Modeling Image Composition for Complex Scene Generation

Zuopeng Yang[1*] Daqing Liu[2*] Chaoyue Wang[3] Jie Yang[1†] Dacheng Tao[2,3]
[1]Shanghai JiaoTong University [2]JD Explore Academy, JD.com [3] The University of Sydney
{yzpeng, jieyang}@sjtu.edu.cn, chaoyue.wang@outlook.com, {liudq.ustc, dacheng.tao}@gmail.com

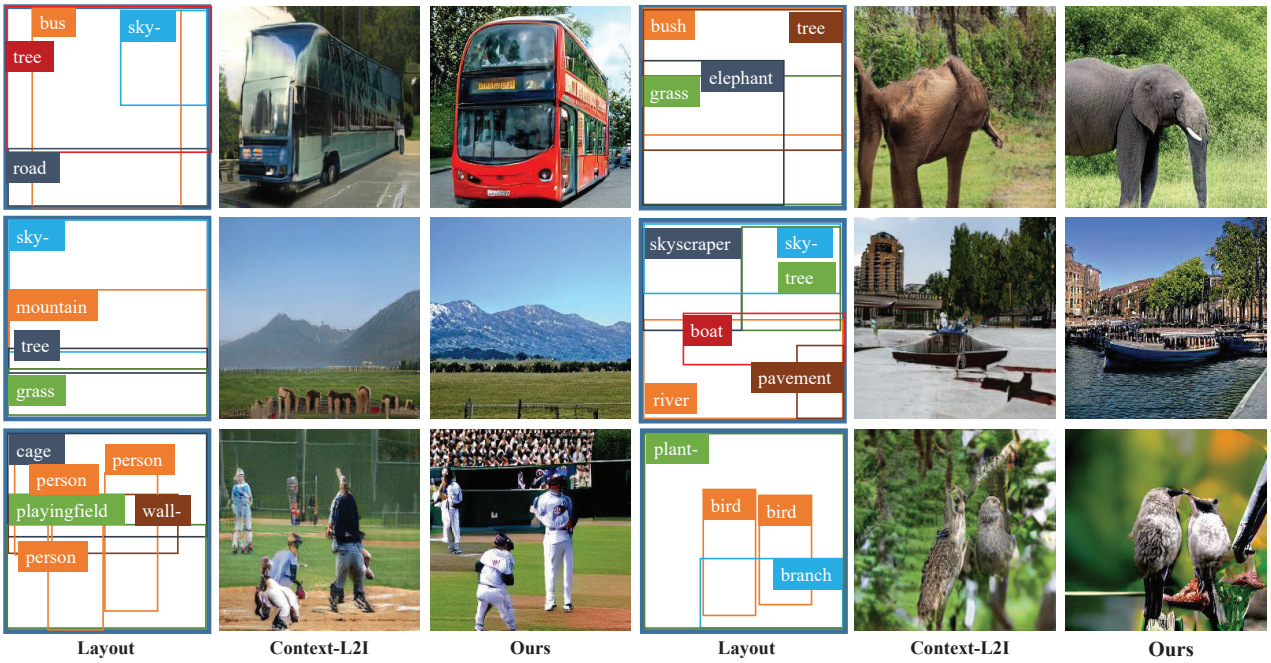

Figure 1. By specifying the layouts, we compared several results generated by recent CNN-based method [10] and our proposed TwFA. Our approach enables transformers to synthesize high-quality images containing **multiple objects with complex structures** from layouts (bounding boxes with categories).

Abstract

We present a method that achieves state-of-the-art results on challenging (few-shot) layout-to-image generation tasks by accurately modeling textures, structures and relationships contained in a complex scene. After compressing RGB images into patch tokens, we propose the Transformer with Focal Attention (TwFA) for exploring dependencies of object-to-object, object-to-patch and patch-to-patch. Compared to existing CNN-based and Transformer-based generation models that entangled modeling on pixel-level&patch-level and object-level&patch-level respectively, the proposed focal attention predicts the current patch token by only focusing on its highly-related tokens that specified by the spatial layout, thereby achieving disambiguation during training. Furthermore, the proposed TwFA largely increases the data efficiency during training, therefore we propose the first few-shot complex scene generation strategy based on the well-trained TwFA. Comprehensive experiments show the superiority of our method, which significantly increases both quantitative metrics and qualitative visual realism with respect to state-of-the-art CNN-based and transformer-based methods. Code is available at https://github.com/JohnDreamer/TwFA.

1. Introduction

Generating photo-realistic images is the ever-lasting goal in computer vision. Despite achieving remarkable progress on image generation for both simple scenario, *e.g.*, faces,

*Equal Contribution. †Corresponding author.

This research is partly supported by NSFC, China (No: 61876107, U1803261), and Dr. Chaoyue Wang is supported by ARC FL-170100117.

cars, and cats [14, 15, 30], and single object, *e.g.*, ImageNet [1, 42], the image generation for complex scenes composed of multiple objects of various categories is still a challenging problem.

In this paper, we focus on one representative complex scene image generation task, layout to image generation [45] (L2I), which aims to generate complex scenes conditioned on specified layouts. The layout, as illustrated in Figure 1, consists of a set of object bounding boxes and corresponding categories, thus providing a sketch of the expected complex scene image. Compared with other conditions for complex scene generation, including textual descriptions [27], scene graphs [13, 23], and segmentation masks [20], layouts are much more user-friendly, controllable and flexible [45]. Ambitiously, we further propose a new few-shot layout to image generation task (few-shot L2I), which aims to generate complex scenes with a novel object category after providing only a few images containing the novel objects.

