SpatialGAN: Progressive Image Generation Based on Spatial Recursive Adversarial Expansion

Lei Zhao*
cszhl@zju.edu.cn
Zhejiang University
Hang Zhou, Zhejiang Province, China

Sihuan Lin
Zhejiang University
Hang Zhou, China
linsh@zju.edu.cn

Ailin Li
Zhejiang University
Hang Zhou, China
linsh@zju.edu.cn

Huaizhong Lin
Zhejiang University
Hang Zhou, China
linhz@zju.edu.cn

Wei Xing†
Zhejiang University
Hang Zhou, China
wxing@zju.edu.cn

Dongming Lu
Zhejiang University
Hang Zhou, China
ldm@zju.edu.cn

ABSTRACT

The image generation model based on generative adversarial networks has recently received significant attention and can produce diverse, sharp, and realistic images. However, generating high-resolution images has long been a challenge. In this paper, we propose a progressive spatial recursive adversarial expansion model(called SpatialGAN) capable of producing high-quality samples of the natural image. Our approach uses a cascade of convolutional networks to progressively generate images in a part-to-whole fashion. At each level of spatial expansion, a separate image-to-image spatial adversarial expansion network (conditional GAN) is recursively trained based on context image generated by previous GAN or CGAN. Unlike other coarse-to-fine generative methods that constraint on generative process either by multi-scale resolution or by hierarchical feature, the SpatialGAN decomposes image space into multiple subspaces and gradually resolves uncertainties in the local-to-whole generative process. The SpatialGAN greatly stabilizes and speeds up the training, which allows us to produce images of high quality. Based on visual Inception Score and Fréchet Inception Distance, we demonstrate that the quality of images generated by SpatialGAN on several typical datasets is better than that of images generated by GANs without cascading and comparative with the state of art methods with cascading.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; **Image representations**.

KEYWORDS

progressive image generation; spatial recursive adversarial expansion; generative adversarial network

ACM Reference Format:
Lei Zhao, Sihuan Lin, Ailin Li, Huaizhong Lin, Wei Xing, and Dongming Lu. 2020. SpatialGAN: Progressive Image Generation Based on Spatial Recursive Adversarial Expansion. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413760

Figure 1: 512×512 images generated by our method using the CELEBA-HQ[41] dataset.

*Corresponding author.
†Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413760

1 INTRODUCTION

Image generative methods that produce novel samples according to high-dimensional data distributions learned from image data set, are being widely used in image synthesis [9, 49, 50], cross-domain image generation [31, 58], image super-resolution [27, 29], image colorization [56] and image inpainting [8, 22, 28, 52, 54, 55]. Currently, the most typical approaches are autoregressive models [40], variational autoencoders (VAE) [26] and generative adversarial networks (GAN) [16]. Each of these methods has its advantages and disadvantages. Autoregressive models are an effective approach to tractably model a joint distribution of the pixels in the image as a product of conditional distributions, which can produce sharp images but are slow to evaluate and do not have a latent representation as they directly model the conditional distribution over pixels, potentially limiting their applicability such as NADE [46], PixelCNN [39]. VAEs are easy to train but tend to produce blurry results due to restrictions in the model, although recent work is improving this [25]. GANs propose a new framework for estimating generative models via an adversarial process, in which two models are simultaneously trained. A generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. GAN can produce diverse, sharp, and realistic images that capture both the fine-grained details and global coherence of images. However, GANs still face many unsolved difficulties: 1) in general, they are very difficult to train, even with many tricks applied in [3, 34, 42, 43]. Balancing the convergence of the discriminator and of the generator is a challenge: frequently, the discriminator wins too easily at the beginning of training [16]. 2) GANs easily suffer from modal collapse, a failure mode in which just one image is learned [13]. 3) The generation of high-resolution images based on GAN is difficult because higher resolution makes it easier to tell the generated images apart from training images [38], thus drastically amplifying the gradient problem. Large resolutions also necessitate using smaller minibatches due to memory constraints, further compromising training stability. Researchers also aim at stretching GAN's limit to generate higher-resolution, photo-realistic images. The higher resolution of the generated image means the higher dimension of the generated image, and its variability of image space increases exponentially with respect to low-resolution images. In order to learn the distribution of the high-resolution image space, the amount of training data needed is large. The distribution learning of the high-resolution images is prone to the phenomenon of underfitting based on existing image datasets. That is to say, the number of images needed to learn the statistical rules of high-resolution images is much larger than that of existing training datasets. Generative networks are easy to learn some features that are not image domain ones, resulting in poor generalization ability. In addition, for the generative adversarial network, the high-resolution image has obvious details, which makes it easy for the discriminator to discriminate the training image and the generated image [38], which will seriously affect the growth of the generator during training. Therefore, the network that generates high-resolution images is prone to unstable training or even unable to converge, and it is also prone to generate many non-sense images. In order to solve the problem of quality and stability in high-resolution image generation, inspired by Collapsed Gibbs sampling [23, 47, 48] and attention mechanism , this paper proposes a spatial recursive generative adversarial network (SpatialGAN). The SpatialGAN first generates the central part of the image, whose spatial dimension (spatial resolution) is relatively small, the variability of the central part content is greatly reduced, which enhances the reliability of the network. The remaining image content is progressively generated by way of recursive fashion. Each spatial extension operates on training data set with the same spatial resolution.

