Image Generation from Layout via Pair-Wise RaGAN

Ting Xu, Kai Liu, Yi Ji(✉), and Chunping Liu

College of Computer Science and Technology, Soochow University, Suzhou, China
{txu7,kliu0923}@stu.suda.edu.cn, {jiyi,cpliu}@suda.edu.cn

Abstract. Despite recent remarkable progress in image generation from layout, synthesizing vivid images with recognizable objects remains a challenging problem, object distortion and color imbalance occasionally happened in the generated images. To overcome these limitations, we propose a novel approach called Pair-wise Relativistic average Generative Adversarial Network (P-RaGAN) which includes a pair-wise relativistic average discriminator for enhancing the generative ability of network. We also introduce a consistency loss into our model to keep the consistency of original latent code and reconstructed or generated latent code for reducing the scope of solution space. A series of ablation experiments demonstrate the capability of our model in the task from layout to image on the complicated COCO-stuff and Visual Genome datasets. Extensive experimental results show that our model outperforms the state-of-the-art methods.

Keywords: Image generation · Layout · Relativistic average Discriminator · Consistency loss

1 Introduction

Image generation from layout is a novel hot topic in computer vision, which requires the function of dealing with incomplete information in the layout and the ability of learning how to handle multi-modal information(vision and language). Existing powerful algorithms can conveniently assist technicians to comprehend visual patterns. Therefore, there exists a number of visual applications, *e.g.* self-driving technology, it is a help for art creation and a novel aid in scene graph generation [25].

Here layout includes semantic layout (bounding boxes and object shapes) and spatial layout (bounding boxes and object category). It can be used not only as

Supported by National Natural Science Foundation of China (61972059, 61773272), The Natural Science Foundation of the Jiangsu Higher Education Institutions of China (19KJA230001), Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (93K172016K08), Suzhou Key Industry Technology Innovation-Prospective Application Research Project SYG201807), the Priority Academic Program Development of Jiangsu Higher Education Institutions.

H. Zhang et al. (Eds.): NCAA 2020, CCIS 1265, pp. 193–206, 2020.
https://doi.org/10.1007/978-981-15-7670-6_17

initial domain in the task from layout to image but also as intermediate representation in other task such as text-to-image (known as T2I). Most promising results in deep learning are based on generative model like Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs). Hong *et al.* [9] proposed a two-step image synthesizing method with constructing a semantic layout from the text, which can generate an image conditioned on the layout and text description. But their methods did not follow the end-to-end training manner. Hinz *et al.* [8] introduced a novel method which added an object pathway to both the generator and discriminator for controlling the location of the objects. Compared with semantic layout, the spatial layout is a coarse-grained description. Image generation from the spatial layout does not need annotate the segmentation masks. There are two main challenges in this task. The first challenge is how to synthesize images which must correspond to layout. Second, from the perspective of human vision, the generated samples need to be real enough. Zhao *et al.* [22] proposed an approach called layout2im based on VAE-GAN network for synthesizing images from the spatial layout. Most their samples generated by layout2im did correspond to the given layout. However, object distortion and color imbalance has occasionally arisen in the generated samples.

Most existing image generation from layout to image tasks are GAN-based methods, but the training of vanilla GANs is notorious due to its uncertainty and instability. Therefore, many researchers are devoted to the method study of how to stabilize the training of GANs. The Wasserstein GAN (WGAN) [4] is a good illustration of stability training, which adopts Earth Mover Distance (EMD) instead of Jensen-Shannon (JS) divergence as an objective. Mao *et al.* [23] propose another method by replacing the cross entropy loss function with the least square loss function. The method also attempted to use different distance measures to build a more stable and converging adversarial network. Although these methods have partially solved the stability problem, some of them such as[3] are required more powerful computing performance than Standard GANs.

Our work aims to improve the generative ability of GAN for the high-quality of generated images. The overall framework of the proposed method is shown in Fig. 1. Consequently, we propose a novel method which contains two new components. One is to introduce a pair-wise relativistic average discriminator[12] for increasing generative capability, the other is to propose a consistency loss function for reducing the scope of solution space. The major contributions of our method are summarized here:

1) We propose a new model for generating realistic image from coarse layout.
2) We introduce a pair-wise relativistic average discriminator, which includes an image-wise discriminator that can generate more recognizable and vivid images and an object-wise discriminator that can address the problem of object distortion.
3) We adopt a consistency loss for narrowing the scope of solution space.

