preprint arXiv:1705.02894* .

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context". In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *ECCV*, 740–755. ISBN 978-3-319-10602-1.

Liu, S.; Wei, Y.; Lu, J.; and Zhou, J. 2018. An improved evaluation framework for generative adversarial networks. *arXiv preprint arXiv:1803.07474* .

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *ECCV*, 852–869. Springer.

Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; and Bousquet, O. 2018. Are gans created equal? a large-scale study. In *NeurIPS*, 700–709.

Ma, L.; Jia, X.; Georgoulis, S.; Tuytelaars, T.; and Van Gool, L. 2018. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145* .

Mittal, G.; Agrawal, S.; Agarwal, A.; Mehta, S.; and Marwah, T. 2019. Interactive Image Generation Using Scene Graphs. In *ICLR: Deep Generative Models for Highly Structured Data Workshop*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *ICLR*.

Miyato, T.; and Koyama, M. 2018. cGANs with Projection Discriminator. In *ICLR*.

Newell, A.; and Deng, J. 2017. Pixels to graphs by associative embedding. In *NIPS*, 2171–2180.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2642–2651.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *CVPR*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NIPS*, 8024–8035. Curran Associates, Inc.

Qi, X.; Chen, Q.; Jia, J.; and Koltun, V. 2018. Semi-parametric image synthesis. In *CVPR*, 8808–8816.

Quinn, M. H.; Conser, E.; Witte, J. M.; and Mitchell, M. 2018. Semantic image retrieval via active grounding of visual situations. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, 172–179.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *ICLR*.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative Adversarial Text to Image Synthesis. In *ICML*, volume 48 of *Proceedings of Machine Learning Research*, 1060–1069.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *NIPS*, 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.

Sharma, S.; Suhubdy, D.; Michalski, V.; Kahou, S. E.; and Bengio, Y. 2018. ChatPainter: Improving Text to Image Generation using Dialogue. In *ICLR Workshop*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Singh, K. K.; Ojha, U.; and Lee, Y. J. 2019. FineGAN: Unsupervised Hierarchical Disentanglement for Fine-Grained Object Generation and Discovery. In *CVPR*.

Sun, W.; and Wu, T. 2019. Image Synthesis From Reconfigurable Layout and Style. In *ICCV*.

Sun, W.; and Wu, T. 2020. Learning Layout and Style Reconfigurable GANs for Controllable Image Synthesis. *arXiv preprint arXiv:2003.11571* .

Sylvain, T.; Petrini, L.; and Hjelm, D. 2020a. Locality and Compositionality in Zero-Shot Learning. In *ICLR*.

Sylvain, T.; Petrini, L.; and Hjelm, R. D. 2020b. Zero-Shot Learning from scratch (ZFS): leveraging local compositional representations. *arXiv preprint arXiv:2010.13320* .

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Teney, D.; Liu, L.; and van Den Hengel, A. 2017. Graph-structured representations for visual question answering. In *CVPR*, 1–9.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* .

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *CVPR*.

Xu, Q.; Huang, G.; Yuan, Y.; Guo, C.; Sun, Y.; Wu, F.; and Weinberger, K. 2018a. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755* .

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018b. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *CVPR*.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*, 10685–10694.

Yin, G.; Liu, B.; Sheng, L.; Yu, N.; Wang, X.; and Shao, J. 2019. Semantics Disentangling for Text-To-Image Generation. In *CVPR*.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2019. Self-Attention Generative Adversarial Networks. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 7354–7363.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks. In *ICCV*.

Zhao, B.; Meng, L.; Yin, W.; and Sigal, L. 2019. Image Generation From Layout. In *CVPR*.