In summary, the contributions of this paper are as follows: 1) This paper proposes a progressive recursive generation model(SpatialGAN), which is different from the progressive model based on different resolutions [41]. The SpatialGAN progressively generates the whole image from the center of the image space.

2) In the process of progressive image generation, the different space scale discriminators are adopted to constrain the intermediate generation process. In this sense, the constraints of our recursive generative process are equivalent to a regularizer, which can stabilize the training of whole networks.

3) The progressive generation method in this paper is similar to the Collapsed Gibbs sampling method. In this sense, the SpatialGAN is the Collapsed Gibbs sampler in high dimensional image space. We fist generate the attention part of the image and then recursively expand the surrounding image space with the previously generated content as the context constraint instead of generating the whole image at once. The SpatialGAN can transform the problem of high-resolution image generation into the problem of multiple low dimensional image generation, which greatly improves image quality (as shown in the Fig.1).

2 RELATED WORK

Generative adversarial networks. Generative adversarial networks (GANs) [16] focus on modeling the natural image distribution by training the deep neural networks to generate samples to be indistinguishable from natural images. GANs have been used in a wide variety of applications such as image generation [4, 57], representation learning [42], image manipulation [24], object detection [29], and video applications [8, 33, 44].

Although GAN and their variants can produce diverse, sharp, and realistic images that capture both the fine-grained details and global coherence of images, there are still three aforementioned problems. Recently a number of more stable alternatives have been proposed, including least squares [32], absolute deviation with margin [57], and Wasserstein distance [4, 17]. In order to generate high-resolution images, Various coarse-to-fine schemes [7] have been proposed [12, 21]. A new training methodology called ProGAN [41] is proposed to grow both the generator and discriminator progressively. ProGAN starts from a low resolution and adds new layers that model increasingly fine details as training progresses. Similar studies include multimodal image style transfer [48, 49]. Inspired by coarse-to-fine method [41], Gibbs sampling [23, 47, 49] and attention mechanism [36, 55], we propose a new progressive image generative method based on spatial recursive adversarial expansion. We first generate the central part of the image (attention part of the image), instead of generating the whole image in once,

and then gradually recursively generate the surrounding image content based on previously generated content.

Image-to-image translation. Many researchers have leveraged adversarial learning for image-to-image translation [42], whose goal is to translate an input image from one domain to another domain given input-output image pairs as training data. Specifically, GANs have proven to be effective means of achieving plausible image-to-image translation results. For instance, pix2pix algorithm [42] uses a GAN conditioned on the source image and imposes an L_1 loss between the generated image and its ground-truth image. This requires the existence of ground-truth paired images from each of the source and target domains. Unpaired image-to-image translation network [24, 58] builds upon pix2pix and removes the paired input data requirement by imposing cycle adversarial constraint, which conserves the overall structure and content of the image.

3 PROGRESSIVE SPATIAL RECURSIVE ADVERSARIAL EXPANSION NETWORK

In this section, we will introduce our model architecture (SpatialGAN) in detail, including overall architecture, and learning schemes, network architecture and loss functions respectively in Sec. 3.1, 3.2 and 3.3.

3.1 Overall Architecture and Learning Schemes

The purpose of our model is to learn a distribution over natural images in stages and generate new images according to learned distribution. For an image space with $n \times n$ pixels, it's probability $p(x)$ can be written as the product of the conditional distribution over pixels:

$$p(x) = \prod_{i=1}^{n^2} p(x_i | x_0, ..., x_{i-1}) \tag{1}$$

The higher the resolution of the image is, the larger the spatial dimension represented by the image is. It is very difficult to model and sample the high spatial dimension, which is the so-called dimension disaster problem in deep learning literature. The higher the spatial resolution of the image is, the more training samples are needed to learn spatial distribution. In terms of the number of samples in the current dataset, the learned distribution can not represent the actual spatial distribution of high-resolution images, resulting in poor generated image quality and other problems. In order to solve this problem, inspired by the collapsed Gibbs sampling, [23, 47, 48] we iteratively generate the whole image from the central part of the image instead of generating the whole one at once, which greatly reduces the spatial dimension of the image we need to learn and generate. So we propose a solution to divide an image into several different non-overlapping image blocks.There are several ways to divide an image into different image patches, inspired by attention mechanism, we will divide the image from the image center. Generally speaking, The central part of an image often contains the most important content of the image, so we call the first patch (central part of the image) attention one and remaining patches spatial expansion ones. The default size of the center patch is 16×16, and the other patches are multiplied by 2, so the size of the patch sequence is 16×16,23×32,64×64,128×128,etc. The joint distribution over image pixels is factorized into a product of conditional patch distributions after the image is divided into many different patches. The probability model of the image is as follows:

$$p(x) = \prod_{i=1}^{m} p(patch(x)_i | patch(x)_1, ..., patch(x)_{i-1}) \tag{2}$$

where m is the number of patches contained in an image, and $patch(x)_i$ denotes i^{th} patch, which contains a series of image pixels. The value $p(patch(x)_i | patch(x)_1, ..., patch(x)_{i-1})$ is joint probability distribution of pixels contained in the i^{th} patch given all the previous patches $patch(x)_1, ..., patch(x)_{i-1}$.