In the remainder of this paper is described below. We shortly present related work about our study in Sect. 2, and in Sect. 3, we illustrate in detail the components of P-RaGAN. Then we focus on the experimental results and analysis in Sect. 4.

2 Related Work

Conditional GANs. The introduction of GAN[6] has achieved promising results in image generation tasks. It has been an increasing attention on conditional image generation or manifold-valued Image Generation [24]. Conditions can be text [17,20,21], image [10], scene graph [11] or layout[22]. However, model collapse does exist with these GAN-based methods. In the Standard Generative Adversarial Networks (SGANs) methods, the discriminator is only responsible for obtaining the probability that the input sample is real data, and the generator is trained to synthesize samples that can "cheat" discriminator. But this way isn't appropriate in the sight of a priori knowledge that half of a mini-batch of data comes from the generated samples [12]. Moreover, from the perspective of human cognition, the discriminator should also output higher scores to generated sample which is more realistic than the real one. [12] proposed a relativistic discriminator to solve these problems and achieved the satisfied results. But it only judges whether the input image is relatively realistic at image level. For this, we discriminate at object level by comparing the authenticity between the input object patches and the opposite type.

Image Generation from Layout. The image generation from layout is formulated as:

$$I_g = G\left(L_S\right). \tag{1}$$

where L_S denotes the original layout which is composed of an appearance latent code z^r, the position of objects in images $B_{1:T}$ and the class label of object c. G is a generator network. I_g refers to the generated image. Since $z^r \in \mathbb{R}^{64}$ and $I_g \in \mathbb{R}^{C\times H\times W}$($C$ refers to the channel, H and W are the height and the width of input image, respectively.) are essentially a low dimensional vector and a high dimensional vector, respectively, the layout-image pair can't be perfectly aligned and it prevents the generator from synthesizing more realistic and detailed images.

In most recent studies [8,9,11,14,22], layout-based image generation methods have been explored. These methods [8,9,11,14] divided T2I task into the following two steps: language → layout and layout → image. These approaches reduce the difficulty of generation because they employ layout as a bridge across a huge semantic gap between images and natural language descriptions. Despite that layout2im [22] takes layout as input, the approach can generate image by combining variational auto-encoders and generative adversarial network. Because of the standard generative adversarial network in Layout2im, the quality of the generated images is relatively poor.

One major challenge in layout to image task is how to generate images with more details and realistic objects without shape abnormality. To this end, we propose a novel Pair-wise Relativistic average Generative Adversarial Network (P-RaGAN). P-RaGAN consists of a pair-wise relativistic average discriminator. We also introduce a consistency loss for promising the consistency of the object latent code pairs (z^r, z'_r) and (z^o, z'_o).

3 Generating Realistic Image from Coarse Layout

It is proved that the training of traditional GANs network such as SGANs or DCGANs is troublesome from existing powerful empirical and theoretical evidence [15]. Therefore the task of generating high-quality images is difficult due to the training instability of traditional GANs. Thus, for generating more high-quality images, we construct a realistic image generation framework. The detailed framework is illustrated in Fig. 1. In order to generate images without object distortion by enhancing the power of GAN, we introduce the pair-wise relativistic average discriminator (PwRaD). And to reduce the scope of solution space, we also introduce the consistency loss function, as revealed in Fig. 1.