As to the complex scene generation, including (few-shot) L2I tasks, the core challenge is how to synthesize a photo-realistic image with reasonable object-level relationships, clear patch-level instance structures, and refined pixel-level textures. Existing attempts to the L2I task can be divided into two categories, *i.e.*, CNN-based [18, 20, 24, 33, 34, 46] and Transformer-based [12], according to their generator. The CNN-based methods deploy an encoder-decoder generator [13, 24] where the encoder transfers the layout into an image feature map, and the decoder upsamples the feature map into the target image. Those methods capture the object relationships in the encoder by a self-attention [35] or a convLSTM [46], and model the instance structures and textures simultaneously in the decoder by upsampling convolutions. In contrast, the Transformer-based methods tokenize the layout into object tokens and employ a pre-trained compression model to quantize the image into a sequence of discrete patch tokens, thus simplifying the image generation task as an image patch composition task implemented by a Transformer. Those methods produce the detailed textures with the compression model, and model both relationships and structures by the Transformer.

However, the entangled modeling on patch-level and pixel-level (CNN-based methods) or object-level and patch-level (Transformer-based methods) prevents the model from capturing inherent instance structures, leading to blurry or crumpled objects, and increases the burden on the few-shot learning because the model must learn the two levels information simultaneously with only a few images. To this end, upon the Transformer-based methods, we propose a Transformer with Focal Attention (TwFA) to separately model image compositions on object-level and patch-level by distinguishing between object and patch tokens. Different from vanilla self-attention, which neglects the composition prior of spatial layouts, our focal attention further constrains each token can only attend on its related tokens according to the spatial layouts. Specifically, to model object relationships, an object token attends on all object tokens to capture the global information. To model instance structures, a patch token attends on the object it belongs to and the patches inner the object bounding box. By the proposed Focal Attention, the TwFA focuses on generating the current patch without any disturbance from other objects or patches thus increasing the data efficiency during training. Therefore, the focal attention makes the TwFA can fast learn the novel object category with only a few images.

We validate the effectiveness of the proposed TwFA on COCO-stuff [2, 22] and Visual Genome [17] datasets. TwFA improves the state-of-the-arts [12,20] FID score from 29.56 to 22.15 (-25.1%) on COCO-stuff, and from 19.14 to 17.74 (-7.3%) on Visual Genome. Morevoer, TwFA demonstrates the superiority on the few-shot L2I task with strong performance and impressive visualizations.

2. Related Work

CNN-based image generation. In recent years, a number of CNN-based generative models have been proposed, and achieved significant progress on (un)conditional image generation tasks. Till now, CNN-based generative models (*e.g.*, GANs [6, 36, 37], VAEs [16]) are good at synthesizing high-resolution and high-fidelity object images, which include but are not limited to flowers, human/animal faces, and buildings [1, 3, 14]. However, generating complex real-world scenes which include multiple instances with variant layout and scale has still been a challenging task [10,20,32]. To ease the difficulty of synthesizing complex scenes, previous works usually break the tasks into several steps. For example, Layout2Im [45,46] models this task as object generation then image generation pipeline, each object is controlled by a certain category code and an uncertain appearance code. For better controllability, LostGANs [32, 33] first synthesize the semantic masks from layouts, and then ISLA-Norm is proposed for generating color images from specific masks and style codes. In addition, some works focus on improving models' generative performance by introducing pseudo supervisions [18, 34], additional annotations [24] or fine-grained control [5]. Although CNN-based layout-to-image generation methods have achieved promising performance on texture synthesis, they may still suffer from accurately modeling the dependencies among pixels (or object parts), which hinders the model generated more realistic scene images.

Transformer-based image generation. Recently, transformers not only demonstrate promising results in computer vision tasks [41, 43], but also show potential on conditional visual content generation [4, 21, 26]. First, a Vector Quantised Generative Adversarial Network (VQ-GAN) [4]

is trained to compress images into finite discrete representations/tokens. Then, an autoregressive transformer is trained to model the dependencies between discrete image tokens. Through modeling together with conditional signals, such as text, class labels and keypoints, transformers demonstrated the strong capability of generating semantic controllable images [4]. For the complex scene generation, [12] made a great attempt on synthesizing the high-resolution image from a given layout. Although promising results have been achieved, synthesizing complex scenes that consist of multiple instances and staff is still a challenge task. Since autoregressive transformers have challenges in handling spatial positions [40], they may not accurately model object-object and object-patch relationships. In this paper, the proposed transformer with focal attention can better model the composition of the complex scenes, and leads to better generation performance.

Few-shot image generation. Few-shot learning is first explored in discriminative tasks. Given limited data from a target domain, neural networks have to overcome training/fine-tuning difficulties, and generalize the pre-trained model to target domain [8, 9, 39]. In few-shot image generation, the generative model is trained to synthesize diverse images from the target domain. Previous few-shot image generation methods mainly focused on generating simple patterns and low resolution results [28, 29]. Recently, [7, 19, 25, 38] extend the few-shot generation to objects with similar structures, such as human faces, buildings and cars. However, it has great challenges in performing few-shot generation on complex scenes or novel classes in complex scenes. To our best knowledge, this paper is the first attempt on few-shot complex scene generation.