We transform the high-dimensional problem into the low-dimensional one through the aforementioned patch-based image segmentation scheme. For each image patch, we model a GAN or CGAN to learn it's distribution and then concatenate all these network models together to form a complete cascade network model (named SpatialGAN), as shown in Fig.2. Our proposed network is a hierarchical deep convolutional neural network and is comprised of many subnetworks: a basic attention adversarial network (BAAN) G_a and a series of spatial adversarial expansion network (SAEN) G_i, where $i \in 1, 2, ..., N$,N is the number of spatial adversarial expansion networks, related to the number of image patches and is also determined by us. These subnetworks are parameterized by $\theta_0, \theta_1, ..., \theta_j$ respectively (these parameters will be made explicit later), where $j \in 0, 1, ..., N$.A basic attention network (BAAN) can be one of the typical GAN, such as DCGAN [1], BEGAN [47], WGAN-GP [23], which mainly responsible for learning the distribution of attention patches. All spatial adversarial expansion networks will be cascaded together recursively. Each spatial adversarial expansion network G_i will be trained to expand pixel space of the image and take the image patch IP_{i-1} generated by the previous network G_{i-1} as an input and output the image patch IP_i whose size is larger than input one IP_{i-1}. The spatial adversarial expansion Networks take an image patch $patch(G_a)$ (generated by G_a) as an input and is trained to generate multiple output images IP_k of increasing sizes,

$$IP_k = f\left(\bigcup_{j=1}^{k} \theta_j, patch(G_a)\right) \tag{3}$$

where θ_j denotes parameter of j^{th} spatial adversarial expansion network, f denotes function expressed by k^{th} spatial adversarial expansion network. At test time, in order to produce sharp, realistic and, diverse high-resolution images, the SpatialGAN generates image content from local (starting from center of image) to whole. BAAN G_a takes noise z_0 as an input and generates the image patch which will be sent to SAEN connected to it. Each SAEN G_i expands image space based on context content generated by previous SAEN and noise z_i until the whole image is completed. During training, we first train basic attention adversarial network(BAAN) G_a and then train spatial adversarial expansion network (SAEN) G_i in the order of their space independently. We then jointly fine-tune all the networks together, as shown in Fig. 2. In theory, we can generate very large images in stages, as long as the resolution of the training dataset supports it. Note that while we illustrate

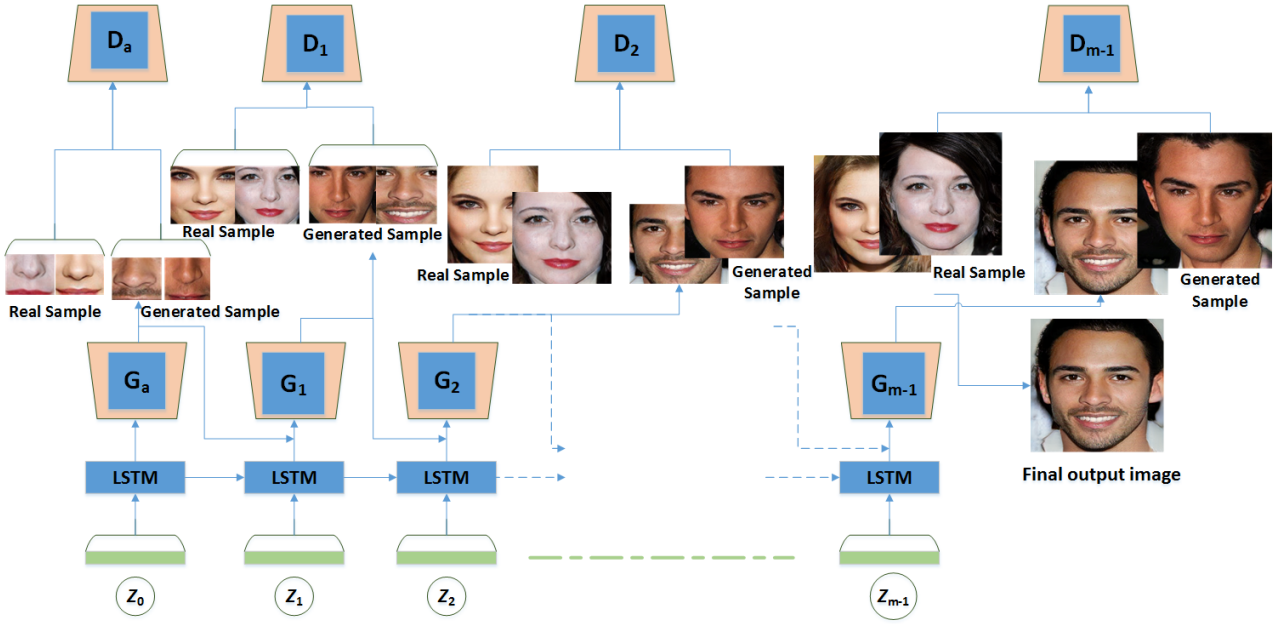

Figure 2: SpatialGAN architecture unfolded to m recursive steps. It mainly consists of one attention GAN Ga and a series of spatial adversarial expansion networks (SAENs). The meaning of each component is explained in the following sections.