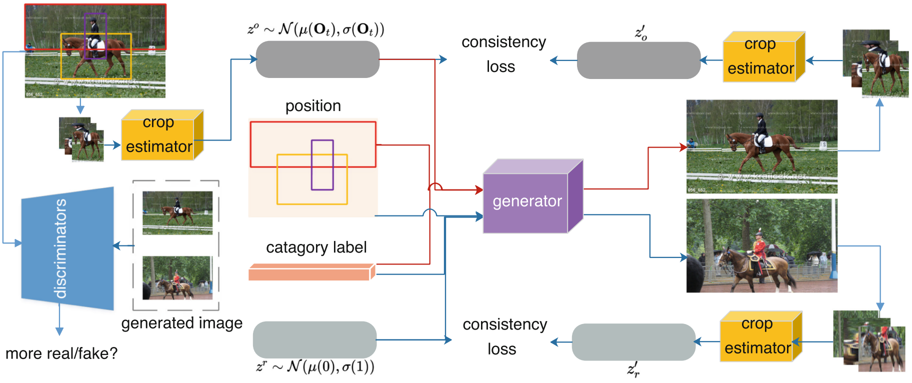

Fig. 1. Overview pipeline of the proposed model. The model first cuts out the objects in the real samples and employs the crop estimator to fit the distribution of these objects patches, then samples two latent codes from the distribution and the standard normal distribution, respectively. The layout consists of the code pair, together with the object's category label and location. Given the layout, it is fed into our generator to synthesize images. We also introduce a pair-wise relativistic average discriminator to generate reasonable and vivid images. In addition, consistency loss is designed to ensure the consistency between latent code pairs for constraining the range of solution space.

3.1 Relativistic Average Generative Adversarial Network

In [12], Jolicoeur-Martineau and Alexia pointed out that GANs based on non-Integral Probability Metrics (IPM) [16] couldn't generate the samples with high-quality, while IPM-based GANs only partially solved this problem. Hence, Jolicoeur-Martineau *et al.* proposed a relativistic discriminator network for stabilizing the training of GANs.

$$D(I_r, I_g) = \mathrm{act}\left(F\left(I_r\right) - F\left(I_g\right)\right) \tag{2}$$

Here, *act* is activation function, $F(\cdot)$ represents the output of the last convolutional layer, I_r means real samples from dataset, I_g describes the false examples from generator.

$$D^{rev}(I_r, I_g) = \mathrm{act}\left(F\left(I_g\right) - F\left(I_r\right)\right) \tag{3}$$

The model (formula (3)) not only increases the possibility of fake sample being more realistic, but also reduces the possibility of real data. To some extent, this makes the JS divergence between the two distributions ($\mathbb{P}_{data}$ and $\mathbb{P}_{fake}$) is minimized [12], which leads to the resolution of gradient disappearance in the GAN.

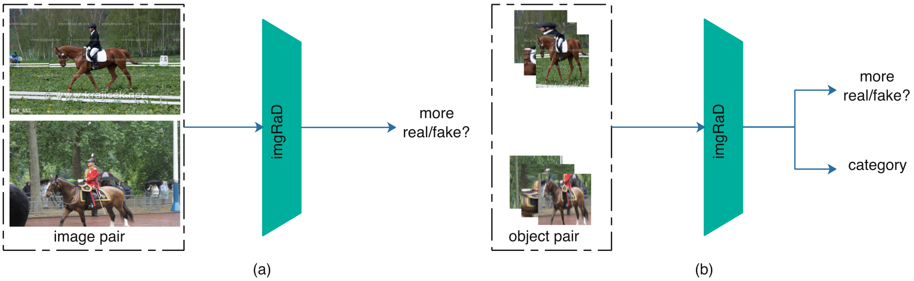

Fig. 2. Image-wise relativistic average discriminator (imgRaD) and object-wise relativistic average discriminator (objRaD). Both of imgRaD and objRaD need to compute the probability that the input is real. The objRaD also needs to output the probability that the input object belongs to each category, which is performed by a fully connected layer.

3.2 Pair-Wise Relativistic Average Discriminator

Inspired by the fact that the relativistic average discriminator (RaD) [12] can offer the promise of generalization ability, we propose a pair-wise relativistic average discriminator (Fig. 2) to estimate the probability of the given input data that is more realistic than the opposite type of input data on average. The discriminator includes an image-wise discriminator and an object-wise discriminator. Different from the standard GANs, the object-wise discriminator requires the similarity of input data and its opposite type in object level. The discriminators we proposed can be defined as:

$$\bar{D}_{img/obj}(\cdot) = \begin{cases} \text{act}\left(F(t) - \mathbb{E}_{I_g \sim \mathbb{P}_{fake}} F(z)\right) & \text{if } t \sim \mathbb{P}_{data} \\ \text{act}\left(F(t) - \mathbb{E}_{I_r \sim \mathbb{P}_{data}} F(I_r)\right) & \text{if } t \sim \mathbb{P}_{fake} \end{cases} \tag{4}$$