3. Approach

3.1. The Framework

As illustrated in Figure 2, the proposed Transformer with Focal Attention (TwFA) for the L2I task generally follows the pipeline of tokenization → composition → generation, *i.e.*, firstly tokenizes the layout/image into sequential discrete object/patch tokens, secondly predicts the distribution of patch tokens in an auto-regressive manner and composes into a discrete patch token sequence, and finally generates the synthesized image from the patch tokens.

Tokenization. Given a layout consists of a set of objects with their bounding boxes and category classes, we directly tokenize it into a sequential object tokens $\boldsymbol{c} = \{(l_i, \boldsymbol{b}_i)_{i=1}^{N}\}$ with N objects, where l_i denotes the i^{th} object's category, and $b_i = [x_{1i}, y_{1i}, x_{2i}, y_{2i}]$ represents its top-left and bottom right corner positions.

Given an image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$, we tokenize it with an encoder of Vector Quantised Generative Adversarial Network (VQ-GAN [4]) that compresses high-dimensional

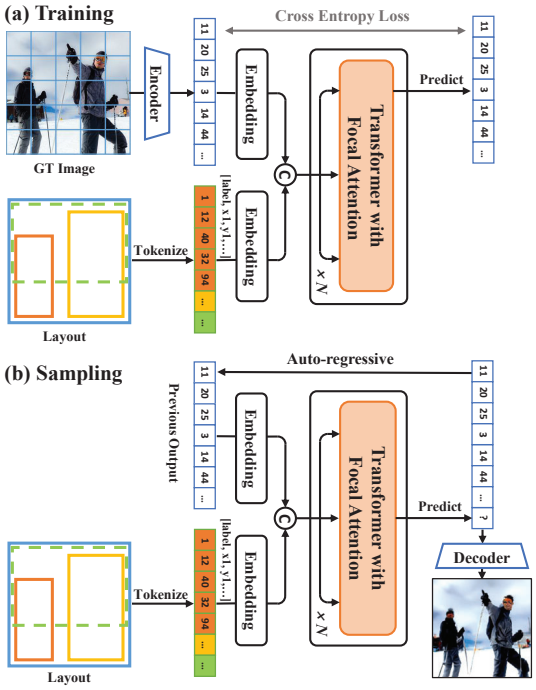

Figure 2. The overview of the proposed Transformer with Focal Attention (TwFA) framework for the L2I task. Given a layout and an image as input, we first 1) tokenize them into sequential discrete object/patch tokens by the embedding/encoder, next 2) predict the next patch token by the TwFA, and then 3) during inference, generate the RGB image from all predicted patch tokens by the decoder.

data into a discretized latent space and reconstructs it. Specifically, the VQ-GAN encoder Enc tokenizes the image $\boldsymbol{x}$ into a collection of indices $\boldsymbol{s}$ of codebook entries:

$$\boldsymbol{s} = \{s_1, s_2, \ldots, s_M\} = Enc(\boldsymbol{x}). \tag{1}$$

Here, the codebook actually is a "vocabulary" of learned representations $C = \{\boldsymbol{e}_i\}_{i=1}^{K}$ where the vocabulary size is K. The VQ-GAN decoder Dec then tries to reconstruct the original image from these latent codes. In our method, we use a pretrained generic VQ-GAN [4] and keep the weights frozen without any fine-tuning in our experiments.

Composition. After tokenizing the input as layout and patch tokens, we simplify the image generation task into an image composition task, where we can only focus on how to produce the final sequential discrete patch tokens with the proposed Transformer with Focal Attention (TwFA). In detail, given the object tokens $\boldsymbol{c}$ and the generated (or ground-truth) patch tokens $\boldsymbol{s}_{<i}$, the TwFA is introduced to model the long-range dependency and predict the probability of the next patch token s_i:

$$p(s_i|\boldsymbol{s}_{<i}, \boldsymbol{c}) = \mathrm{TwFA}(\boldsymbol{s}_{<i}, \boldsymbol{c}). \tag{2}$$

In an auto-regressive manner, TwFA generates the final patch tokens $\boldsymbol{s}$ step-by-step.

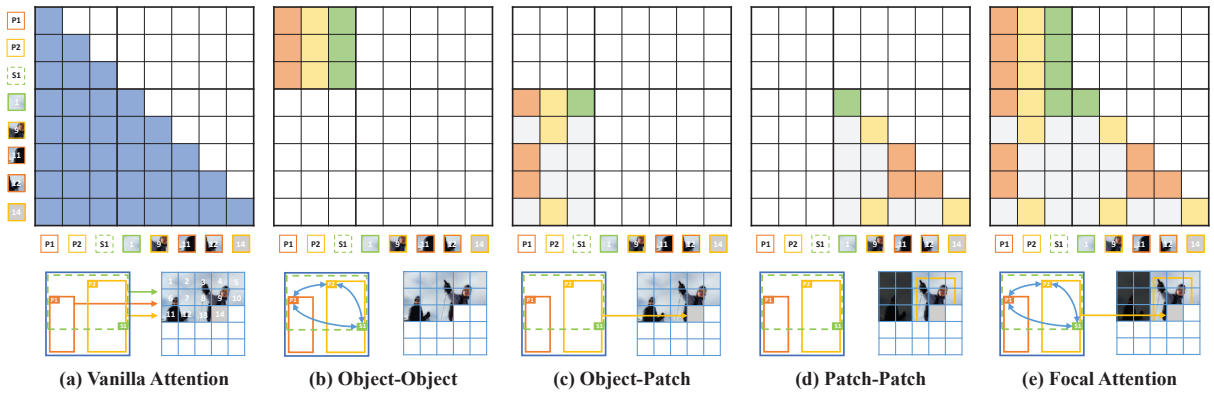

Figure 3. The illustration of different attention mechanisms with connectivity matrix. (a) Vanilla attention follows a casual mask, neglecting different interactions, (b) Object-object interaction enhances the modeling of object-level relationships; (c) Object-patch interaction makes patches to realize the object categories, (d) Patch-patch interaction introduces the composition prior of layouts, (e) Therefore our focal attention captures both object-level relationships and patch-level structures. Colors(orange, yellow, green) correspond to different objects.