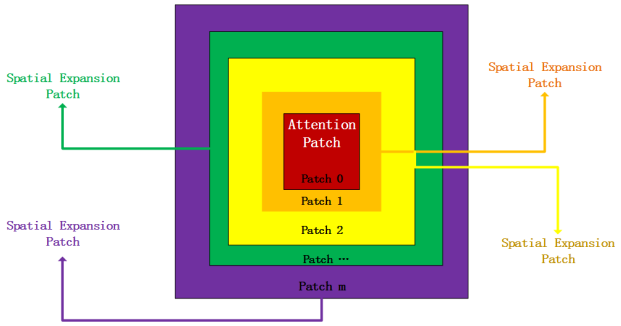

Figure 3: Patch-based image partitioning scheme

the process using a four-level hierarchy, the same concept can be extended recursively to enable the generation of progressively larger images. Similar ideas but different architectures could be found in recent unconditional GANs [12, 21] and conditional image generation [9, 19].

3.2 Network Architecture

3.2.1 ***Basic Attention Adversarial Network (BAAN).*** As mentioned in the sec 3.1, the SpatialGAN transforms the high-dimensional problem into the low-dimensional one through patch-based image segmentation scheme. So we hope that SpatialGAN could first generate the attention content of the image whose size is much smaller than that of the whole image. Since the size of attention content is small, the existing basic GAN can be used to generate realistic attention content very well. We use DCGAN [1] , BEGAN [11] WGAN [17] respectively as basic attention adversarial network to do ablated study. The attention content is very important since all subsequently generated content is based on it as context constraint. The diversity and quality of the attention content generated by BAAN directly determine the diversity and quality of the whole image. In order to improve the diversity and quality of attention content generated by BAAN, we modify the discriminator of current typical GAN. Instead of letting the discriminator distinguish the true and false of an image, we let the discriminator distinguish the true and false of two images concatenated together, which are randomly sampled from the training data set and generated data set, respectively. Zinan [59] had proved that sending multiple images into the discriminator together greatly improves the generated image diversity.

3.2.2 ***Spatial Adversarial Expansion Network (SAEN).*** As mentioned in the sec 3.2.1, the SpatialGAN first generates an image patch (also called attention patch) and then progressively expands image space based on the previously generated patch as context constraint. Through the attention patch is very important since all subsequent generation processes are based on it, the subsequent expansion operations are also a non-trivial factor in determining image quality and diversity. Each spatial adversarial expansion network (SAEN) G_i takes a noise z_i and image patch IP_{i-1} (where IP_0 is $patch(G_a)$ generated by G_a)) as an input and output an image patch IP_i (IP_{m-1} is the final output image of SpatialGAN, where m is the number of patches). The noise z_i is used to increase image diversity. The SAEN's architecture is shown as Fig. 4. It is similar to one proposed by Brock [6] for image generation based on ImageNet dataset. Each SAEN G_i contains two parts. One is residual spatial expansion network (RSEN), which is the main part of the whole

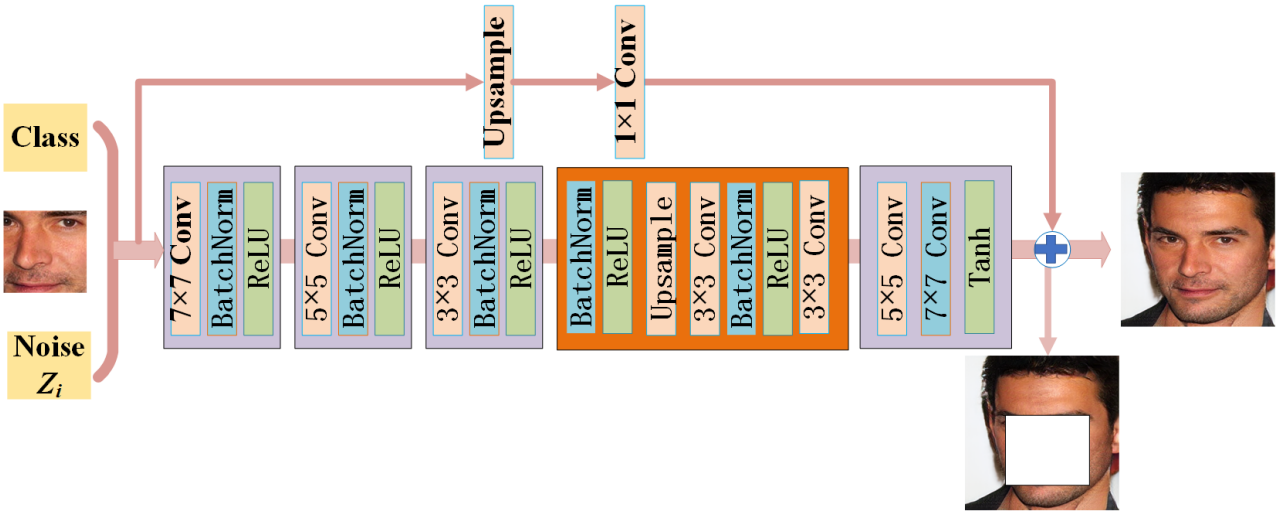

Figure 4: Architecture of the Spatial Adversarial Expansion Network (SAEN)