And the relativistic loss function of discriminator is:

$$\begin{aligned} L_{D_{img/obj}} = & -\mathbb{E}_{I_r \sim \mathbb{P}_{data}}\left[\log\left(\bar{D}_{img/obj}(I_r)\right)\right] \\ & -\mathbb{E}_{I_g \sim \mathbb{P}_{fake}}\left[\log\left(1 - \bar{D}_{img/obj}(I_g)\right)\right] \end{aligned} \tag{5}$$

In the image-wise discriminator (D_{img}), I_r and z represent the ground-truth and the synthesized samples respectively. In the object-wise discriminator (D_{obj}), I_r and z refer to ground-truth and synthesized images patches containing objects, respectively. Neither of the two discriminators is added to the batch normalization layer. The proposed discriminator improves the corresponding loss function and ensures that the input data is more real than its opposite type on average, not only more real than a small part. Experiments show that the proposed model can generate pseudo-real-world image.

3.3 Consistency Loss

Although the mapping function between layout and image can be learned through the mutual confrontation between the generator and the discriminator. It is difficult to obtain a perfect mapping function between the input and the output. In order to further reduce the space of mapping function, the object latent code $\boldsymbol{z_o'}$ of the reconstructed image should be as close as possible to the object latent code $\boldsymbol{z^o}$ of the real sample. Meanwhile, the latent code $\boldsymbol{z_r'}$ from the crop estimator should be essentially consistent with the code $\boldsymbol{z_r}$ from the standard normal distribution sampling. So we explore a consistency loss, which is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{con}} = & \mathbb{E}_{\left(z^o, z_o'\right) \sim \left(\mathbb{P}_{O_r}, \mathbb{P}_{O_g}\right)}\left[\|z^o - z_o'\|_1\right] \\ & + \mathbb{E}_{\left(z^r, z_r'\right) \sim \left(\mathbb{P}_{identity}, \mathbb{P}_{O_g}\right)}\left[\|z^r - z_r'\|_1\right] \end{aligned} \tag{6}$$

where $\mathbb{P}_{O_r}$ is the distribution of objects in the ground-truth images, $\mathbb{P}_{O_g}$ represents the distribution of objects in the generated.

3.4 Total Loss Function

Our final loss function is a weighted sum of these loss functions, which is defined in detail:

$$\mathcal{L}_{total} = \lambda_0 \mathcal{L}_{\text{KL}} + \lambda_1 \mathcal{L}_{con} + \lambda_2 \mathcal{L}_{\text{img}} + \lambda_3 \mathcal{L}_{\text{obj}} + \lambda_4 \mathcal{L}_{\text{AC}}^{\text{obj}} \tag{7}$$

where $\mathcal{L}_{\text{KL}}$ is the KL-divergence between the normal distribution and the distribution $\mathbb{P}_{O_t}$, $\mathcal{L}_{con}$ relates to the consistency loss, $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{obj}}$ describe the relativistic loss from discriminators, $\mathcal{L}_{\mathrm{AC}}^{\mathrm{obj}}$ is the same as the classifier loss in layout2im. The λ_i is the hyper-parameter for balancing the proportion of various losses. We set λ_1 to 0.1, 0.5, 1 and 5, respectively, in the experiments. We find that when we set $\lambda_0 = 0.1$, $\lambda_1 = 0.5$, $\lambda_2 = 1$, $\lambda_3 = 1$ and $\lambda_4 = 1$, we get the best results.

4 Experimental Results and Analysis

In this section, we have carried out extensive experiments on the datasets with multi-objects and complex scene images, *e.g.* COCO-Stuff [5] and Visual Genome [13] datasets. At the same time, we also evaluate the performance of the proposed method from three aspects: the recognizability and stability of the generated image, the consistency with the image distribution in the datasets and the structural similarity of human visual perception. Finally, we further analyze the role of the pair-wise relativistic average discriminator and the consistency loss function for improving the quality of generated image through ablation experiments.