Generation. With the generated patch tokens $\boldsymbol{s}$, we further reconstruct it into a real image by VQ-GAN decoder, as:

$$\hat{x} = Dec(\boldsymbol{s}). \tag{3}$$

Since arbitrary appearances of different objects can be encoded into the codebook, VQ-GAN is a useful tool for modeling the texture information.

Training and Sampling. To train the TwFA, we directly employ a cross-entropy loss for the sequence prediction task:

$$\mathcal{L} = -\sum_{i=1}^{M} \log(p(s_i|\boldsymbol{s}_{<i}, \boldsymbol{c})), \tag{4}$$

where M is the token length of images, s_i and $\boldsymbol{s}_{<i}$ are tokens tokenized from ground-truth images. During inference, the ground-truth image and its patch tokens are not available. We leverage multinomial resampling strategy to generate diverse images for the same layout.

3.2. The Attention

In this part, we first revisit the vanilla attention of Transformer [35], and then elaborate the proposed Focal Attention as illustrated in Figure 3.

3.2.1 Vanilla Attention

Given the tokenized object tokens $\boldsymbol{c}$ and patch tokens $\boldsymbol{s}$, we embed them into the feature $\boldsymbol{F} = [\mathrm{Emb}_c(\boldsymbol{c}); \mathrm{Emb}_s(\boldsymbol{s})]$, where $\mathrm{Emb}_{c/s}$ are two embedding layers, and $[;]$ denotes the concatenate operation. Then the feature $\boldsymbol{F}$ is fed into a multi-layer decoding Transformer, which adopts attention mechanism with Query-Key-Value (QKV) model.

Given the queries $\boldsymbol{Q} = \boldsymbol{W}_Q\boldsymbol{F}$, keys $\boldsymbol{K} = \boldsymbol{W}_K\boldsymbol{F}$, and values $\boldsymbol{V} = \boldsymbol{W}_V\boldsymbol{F}$, the vanilla attention is given by:

$$\mathrm{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D_k}} \circ \boldsymbol{M}\right)\boldsymbol{V}, \tag{5}$$

where $\boldsymbol{M}$ is called connectivity matrix, and $\circ$ denotes element-wise product. In the standard self-attention mechanism, every token needs to attend to all other generated tokens, *i.e.*, $\boldsymbol{M}$ is a causal mask as shown in Figure 3 (a), given by:

$$\boldsymbol{M}[i,j] = \begin{cases} 1, & \text{if } j \le i, \\ -\infty, & \text{else.} \end{cases} \tag{6}$$

However, the vanilla attention neglects the different type tokens, *i.e.*, object tokens and patch tokens, in our L2I task, hindering the model to well capture the object-level relationships and patch-level instance structures. For example, while generating the 14th patch in Figure 3 (a), the token attends to all object tokens, including P1, P2, and S1, even though the 14th patch doesn't belong to P1 and S1. Meanwhile, the token also attends to all generated patch tokens, even though not all patches are related to the 14th patch.

3.2.2 Focal Attention

To address the above issues in vanilla attention, We carefully design the connectivity matrix to guide the transformer to focus on the related tokens. To better demonstrate mechanism of our focal attention, we further decompose the connectivity matrix $\boldsymbol{M}$ into three areas, *i.e.*, object-object, object-patch, and patch-patch interaction.

Object-Object Interaction. To model the object-level relationships and learn the global context for each object, we design the object-object interaction as shown in Figure 3 (b). The dense interaction makes each object can interact with each other to capture the relationships by the multi-layer transformer, which is essential for object structure reasoning, for example, to generate a man kicking a soccer, the human's action can be predicted by the relative position between him and the ball. While vanilla attention models the

context in a single direction, in other words, the first object never know others. Hence, the connectivity matrix $\boldsymbol{M}_{oo}$ in the object-object area can be written as:

$$\boldsymbol{M}_{oo}[i,j] = 1. \tag{7}$$

Object-Patch Interaction. To make the patch aware of the object category it belongs to, we design the object-patch interaction as shown in Figure 3 (c). We stipulate that a patch of an object attends only on the corresponding object token to enhance the representation of the class, and a patch of a stuff or background can attend on all object tokens to make sure the image surroundings is consistent with the complex scene. Formally, the connectivity matrix $\boldsymbol{M}_{op}$ in the object-patch area is:

$$\boldsymbol{M}_{op}[i,j] = \begin{cases} 1, & \text{if } \boldsymbol{s}_i \text{ relates to } \boldsymbol{c}_j, \\ -\infty, & \text{else.} \end{cases} \tag{8}$$

Here, we explicitly distinguish the instance objects, *e.g.*, man and bus, and stuff objects, *e.g.*, sky and grass. If $\boldsymbol{s}_i$ locates in an instance, we define $\boldsymbol{s}_i$ only relates to this one instance. If $\boldsymbol{s}_i$ locates in a stuff, we define $\boldsymbol{s}_i$ relates to every objects for we hypothesize the stuff area generation relies on the global information.