network, the other is global skip connec- tion, which we refer to as ResContext. The RSEN is ResBlock, which consists of two 3×3 convolution layers, two instance normalization layers [45], and two ReLU [37] activation layers and one Upsample layer. The details of ResBlock are in Fig. 4. Each RSEN learns spatial expansion based context content and output image patch IPR_i, so $IP_i = IP_{i-1} + IPR_i$, where IP_i, IP_{i-1} is the output image patch and the input image patch of SAEN G_i , respectively and IPR_i is output image of RSEN. We find that such SAEN model architecture makes training faster and generalizes better. Through the use of global skip connection of input image, we ensure that the main part of our SAEN network only cares about the generation of surrounding space content based on the contextual content constraint. Through the cascading of SAENs, we can continue to expand in the image space until the content of the whole image is generated. In order to increase the diversity of spatial expansion, we add noise z_i to each spatial expansion network as an input. In order to further improve the quality of the generated image, one self-attention layer [6] is added before the final SAEN. During the training phase, we define a multi-scale critic network, and each critic network is Wasserstein GAN [4] with Gradient Penalty [17] to which we refer as WGAN-GP. In order to improve global and local consistency, the architecture of the critic network is identical to [22].

3.2.3 ***Spatial Connections.*** The SpatialGAN has two kinds of spatial connections – informally speaking, one on 'top' and one on 'bottom'. The 'top' connections perform the act of sequentially spatial expansion. The 'bottom' connections are constructed by a LSTM on the noise vectors $z_0, z_1, \ldots, z_{m-1}$. Intuitively, this noise-vector-LSTM provides information to the spatial adversarial expansion network (SAEN) about what else has been generated in the past. Besides, when there are multiple spatial adversarial networks, we will send the output content generated by the previous network to the next network as an input. In this way, the model is able to 'see' previously generated image content.

Figure 5: images sampled using different generative models. From top to bottom: SpatialGAN-BEGAN, SpatialGAN-WGAN,SpatialGAN-DCGAN with 512×512 and BEGAN [11], WGAN [17], DCGAN [1] with 64×64.

3.3 The Joint Loss Function

As mentioned in sec 3.2. The SpatialGAN consists of many subnetworks, each subnetwork first independently is trained and, then

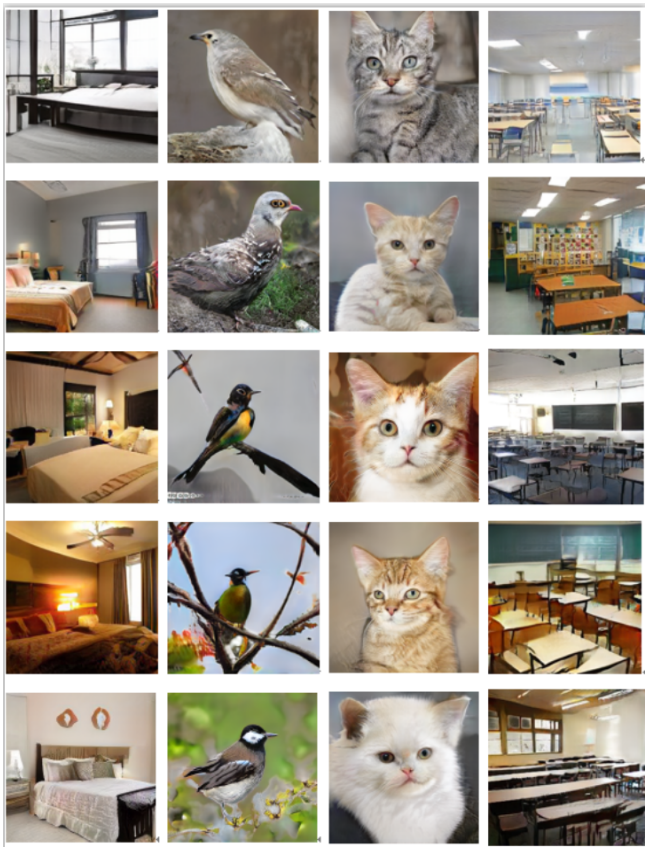

Figure 6: Selection of 256×256 images generated from different LSUN categories. From left to right, they are classroom, bird, cat1 and classroom.

whole SpatialGAN network jointly is trained. The joint loss function is as follows:

$$L = L_{G_a} + \sum_{i=1}^{m} \lambda_i L_{G_i} \tag{4}$$

Where $L_{G_a}$is the loss function of basic attention adversarial network (BAAN), and L_{G_i} is the loss function of spatial adversarial expansion network (SAEN) G_i. $\lambda_i$is weighted hyper parameters of i^{th} SAEN loss. These loss functions are described in detail below.

3.3.1 ***Loss Function of BAAN.*** The basic attention adversarial network may be any one of vanilla GAN, such as DCGAN [1], BEGAN [11], WGAN [17],SAGAN [18]. Our loss function of BAAN is similar to the adopted vanilla GAN. The difference is that our discriminator distinguishes the true and false of two images concatenated together, which are randomly sampled from the training data set and generated data set, respectively.

3.3.2 ***Loss Function of SAEN.*** The SpatialGAN progressively generates high-resolution images by spatial expansion operation. The main body of spatial expansion operation is a series of SAENs. Each SAEN takes as an input noise z_i and context image patch IP_{i-1} and output image patch IP_i. In this sense, the SAEN is similar to image-to-image network [42]. The difference is that the input and output images belong to the same domain and the size of the input image is smaller than that of the output image. In order to implement the spatial expansion, we require SAEN to generate image content in surrounding space with an input image patch as a constraint. The SpatialGAN adopts adversarial loss as the loss function of each SAEN network. In order to ensure the consistency of texture and content of image patch generated by spatial extension network, data sets with the same size are used to adversarial constraints in each SAEN.