4.1 Evaluation Metrics

We adopt three evaluation indicators for evaluating the performance of our layout-conditional image generation method: Inception Score (IS) [18], Fre'chet Inception Distance (FID) [7], and structural similarity index (SSIM) [19]. The evaluation of inception scores on visual quality is considered to be related to human perspective [2]. For measuring the recognizability and diversity of synthesized samples, we apply the pre-trained classifier (VGG-net [1]) to all images generated by our model and baseline to study the statistical characteristics of its score distribution for computing the inception scores. The higher the score, the better. Unlike Inception Score, the FID score can measure whether the generated image is in the same distribution as the image in the dataset. We prefer getting lower scores for better performance. Evaluation based on Fre'chet Inception Distance (FID) is beneficial for reality evaluation. But when an image has its own feature (for example, as long as it has the feature of human face, even though the position of eyes and mouth is changed), IS and FID will still give it a high evaluation, which is obviously contrary to human cognition. We choose SSIM as another evaluation metric because SSIM is considered to be correlated with the image quality perception of the human visual system (HVS) [19]. The higher the quality of the image, the greater the SSIM value. Here we show it in percentage form.

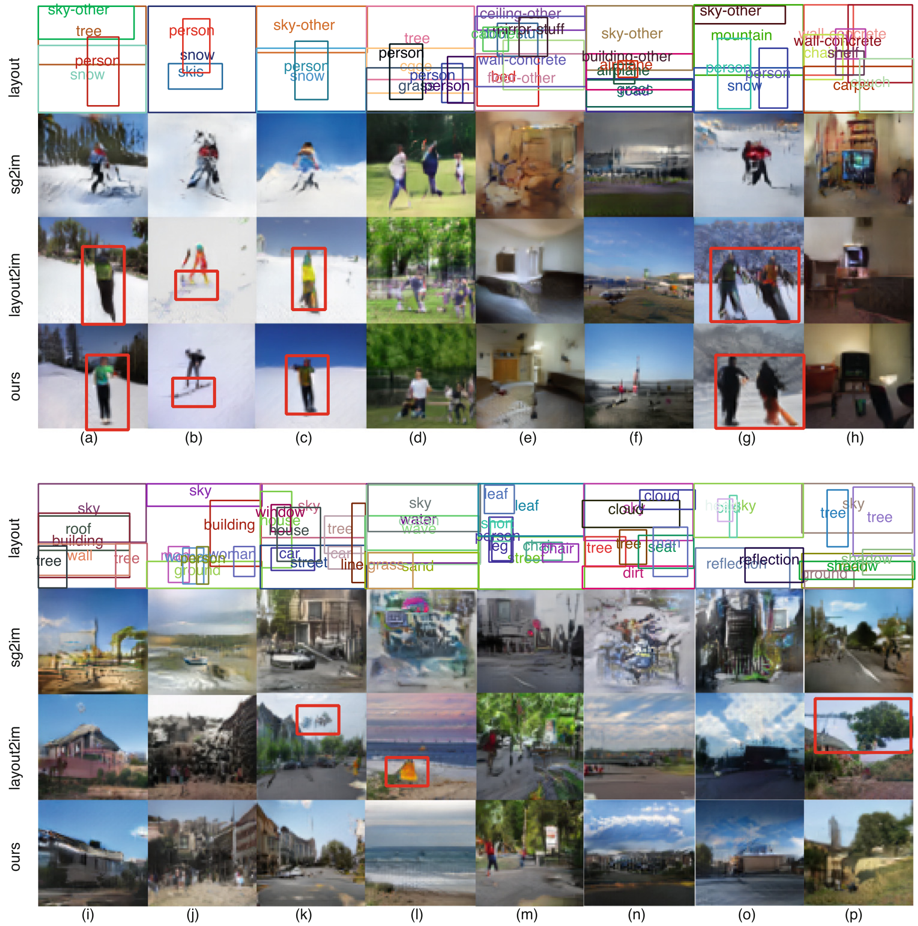

Fig. 3. The partial samples of generated image using given layout. The results of the COCO-Stuff and Visual Genome datasets are shown in rows 1 to 4 and 5 to 8, respectively. For sg2im and layout2im, we use the pre-trained model to generate images.

Table 1. Quantitative evaluation results. GT indicates ground-truth layout. layout2im-p refers to the pre-trained model. layout2im-o: we train the layout2im from scratch.