Patch-Patch Interaction. To ensure the generative consistency in the local area, the relationships between patches need considerations. As shown in Figure 3 (d), the patch-patch interaction realize the isolation between the instance and background and meanwhile keep the local consistency. Formally, the connectivity matrix $\boldsymbol{M}_{pp}$ in the patch-patch area is:

$$\boldsymbol{M}_{pp}[i,j] = \begin{cases} 1, & \text{if } \boldsymbol{s}_i \text{ and } \boldsymbol{s}_j \text{ are neighbors}, j \leq i, \\ -\infty, & \text{else.} \end{cases} \tag{9}$$

Thanks to the composition prior from the layout, here we define two patches are neighbors if they belong to the same object, or they are both belong to stuffs or background.

By combining the above three types of interaction mechanisms, the connectivity matrix $\boldsymbol{M}$ of focal attention can be written as:

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{oo} & -\infty \\ \boldsymbol{M}_{op} & \boldsymbol{M}_{pp} \end{bmatrix}. \tag{10}$$

Finally, by unambiguously dealing with the interactions of different types, we increase the data efficiency and make it possible for the complex scene few-shot learning.

3.3. The Few-Shot L2I

Upon the well-trained TwFA, we are ambitious on few-shot layout to image generation. Given a novel object class and a few images containing this novel object, we aim at learning a model which can synthesize the complex scene image containing the novel class, while keeping the performance for the base ones.

The few-shot framework is based on the TwFA, we 1) append a new instance embedding to the input layer for the novel class, 2) split the last transformer layer into two, where the first one is for the base classes, and the second one is for the novel class, and 3) insert a token fusion module, which fuse the tokens from two last transformer layer according to the spatial layouts, to produce the final sequential patch tokens.

To train the few-shot framework, we initialize the new instance embedding with the well-trained embedding of its superclass, the second transformer layer with the first one, and fine-tune the new instance embedding and the second last transformer layer while keeping the parameters of pre-trained TwFA frozen. It's worth noting that thanks to the separate modeling on the hierarchical object-patch-pixel level and the sparse focal attention, our TwFA can fast adopts to new class with only a few images and achieves impressive performance.

4. Experiments

In this section, we first introduced our experimental settings that include training/testing datasets, and evaluation metrics. Then, we carried out quantitative and qualitative comparisons between our method and state-of-the-art CNN-based and transformer-based layout-to-image generation methods. Ablation studies are performed to validate effectiveness of the proposed focal attention. Finally, we performed few-shot complex scene image generation.

4.1. Experimental settings

Datasets. Following previous layout to image generation papers, we validate the proposed TwFA and state-of-the-art methods on two datasets: COCO-stuff [2, 22] and Visual Genome [17]. COCO-stuff dataset is an expansion of the Microsoft Common Objects in Context (MS-COCO) dataset, which includes 91 stuff classes and 80 object classes. Visual Genome (VG) is a complex scene understanding dataset that contains annotations such as bounding boxes, object attributes, relationships, *etc*. Following existing works [10, 20, 33], we only employed scene images, bounding boxes and labels in both datasets. In this paper, for fair comparisions, all models are trained on the resolution of 256×256.

Implementation Details. In the initial stage, the shot edge of each training image is first scaled to 296 pixels, keeping the image's aspect ratio unchanged. Then the pre-trained VQ-GAN is utilized to tokenize each 256×256 px crop into 16×16 tokens. The codebook size is set to 8192. In the second stage, we leverage these tokens to train our TwFA (24 layers, 16 attention heads, and 1024 embedding dimensions) and only implement the focal attention on the