High-resolution image synthesis poses a great challenge to the GAN discriminator design. To differentiate high-resolution real and synthesized images, the discriminator needs to have a large receptive field. That would require either a deeper network or larger convolutional kernels. As both choices lead to increased network capacity, overfitting would become more of a concern. Also, both choices require a larger memory footprint for training, which is already a scarce resource for high-resolution image generation. To address the issue, we adopt the multi-scale discriminators proposed in [48]. We use three discriminators that have an identical network structure but operate at different image scales. We will refer to the discriminators as D_1, D_2 and D_3. Specifically, we downsample the real and synthesized high-resolution images by a factor of 2 to create an image pyramid of three scales. The discriminators D_1, D_2 and D_3 are trained to differentiate real and synthesized images at the three different scales, respectively. The discriminators that operate at the coarsest scale have the largest receptive field, a more global view of the image, and can guide the SAENs to generate globally consistent images. The discriminator operating at the finest scale guides the SAENs to produce finer details. In order to generate high-resolution images, SpatialGAN first generates the attention part of images, which results in reducing the diversity of images. In order to improve the diversity of SpatialGAN without losing image quality, our multi-scale discriminators are modified to distinguish the real and generated of two images concatenated together, which are randomly sampled from training data set and generated data set at the three different scales respectively.

$$L_{G_i} = L_i^{adv} = min_{G_i} max_{D_1,D_2,D_3} \sum_{k=1}^{3} L_{GAN}(G_i, D_k) \tag{5}$$

DATASET	FID(ProGAN)[41]	FID (SpatialGAN best)
LSUN-classroom	20.36	18.52
LSUN-bedroom	8.34	7.63
LSUN-cat	37.52	32.54
LSUN-bird	29.91	27.21
CELEBA-HQ	7.30	7.10
CIFAR 10(supervised)	28.49	25.36

Table 2: Fréchet Inception distance (FID), lower is better.

Figure 7: Samples(bird) generated by our SpatialGAN model at 256×256 from LSUN categories.

Figure 8: Samples(classroom) generated by our SpatialGAN model at 256×256 from LSUN categories.

DATASET	FID(SpatialGAN)	FID (SpatialGAN-OPAC)
LSUN-cat	32.54	36.34
LSUN-bird	27.21	28.51
	IS(SpatialGAN)	IS (SpatialGAN-OMD)
CIFAR10	8.89	8.78

Table 3: Ablation experiments on LSUN and CIFAR10.

4 EXPERIMENTS

In this section, we perform experiments on four datasets, including CELEBA, CELEBA-HQ (Tero Karras et al.,2018), CIFAR-10 (A. Krizhevsky et al.,2009), and LSUN. Firstly, we did various ablated experiments for different attention GAN. For ease of explanation, the SpatialGAN based on different vanilla GANs, such as BEGAN [11], WGAN [17], DCGAN [1],SAGAN [6] are called SpatialGAN-BEGAN, SpatialGAN-WGAN, SpatialGAN-DCGAN,SpatialGAN-SAGAN. Fig. 5 shows the images generated by DCGAN, WGAN, BEGAN, respectively, at 64×64 resolution and SpatialGAN-DCGAN, SpatialGAN-WGAN, SpatialGAN-BEGAN at 512×512 resolution respectively on CELEBA and CELEBA-HQ. In order to further improve the quality and diversity of the SpatialGAN, we employ hinge loss [30] GAN objective. At the same time, we provide class information to each SAEN with class-conditional BatchNorm [14] and to discriminator with projection [35]. The optimization settings follow Zhang etal. [18](employing Spectral Norm in G) with the modification that we halve the learning rates and take two D steps per G step. We use Orthogonal Initialization [2] during training SpatialGAN. For each spatial adversarial expansion network G_i. We train them independently according to the loss function given in equation 5. Finally, the joint loss function given in equation 4 is used to fine tune the whole network model to make it globally optimal. We take generating 256×256 image as an example, our model first generates 32×32 image patch, and then recursively generates 64×64, 128×128, 256×256 images. In the experiments, $\lambda_1 = 1.25$, $\lambda_2 = 2$, $\lambda_3 = 3$. To ensure the image quality with large spatial resolution, the larger the network output spatial resolution, the greater the hyper-parameter corresponding to its loss function. The parameters of one spatial adversarial expansion network model is 24.21M. We take generating 256*256 image as an example, our model first generates 32 * 32 image patch, and then recursively generates 64*64, 128*128, 256*256 images. There are 3 spatial adversarial expansion networks. The parameters of all spatial adversarial expansion networks is 72.63M. The parameters of the whole network model is the sum of the parameters of vanilla model and spatial adversarial expansion network models. As can be seen from Fig. 5, Fig. 7, Fig. 8, and Fig. 9, the quality of images generated by our algorithm is better than that of images generated by corresponding attention GAN. In addition, we also find that our model has similar performance with the corresponding original vanilla GAN. For example, WGAN can prevent mode collapse very well, but the quality of the generated image is relatively poor, which is consistent with our SpatialGAN-WGAN model. Through BEGAN could generate high-quality images, the diversity of images is poor, which is consistent with our SpatialGAN-BEGAN model. The Fig. 6 shows the images generated by SpatialGAN on LSUN. We do experiments based on

Figure 9: Samples(cat) generated by our SpatialGAN model at 256×256 from LSUN categories.