	IS		FID		SSIM	
Method	COCO	VG	COCO	VG	COCO	VG
real image (64 × 64)	16.3 ± 0.4	13.9 ± 0.5	-	-	-	-
pix2pix [10]	3.5 ± 0.1	2.7 ± 0.02	121.97	142.86	-	-
sg2im(GT) [11]	7.3 ± 0.1	6.3 ± 0.2	67.96	74.61	-	-
layout2im [22]	9.1 ± 0.1	8.1 ± 0.1	**38.14**	**31.25**	-	-
layout2im-p	9.1 ± 0.1	8.1 ± 0.1	42.80	40.07	24.2	36.1
layout2im-o	9.1 ± 0.1	8.1 ± 0.1	43.80	39.39	24.1	36.1
ours	**10.1 ± 0.1**	**8.5 ± 0.1**	38.70	34.13	**25.7**	**38.2**

4.2 Comparing Analysis

According to the three evaluation indexes, we compare with the experimental results of three state-of-the-art image generation methods based on layout, and the quantitative evaluation results are summarized in Table 1. We present three evaluation results based on IS, FID and SSIM. Pix2pix [10] has achieved promising results in image translation. Here, we set the original domain as the layout and the real image as the target domain. Sg2im [11] adopts graph convolution to handle scene graph in the T2I. It can be applied to generate image from layout only by using ground-truth layout. Layout2im [22] is originally trained for taking layout as input to generate images.

From the IS index of two experimental datasets, our method is improved from 9.1 to 10.1 and from 8.1 to 8.5 respectively. Since this index represents the recognizability and diversity of object in the generated image, this demonstrates that our method can generate more recognizable and realistic objects. FID is more suitable to describe the diversity of GAN, and has better robustness to noise. The FID of our model is slightly lower than that reported by layout2im, because the images generated by our model are more the same. Our approach increases SSIM from 24.2 to 25.7 and from 36.1 to 38.2. SSIM is a full reference image quality evaluation index. It measures the similarity of two images from three aspects of brightness, contrast and structure. Its design takes the visual characteristics of the human visual system (HVS) into account, which is more consistent with the visual perception of the human eye than the traditional way. This also explains that from a human point of view, our model produces images that are more strongly correlated with the real samples in the dataset. In other words, it indicates that our method is easier to generate images that are better aligned with real samples. And our proposed approach significantly outperforms these existing approaches [10,11,22].

In addition, we respectively illustrate the visual effect of the generated image in Fig. 3 and Fig. 4. Given the coarse layout, Fig. 3 shows the partial generated image results of sg2im, layout2im and the proposed methods on two experimental datasets. Due to the combination of variational auto-encoders and generative adversarial network, layout2im method does synthesize images that correspond with layout. However, the object shape is distorted in Fig. 3 (a), (c), (g) and (j). Furthermore, compared with the object location of given layout, some objects‘ location in the generated image is wrong in Fig. 3 (k), (l) and (p), which makes the generated images look fake. Since their model ignores the fact that half of the input data is fake, this inevitably leads to the distortion in the generated images. In Fig. 3 (e) and (n), the reason why the images generated by sg2im look chaotic is that the information that scene graphs can express is very limited, and only six relationships are defined in the experiment.

In contrast, our model applies the consistency loss penalty model to ensure that the generated image is close enough to the ground-truth, and on this basis, we make full use of the fact that half of the data fed into the discriminator is fake, so the generated image with high recognition is more close to the ground-truth image.