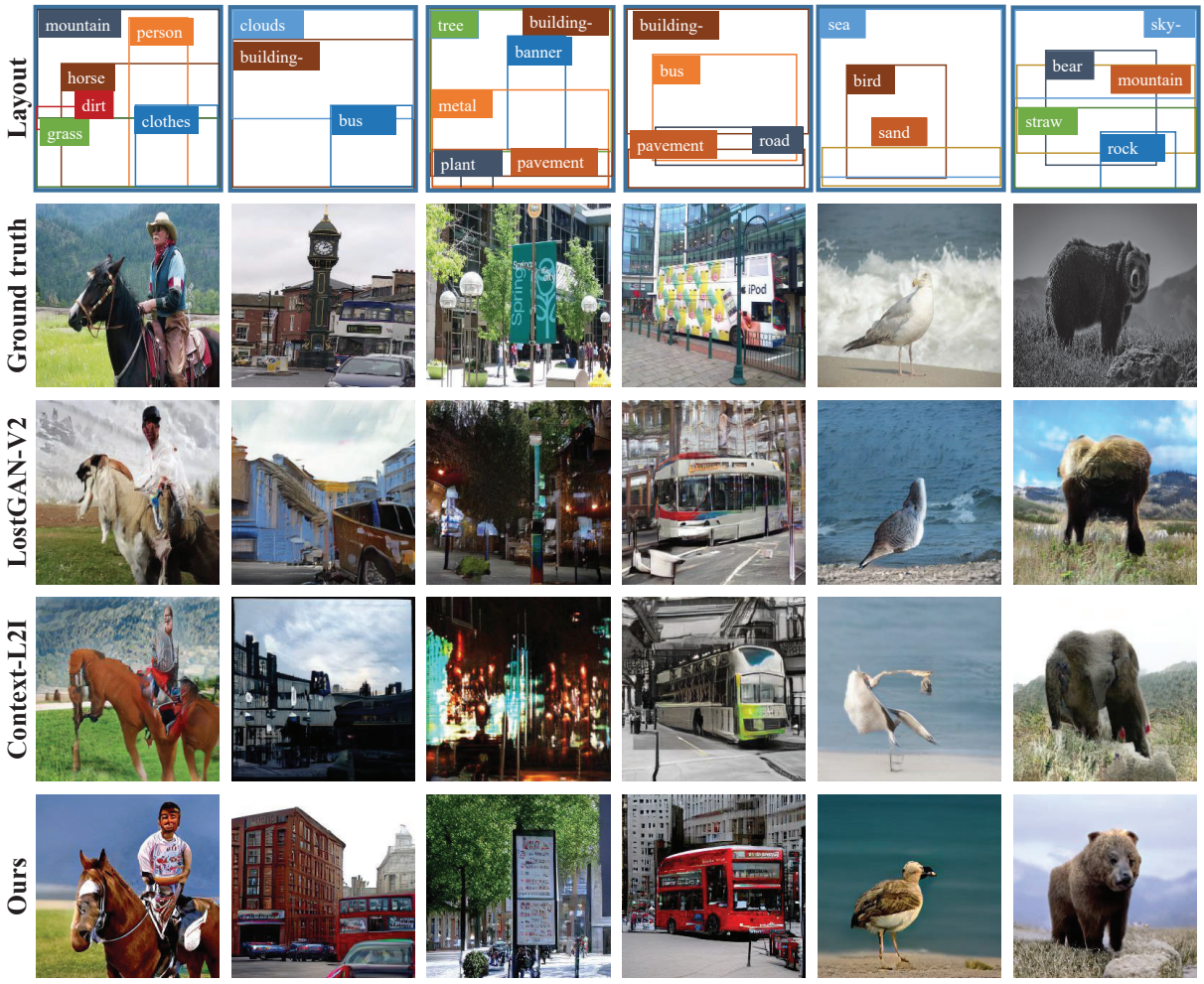

Figure 4. Samples generated from the layouts in COCO-stuff [2] by our method against the most representative baseline model, *i.e.* LostGAN-V2 [33] and the state-of-the-art existing model in Table 1, *i.e.* Context-L2I [10]. For all the different scenes, TwFA outperforms the state-of-the-art model with finer instance structures. More results are demonstrated in supplementary materials.

instance objects. The dropout rate is set to 0.1. When testing, we directly resize the images to 256×256 without any other augmentation.

Evaluation Metrics. To evaluate the generation performance over all comparison methods and our TwFA, we adopted five metrics to evaluate the visual realism and diversity of generated complex scene images. They are Inception Score (IS) [31], Frechet Inception Distance (FID) [11], SceneFID [34], Diversity Score (DS) [44], and YOLO Scores [20]. Inception Score (IS) is one of the earliest metrics for automatically evaluating the quality of image generative models. For the FID score, it first extracts image features from a pretrained backbone network (*e.g.* Inception V3 trained on ImageNet dataset), then computes the 2-Wasserstein metric between real-world images and generated images. Similarly, SceneFID is proposed for complex scene generation tasks. It computes the Frechet Inception Distance (FID) on the crops of all objects instead of the whole image. Different from measuring the distribution of generated images, Diversity Score (DS) compares the difference between the generated image and the real image from the same layout. Additionally, YOLO Scores are employed as an evaluation metric to measure the consistency of generated images' layouts to conditions.

4.2. Comparisons with Existing Methods

To validate the effectiveness of the proposed TwFA, we compared our model with both CNN-based and Transformer-based complex scene generation methods. Among them, CNN-based methods include LostGAN-V2 [33], OCGAN [34], LAMA [20], Context-L2I [10] and the only transformer-based method is HCSS [12]. For a fair comparison, we adopt their official released pre-trained models or the official reported scores in their papers.

	FID ↓		SceneFID ↓		Inception Score ↑		Diversity Score ↑	
	COCO	VG	COCO	VG	COCO	VG	COCO	VG
LostGAN-V2 [33]	42.55	47.62	22.00	18.27	18.01±0.50	14.10±0.38	0.55±0.09	0.53±0.09
OCGAN [34]	41.65	40.85	-	-	-	-	-	-
HCSS [12]	33.68	19.14	13.36	8.61	-	-	-	-
LAMA [20]	31.12	31.63	18.64	13.66	-	-	0.48±0.11	0.54±0.09
Context-L2I* [10]	29.56	-	14.40	-	18.57±0.54	-	0.65±0.00	-
Ours	**22.15**	**17.74**	**11.99**	**7.54**	**24.25±1.04**	**25.13±0.66**	**0.67±0.00**	**0.64±0.00**

Table 1. Quantitative results on COCO-stuff [2] and Visual Genome (VG) [17]. For fair comparisons, all the results are taken from the original papers and based on the resolution of 256×256. '-' means the related value is unavailable in their papers. '*' denotes results on samples from trained models with the official implementation.