Method	IS[42]
Infusion training[5]	4.62± 0.06
ALI[13]	5.34± 0.05
DCGAN[17]	5.40± 0.08
BEGAN[11]	5.62± 0.05
GMAN [15]	6.00± 0.19
EGAN-Ent-VI [10]	7.07± 0.10
LR-GAN [53]	7.17± 0.07
WGAN[1]	7.68± 0.07
Splitting GAN[1]	7.90± 0.09
SGAN[51]	8.59± 0.12
SAGAN [18]	8.61± 0.09
ProGAN[41]	8.80± 0.05
SpatialGAN-DCGAN†	5.52± 0.06
SpatialGAN-BEGAN†	5.70± 0.05
SpatialGAN-WGAN†	5.40± 0.0.10
SpatialGAN-DCGAN$^{\Gamma}$	7.01± 0.05
SpatialGAN-BEGAN$^{\Gamma}$	7.56± 0.06
SpatialGAN-WGAN$^{\Gamma}$	7.89± 0.10
SpatialGAN-SAGAN$^{\Gamma}$	8.60± 0.10
Our Method(best run)	8.89± 0.06
Real data	11.24± 0.12

Table 1: CIFAR10 inception scores, higher is better.
† denotes without multi-scale discriminator.
$^{\Gamma}$ denotes with multi-scale discriminator to distinguish two concatenated real and generated images.

DCGAN without multi-scale discriminator and two-image concatenation, compared with the DCGAN, the IS is improved by 0.9. We do albation study based on SAGAN, our model without multi-scale discriminator is called SpatialGAN-OMD, and the model without two-image concatenation is called SpatialGAN-OPAC. The results are shown in the Table 3.

4.1 Comparison with The State of The Art

Generative adversarial networks lack an objective function, which makes it difficult to compare the performance of different models. We choose the Inception score (IS) [42] and the Fréchet Inception distance (FID) [20] for quantitative evaluation. Higher Inception score indicates better image quality. However, it is important to understand that Inception score has serious limitations—it is intended primarily to ensure that the model generates samples that can be confidently recognized as belonging to a specific class and that the model generates samples from many classes, not necessarily to assess the realism of details or intra-class diversity. FID is a more principled and comprehensive metric and has been shown to be more consistent with human evaluation in assessing the realism and variation of the generated samples [20]. Lower FID values mean closer distances between synthetic and real data distributions. In all our experiments, 50k samples are randomly generated for each model to compute the Inception score and FID.

Here, we compare progressive SpatialGAN with other state-of-the-art generative models on CIFAR-10 and LSUN dataset. The visual quality of generated images is measured by the widely used metric, Inception score [42], and FID [20]. As shown in Tab.1, SpatialGAN obtains a score of 8.89 ±0.06, and is comparable to those of ProGAN [41] (8.80 ±0.05). We report our scores in the mean and standard deviation computed from the highest scores seen during training. Table 1 compares against the previous method in terms of inception scores.

In order to measure the diversity of generated images, we adopt the Fréchet Inception distance (FID) [20]. The FID scores are shown in Table 2 on CIFAR 10 and LSUN.

5 CONCLUSION

Generative adversarial networks are a most promising class of generative models that have so far been held back by unstable training and mode collapse especially when generating high-resolution images. This work presents partial solutions to both of these problems. We propose a coarse-to-fine technique to stabilize training, which transforms high dimensional issues into low dimensional ones. Our model progressively generates the image in a part-to-whole fashion, which greatly reduces the variability of image space compared with the whole image. We apply our technique to the problem of distribution learning, achieving state-of-the-art results on a number of different data sets, such as CELEBA, CELEBA-HQ, CIFAR10, LSUN. We hope to develop a more rigorous theoretical analysis in future work.

Acknowledgments. This work was supported in part by the program (No:2019C03137, LGF18F020006, LY19F020049,2019008), and the key scientific research base for digital conservation of cave temples in Zhejiang university, state administration for cultural heritage of china.

REFERENCES

[1] and Chintala Alec, Luke. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. (2015).

[2] James McClelland Andrew Saxe. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. (2014).

[3] Michael Arbel, Dougal J. Sutherland, Mikołaj Bińkowski, and Arthur Gretton. 2018. On gradient regularizers for MMD GANs. (2018).

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. (2017).

[5] Florian Bordes, Sina Honari, and Pascal Vincent. 2017. Learning to Generate Samples from Noise through Infusion Training. (2017).

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).

[7] Peter J. Burt and Edward H. Adelson. 1987. The Laplacian Pyramid as a Compact Image Code. *Readings in Computer Vision* 31, 4 (1987), 671–679.

[8] Yang Chao, Lu Xin, Lin Zhe, Eli Shechtman, Oliver Wang, and Li Hao. 2016. High-Resolution Image Inpainting Using Multi-scale Neural Patch Synthesis. (2016).

[9] Qifeng Chen and Vladlen Koltun. 2017. Photographic Image Synthesis with Cascaded Refinement Networks. (2017).

[10] Zihang Dai, Amjad Almahairi, Philip Bachman, Eduard Hovy, and Aaron Courville. 2018. Calibrating Energy-based Generative Adversarial Networks. (2018).