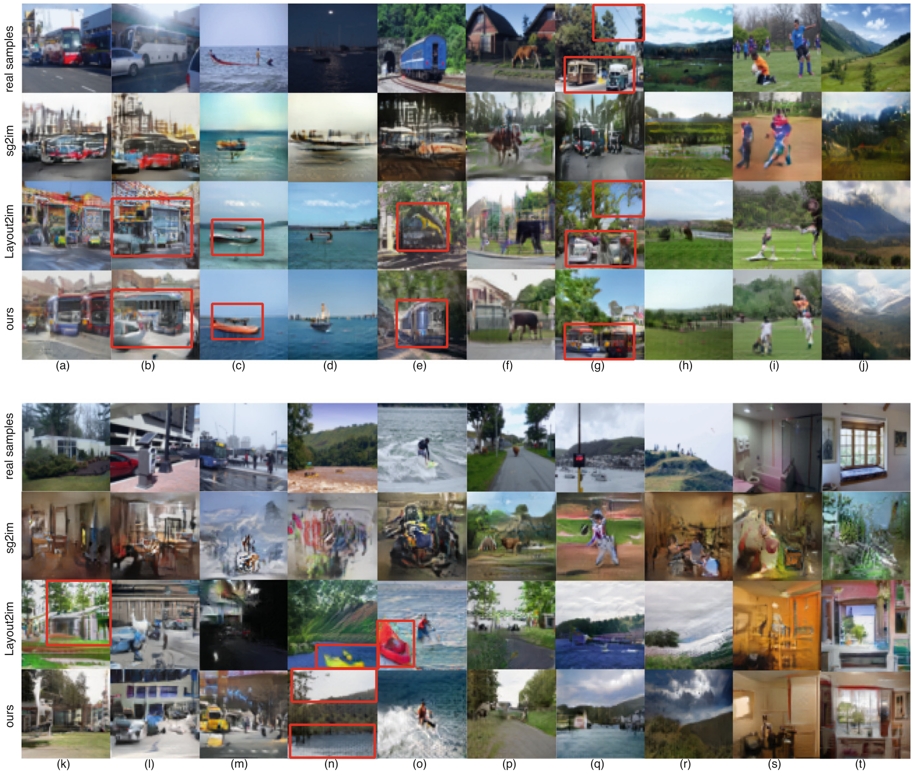

Fig. 4. Partial samples of generated images on the COCO-Stuff (top) and Visual Genome (bottom) datasets. The ground-truth images are shown in first row. For sg2im (the second row) and layout2im (the third row), we use the pre-trained model to generate samples. (Color figure online)

Table 2. Ablation Study. We trained baseline from scratch. CL is the consistency loss, imgRaD and objRaD refer to image-wise and object-wise relativistic average discriminator, respectively, the PwRaD is the pair-wise relativistic average discriminators.

	IS		FID		SSIM	
Method	COCO	VG	COCO	VG	COCO	VG
baseline-o	9.1 ± 0.1	8.1 ± 0.1	43.80	39.39	24.1	36.1
baseline+CL	9.47 ± 0.2	8.2 ± 0.1	39.72	35.46	24.9	36.9
baseline+imgRaD	8.9 ± 0.2	7.8 ± 0.3	40.13	38.51	24.1	36.5
baseline+objRaD	9.4 ± 0.1	8.1 ± 0.1	39.83	37.65	24.2	37.1
baseline+PwRaD	9.7 ± 0.1	8.3 ± 0.1	39.14	35.38	25.0	37.7
baseline+PwRaD+CL	**10.1 ± 0.1**	**8.5 ± 0.1**	**38.70**	**34.13**	**25.7**	**38.2**

To further prove that our proposed model doesn't memorize the images, but has ability to generate images. We present the experimental results based on given layout and given the ground-truth images in Fig. 4. As shown in Fig. 4 (b), our model changes the appearance and direction of the bus, the image generated by layout2im is not so real, although it looks like a car. The shape of the car in sg2im is excessively distorted. As depicted in Fig. 4 (c), without changing the background of the sea, our model "upgrades" the ship in the image. In Fig. 4 (n), our model has changed the color of the sky and the river, and layout2im has changed the color of both, but it has turned the river yellow, which is obviously not in line with the cognition of the human visual system. Sg2im produces incomprehensible images.

From Fig. 4, we can also see that our model can better overcome the problems of the object shape distortion and color imbalance in sg2im and layout2im. For shape distortion, layout2im composes a tree in the wrong place in Fig. 4 (g) and (k). Sg2im directly ignores the existence of trees. Figure 4 (a), the car from layout2im looks more like a wall. The car synthesized by sg2im can't be recognized by human eyes. For color imbalance, as illustrated in Fig. 4 (o), the image generated by layout2im shows an unknown red object, sg2im struggles with the object shape and color. From these examples, we can see that our model can generate images with correct colors and shapes: Fig. 4 (g) shows two buses, (o) doesn't contains strange stuff.

Based on the analysis of the three evaluation indexes and the corresponding visualization results, we can see that our proposed method is effective to overcome the problems of shape distortion and color distortion of objects, and our model has the ability to generate images containing complex objects. It is undeniable that our model is not very good for some slender and tiny objects. We consider that our input is set to 64 × 64 low resolution images. How to generate high resolution image with clear tiny objects, obviously, is an area of special interest in our future work.