	Grid2	Grid4	Grid16	Ours
FID ↓	27.64	29.01	27.04	**22.15**
S-FID ↓	17.26	19.43	16.80	**11.99**
IS ↑	21.84±0.88	21.50±0.59	22.73±0.65	**24.25±1.04**

Table 2. Comparison of different attention configurations. Grid16 is equivalent to vanilla attention. S-FID denotes SceneFID.

YOLO Scores	Grid16	Ours w/o oo	Ours w/o op	Ours w/o pp	Ours Full
AP_{50} ↑	25.97%	25.01%	27.59%	26.00%	**28.20%**
AP_{75} ↑	17.45%	17.38%	19.00%	17.93%	**20.12%**

Table 3. Comparison of different interaction configurations in the Connectivity Matrix. "oo", "op", and "pp" denote the object-object, object-patch, and patch-patch interaction respectively.

The quantitative results by the involved competitors on both the COCO-stuff and Visual Genome datasets are reported in Table 1. Among existing methods, Context-L2I [10] achieved the best overall performance. HCSS [12] employed a transformer with self-attention to perform the complex scene composition modeling task. Since we employed the same texture tokenization strategy, the generation performance depends on how well the transformer can model the composition of complex scenes. Compared to them, ours achieved significant improvement on all metrics.

In Figure 4, we provide several visual comparisons of the complex scene images generated by different methods based on the same layout. According to the visual comparison, we can observe that pervious methods are capable of generating reasonable texture and patches. According to the generated texture, we can roughly understand the synthesized scene. However, they failed to accurately model the instance structures. For example, the bear generated by LostGAN looks like a collection of the animal fur. According to the visual examples, the proposed TwFA performs better on constructing relationships between objects and modeling structure of instances. More examples can be found in our supplementary material.

Overall, the superiority of the proposed TwFA is validated on both quantitative metrics and qualitative visual comparison. The metrics such as FID, sceneFID and IS demonstrated the distribution of TwFA-generated images are statistical better then other methods. And TwFA largely improved the visually quality of complex scene generation.

4.3. Ablation Study

Here, we aim to explore how the attention mechanisms will influence transformer modeling of the complex scene generation. Besides the proposed focal attention, we performed other three attention configurations. First, we employed the global self-attention with causal mask that is widely used in language and image generation tasks. When predicting the current token, it performs self-attention with all given tokens, *i.e*. the layout conditions and all patch tokens before it. Inspired by the recent popular local attentions proposed in vision transformer, we test additional two attention mechanisms with sliding window size 2 and 4, *i.e*. 'Grid2' and 'Grid4'. Specifically, when predicting the current patch token, besides layout conditions, the model only attended on the given patch tokens within the 2D sliding windows. In our model, since the size of compressed token map is 16×16, therefore the setting of global self-attention is equivalent to 'Grid16'.

The quantitative results are reported in Table 2. We can see that our focal attention achieved the best performance on all metrics. Interestingly, we find that both Grid16 and Grid2 perform better than Grid4. It means either modeling global dependencies or local dependencies would contribute to increasing the performance, yet selecting a window size larger than 2 may decrease the performance. A possible explanation is that the relatively large window size is easier to break down the dependencies within and outside the bounding box, and meanwhile, fail to model global relationships. Compared with them, the proposed focal attention better utilizes the information provided by layouts. The same conclusion can be drived from Figure 5. As illustrated in the second row, the models with a larger window size failed to generate another face correctly.

Additionally, we ablate different connectivity matrix components, *i.e.*, object-object, object-patch, and patch-

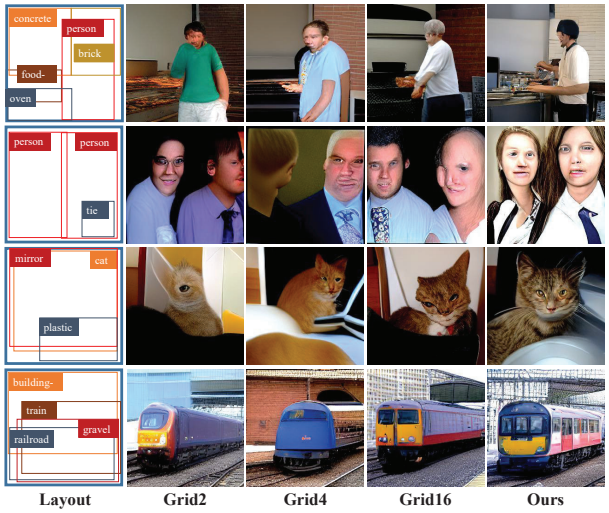

Figure 5. Qualitative ablative results of different attention configurations. Our focal attention achieves the best results.

patch interaction, to investigate their effect to object composition modeling in complex scene generation. Here, we utilize YOLO Scores to evaluate the alignment and fidelity of generated objects. As shown in Table 3, we have the following observations: **1)** The lack of the object-object interaction leads to a huge decrease. It indicates the global context greatly influences the object structure reasoning; **2)** Both "Ours w/o op" and "Ours w/o pp" performs better than Grid16, which suggests the well-designed object-patch/patch-patch interaction is essential to generating an photorealistic object; **3)** By integrating all the interactions, the TwFA improves Grid16 from 17.45% to 20.12% (*+15.3%*) and from 25.97% to 28.20% (*+8.6%*) on AP_{75} and AP_{50}, respectively.

4.4. Few-shot Complex Scene Generation

As aforementioned, benefiting from the accurate relationship modeling, the well-trained TwFA has the potential to perform few-shot complex scene image generation. Specifically, through fine-tuning on a few of images that contain unseen objects, we hope our model can be trained to generate this kind of objects giving new layouts.