[11] Metz david Berthelot, Tom. 2017. BEGAN: Boundary equilibrium generative adversarial networks. (2017).

[12] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. In *International Conference on Neural Information Processing Systems*.

[13] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, and Aaron Courville. 2016. Adversarially Learned Inference. (2016).

[14] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2016. A Learned Representation For Artistic Style. (2016).

[15] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. 2016. Generative Multi-Adversarial Networks. (2016).

[16] Ian J. Goodfellow, Jean Pougetabadie, Mehdi Mirza, Xu Bing, David Wardefarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *Advances in Neural Information Processing Systems* 3 (2014), 2672–2680.

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved Training of Wasserstein GANs. (2017).

[18] Zhang Han, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-Attention Generative Adversarial Networks. (2018).

[19] Zhang Han, Xu Tao, and Hongsheng Li. 2016. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. (2016).

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. (2018).

[21] Xun Huang, Ming Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-Image Translation. (2018).

[22] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. 36, 4 (2017), 1–14.

[23] Mohammad Maminur Islam, Mohammad Khan Al Farabi, and Deepak Venugopal. 2017. Adaptive blocked Gibbs sampling for inference in probabilistic graphical models. In *International Joint Conference on Neural Networks*.

[24] Justin Johnson, Alexandre Alahi, and Li Feifei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution.. In *European Conference on Computer Vision*.

[25] Diederik P Kingma, Tim Salimans, and Max Welling. 2017. Improving Variational Inference with Inverse Autoregressive Flow. (2017).

[26] Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. (2013).

[27] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2016. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. (2016).

[28] Tero Karras Timo Aila Samuli Laine Jaakko Lehtinen. 2016. Context Encoders: Feature Learning by Inpainting. In *IEEE Conference on iclr*.

[29] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Perceptual Generative Adversarial Networks for Small Object Detection. (2017).

[30] Jae Hyun Lim and Jong Chul Ye. 2017. Geometric GAN. (2017).

[31] Ming Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. (2017).

[32] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Wang Zhen, and Stephen Paul Smolley. 2016. Least Squares Generative Adversarial Networks. (2016).

[33] Michael Mathieu, Camille Couprie, and Yann Lecun. 2015. Deep multi-scale video prediction beyond mean square error. (2015).

[34] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. (2018).

[35] Takeru Miyato and Masanori Koyama. 2018. cGANs with Projection Discriminator. (2018).

[36] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. 3 (2014).

[37] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *International Conference on International Conference on Machine Learning*.

[38] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2016. Conditional Image Synthesis With Auxiliary Classifier GANs. (2016).

[39] Aaron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel Recurrent Neural Networks. (2016).

[40] Aaron Van Den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. (2016).

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. PROGRESSIVE GROWING OF GANS FOR IMPROVED QUALITY , STABILITY , AND VARIATION. In *IEEE Conference on Computer Vision Pattern Recognition*.

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Chen Xi. 2016. Improved Techniques for Training GANs. (2016).

[43] Che Tong, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. 2016. Mode Regularized Generative Adversarial Networks. (2016).

[44] Sergey Tulyakov, Ming Yu Liu, Xiaodong Yang, and Jan Kautz. 2017. MoCoGAN: Decomposing Motion and Content for Video Generation. (2017).

[45] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. (2016).

[46] Benigno Uria, Marc Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. 2016. Neural Autoregressive Distribution Estimation. *Journal of Machine Learning Research* 17, 1 (2016), 7184–7220.

[47] Deepak Venugopal and Vibhav Gogate. 2013. Dynamic Blocking and Collapsing for Gibbs Sampling. *Computer Science* (2013).

[48] Tongzhou Wang, Wu Yi, David A. Moore, and Stuart J. Russell. 2017. Neural Block Sampling. (2017).

[49] Ting Chun Wang, Ming Yu Liu, Jun Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2017. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. (2017).

[50] Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan Fang Wang. 2017. Multimodal Transfer: A Hierarchical Deep Convolutional Neural Network for Fast Artistic Style Transfer. (2017).

[51] etc Xun. 2017. Stacked Generative Adversarial Networks. In *International Conference on International Conference on Machine Learning*.

[52] Zhaoyi Yan, Xiaoming Li, Li Mu, Wangmeng Zuo, and Shiguang Shan. 2018. Shift-Net: Image Inpainting via Deep Feature Rearrangement. (2018).

[53] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. 2017. LR-GAN: Layered Recursive Generative Adversarial Networks for Image Generation. (2017).

[54] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawajohnson, and Minh N. Do. 2016. Semantic Image Inpainting with Deep Generative Models. (2016).

[55] Jiahui Yu, Lin Zhe, Jimei Yang, Xiaohui Shen, Lu Xin, and Thomas S. Huang. 2018. Generative Image Inpainting with Contextual Attention. (2018).

[56] Cao Yun, Zhiming Zhou, Weinan Zhang, and Yu Yong. 2017. Unsupervised Diverse Colorization via Generative Adversarial Networks. (2017).

[57] Junbo Zhao, Michael Mathieu, and Yann Lecun. 2017. Energy-based Generative Adversarial Network. (2017).

[58] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision*.

[59] Fanti Zinan, Khetan. 2018. PacGAN: The power of two samples in generative adversarial networks. (2018).