4.3 Ablation Study

Compared with the framework of baseline model, our proposed framework of image generation from coarse layout includes two new components: pair-wise relativistic average discriminators and a consistency loss function. In order to further fully demonstrate that the role of each component in improving the quality of the generated images, we conduct ablation experiments on two experimental datasets. The ablation results are summarized in Table 2.

Pair-Wise Relativistic Average Discriminators. Our pair-wise relativistic average discriminator includes image-wise relativistic average discriminator (imgRaD) and object-wise relativistic average discriminator(objRaD). Through the combination of baseline model and the different components, the evaluation results of two experimental datasets under three evaluation indexes are given in Table 2. We test the compact of removing the relativistic average discriminators.

Specifically, the model trained without objRaD which decreases inception score obtains poor performance, since the model cannot generate recognizable objects. The SSIM score is still high because of the image-wise discriminator. Without the constraint of image relativistic adversarial loss, the model leads to lower score as it ignores the priori knowledge that only half of the input data is real, and the model does not reduce the possibility of discriminating the real data as real. The pair-wise relativistic average discriminator-based GAN achieves higher scores, showing that the synthesized images include more vivid structures. For example, Fig. 3 (b) shows the skis, (p) sets tree in correct position.

Consistency Loss. A consistency loss term in total loss function promises the alignment between the two pairs of latent codes (z^r, z'_r) and (z^o, z'_o). Does this result in better performance? Table 2 reports the scores of this variant. The model with consistency loss increases IS score from 9.1 to 9.4, decreases FID score to 39.7, and improves SSIM value to 24.9. In Fig. 4 (e), compared with sg2im and layout2im, the generative locomotive can be easily identified. Figure 4 (r) displays the mountain and sky which are very similar to those in ground-truth image. This demonstrates that the consistency loss penalizes the distance between the original latent codes z^r, z^o and reconstructed z'_r or generated z'_o latent code. Because L1 norm is more capable of handing the exception value, we use L1 norm to constrain the final loss function in the process of the image generation. And we also try L2 norm in the experiment, but the experimental results are not ideal.

5 Conclusion

In this paper, we propose an approach for image generation from layout which can synthesize more recognizable objects and vivid images. We improve the baseline model layout2im by introducing a pair-wise relativistic average discriminator and a consistency loss function. The quality of image generation is improved using the proposed image-wise and object-wise discriminators. We also design a consistency loss for constraining the scope of solution space. The proposed method Pair-wise Relativistic average Generative Adversarial Network (P-RaGAN) can achieve the better performance based on the objective evaluation metrics (IS, FID and SSIM) and the visual perception evaluation of generated image. Our future work is further enhance the tiny object of image generation.

References

1. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR, Proceedings of the International Conference Learning Representation, pp. 1–14 (2015)
2. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GAN. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 2642–2651 (2017)

3. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
4. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, Proceedings of Machine Learning Research, vol. 70 pp. 214–223, PMLR (2017)
5. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1209–1218 (2018)
6. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local NASH equilibrium. In: Advances in Neural Information Processing Systems, pp. 6626–6637 (2017)
8. Hinz, T., Heinrich, S., Wermter, S.: Generating multiple objects at spatially distinct locations. In: ICLR (Poster). OpenReview.net (2019)
9. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7986–7994 (2018)
10. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
11. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1219–1228 (2018)
12. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net (2019)
13. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**(1), 32–73 (2017)
14. Li, W., et al.: Object-driven text-to-image synthesis via adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12174–12182 (2019)
15. Martin, A., Lon, B.: Towards principled methods for training generative adversarial networks. In: NIPS 2016 Workshop on Adversarial Training. In Review for ICLR, vol. 2016 (2017)
16. Müller, A.: Integral probability metrics and their generating classes of functions. Adv. Appl. Probab. **29**(2), 429–443 (1997)
17. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning, pp. 1060–1069 (2016)
18. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
20. Xu, T., et al.: ATTNGAN: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1316–1324 (2018)

21. Zhang, H., et al.: StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5907–5915 (2017)
22. Zhao, B., Meng, L., Yin, W., Sigal, L.: Image generation from layout. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8584–8593 (2019)
23. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)
24. Huang, Z., Jiqing, W., Van Gool, L.: Manifold-valued image generation with Wasserstein generative adversarial nets. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3886–3893 (2019)
25. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1969–1978 (2019)