For providing both quantitative and qualitative analysis, we trained a baseline (*i.e.* Grid16*) and a Ours* using COCO-stuff training data that removed all zebra images. Thus, zebra images in the original COCO-stuff dataset can act as the additional images for few-shot learning. Here, we choose zebra images for two reasons, (i) zebra is a kind of challenging target to synthesize, since it has a unique texture and complex structure/pose; (ii) there are relatively large amount of zebra images (∼1500) in COCO-stuff datasets, which contributes to more accurate quantitative measurement (*e.g.* FID prefer testing on more images).

Figure 6. Examples of few-shot results. The novel classes are the Christmas tree, penguin, and hot air balloon. TwFA outperforms all the baseline model with finer structures and details.

	Grid16*		Ours*	
# of Shot	FID ↓	Obj-FID ↓	FID ↓	Obj-FID ↓
20	39.47	35.62	**36.34**	**31.17**
30	39.32	34.59	**34.87**	**29.33**
40	37.36	31.34	**34.30**	**28.87**
50	37.47	32.81	**31.53**	**26.73**
Full trained	30.28	21.96	**24.33**	**21.66**

Table 4. Qualitative few shot results. Obj-FID only computes the FID score on the crops of the novel class with a size of 224×224.

As shown in Table 4, all experiments on 20/30/40/50 shots show the superiority of our Ours*. Meanwhile, accompany with the increasing of shot number, better generation performance can be achieved. In Figure 6, three new classes (Christmas tree, penguin, and hot air balloon) with 2 samples are employed to fine-tune both the baseline and our TwFA. From the visual examples, we can see our TwFA show better instance structure compared to the baseline model. Since space limitation, more few-shot generation results are reported in the supplementary material.

5. Conclusion

In this paper, we presented a novel Transformer with Focal Attention (TwFA) to disentangle the modeling between object-level relationships and patch-level instance structures, and introduce the composition prior from spatial layouts into image compositions. Compared with CNN-based and Transformer-based methods, TwFA enables the model to capture the inherent instance structures, and increase the data efficiency to alleviate the burden on few-shot learning with limited data. With extensive experiments and visualizations on both COCO-stuff and Visual Genome datasets, the proposed TwFA demonstrates its superiority over the SoTA methods on both L2I and few-shot L2I tasks.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 2, 5, 6, 7

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 2

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 3

[5] Stanislav Frolov, Avneesh Sharma, Jörn Hees, Tushar Karayil, Federico Raue, and Andreas Dengel. Attrlostgan: Attribute controlled image synthesis from reconfigurable layout and style. In *The German Association for Pattern Recognition*, 2021. 2

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[7] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao. Lofgan: Fusing local representations for few-shot image generation. In *ICCV*, 2021. 3

[8] Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems*, pages 1143–1152, 2019. 3

[9] Fengxiang He, Bohan Wang, and Dacheng Tao. Piecewise linear activations substantially shape the loss surfaces of neural networks. In *International Conference on Learning Representations*, 2020. 3

[10] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *CVPR*, 2021. 1, 2, 5, 6, 7

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[12] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. In *CVPRW*, 2021. 2, 3, 6, 7

[13] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, 2018. 2

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 2, 5, 7

[18] Yandong Li, Yu Cheng, Zhe Gan, Licheng Yu, Liqiang Wang, and Jingjing Liu. Bachgan: High-resolution image synthesis from salient object layout. In *CVPR*, 2020. 2

[19] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *NIPS*, 2020. 3

[20] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *ICCV*, 2021. 2, 5, 6, 7

[21] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*, 2021. 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5

[23] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 2

[24] Ke Ma, Bo Zhao, and Leonid Sigal. Attribute-guided image generation from layout. In *BMVC*, 2020. 2

[25] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, 2021. 3

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 2

[27] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 2

[28] Scott Reed, Yutian Chen, Thomas Paine, Aäron van den Oord, SM Ali Eslami, Danilo Rezende, Oriol Vinyals, and Nando de Freitas. Few-shot autoregressive density estimation: Towards learning to learn distributions. In *ICLR*, 2018. 3

[29] Danilo Rezende, Ivo Danihelka, Karol Gregor, Daan Wierstra, et al. One-shot generalization in deep generative models. In *ICML*, 2016. 3

[30] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 2

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 6

[32] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *ICCV*, 2019. 2

[33] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *TPAMI*, 2021. 2, 5, 6, 7

[34] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. In *AAAI*, 2021. 2, 6, 7

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 4

[36] Chaoyue Wang, Chaohui Wang, Chang Xu, and Dacheng Tao. Tag disentangled generative adversarial network for object image re-rendering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2901–2907, 2017. 2

[37] Chaoyue Wang, Chang Xu, Xin Yao, and Dacheng Tao. Evolutionary generative adversarial networks. *IEEE Transactions on Evolutionary Computation*, 23(6):921–934, 2019. 2

[38] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *CVPR*, 2020. 3

[39] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020. 3

[40] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *ICCV*, 2021. 3

[41] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. In *Advances in Neural Information Processing Systems*, 2021. 2

[42] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 2

[43] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv preprint arXiv:2202.10108*, 2022. 2

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[45] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *CVPR*, 2019. 2

[46] Bo Zhao, Weidong Yin, Lili Meng, and Leonid Sigal. Layout2image: Image generation from layout. *IJCV*, 2020. 2