Pattern Recognition 123 (2022) 108375

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

GL-GAN: Adaptive global and local bilevel optimization for generative adversarial network

Ying Liu[a,b], Heng Fan[c], Xiaohui Yuan[c], Jinhai Xiang[a,b,*]

[a] College of Informatics, Huazhong Agricultural University, Wuhan, 430070, China
[b] Hubei Engineering Technology Research Center of Agricultural Big Data, Huazhong Agricultural University, Wuhan, 430070, China
[c] Department of Computer Science and Engineering, University of North Texas, Denton, Texas 76203, USA

ARTICLE INFO

Article history:
Received 18 November 2020
Revised 9 September 2021
Accepted 15 October 2021
Available online 20 October 2021

Keywords:
Generative adversarial networks (GAN)
Global and local bilevel optimization
Ada-OP
Image generation

ABSTRACT

Although Generative Adversarial Networks (GAN) have shown remarkable performance in image generation, there exist some challenges in instability and convergence speed. During the training, the results of some models display the imbalances of quality within a generated image, in which some defective parts appear compared with other regions. Different from general single global optimization methods, we introduce an adaptive global and local bilevel optimization model (GL-GAN). The model achieves the generation of high-resolution images in a complementary and promoting way, where global optimization is to optimize the whole images and local is only to optimize the low-quality areas. Based on DCGAN, GL-GAN is able to effectively avoid the nature of imbalance by local bilevel optimization, which is accomplished by first locating low-quality areas and then optimizing them. Moreover, through feature map cues from discriminator output, we propose the adaptive local and global optimization method (Ada-OP) for interactive optimization and observe that it boosts the convergence speed. Compared with the current GAN methods, our model has shown impressive performance on CelebA, Oxford Flowers, CelebA-HQ and LSUN datasets.

1. Introduction

Generative adversarial networks [1] are a powerful class of image generation compared with VAE [2] and flow models [3]. However, they may suffer from several issues such as model collapse, non-convergence and instability. Recently, in order to address above mentioned challenges, several solutions have been proposed based on the new network architectures [4–6] and the implementation of stability techniques [7–9]. Most solutions employ single global optimization to generate image and neglect the local parts, which causes some patches of image are low quality. Self-Attention GAN(SAGAN) [10] incorporates a self-attention mechanism into the design of both networks in order to capture local and global dependencies of the target distribution. In this work, we propose an adaptive global and local bilevel optimization to balance the global and local distribution of the target.

1.1. Problems and motivations

Generally, the synthetic image quality is variable. From the lower quality synthetic images, we observe that some images show quality imbalance performance within a sample [4,7,11]. In short, the other areas of the generated image show impressive performance except for some small range poorly portions(We call it the phenomenon of low quality images). Furthermore, this phenomenon exists during the training in some GAN model(see Fig 1). We realize that this phenomenon is an obstacle to generating high-quality images and keeping optimization stability. And it is also one of the factors for the decrease in convergence speed. On the other hand, the low training efficiency has always been a challenge in the GAN field [12,13]. Pointing to the quality unbalance issue, the paper makes an intuitive analysis: the reason is that the existing models generally use a single measurement standard (known as the global optimization mode) to evaluate the quality of the whole image, so it is relatively difficult to optimize some details. For example, some models [14,15] excel at synthesizing images with global structure (e.g., image outline, the location of eyes and the hairstyle in face); while it fails to focus on some details, such as artifact, distorted and uncoordinated regions. In practice, this in turn also can be a reason why some early GAN models [4,7,11] only can generate relatively low-quality images. Some of the later models [12,13,16] well leverage the structural superiority to pay more attention to small low-quality areas by increasing structural complexity but it is at the cost of low computational efficiency. Thus, there is a trade-off between high-quality image

* Corresponding author.
E-mail address: jimmy_xiang@mail.hzau.edu.cn (J. Xiang).

https://doi.org/10.1016/j.patcog.2021.108375
0031-3203/

Fig. 1. Location about low-quality areas of generated images on CelebA-HQ256 dataset. The color from blue to red indicates that the quality of region is from good to bad. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

generation and high training efficiency. It is of great significance in the field of image generation to improve training efficiency on the premise of ensuring the quality of image generation.

1.2. Contributions

In this work, we propose an adaptive global and local bilevel optimization model(GL-GAN), which applies the local bilevel optimization model with a traditional global optimization model. The local bilevel optimization model is only to optimize the poor quality portion in a image and the global model optimizes the whole image. Local optimization and global optimization are carried out interactively, hence the quality of local and global keeps balance and the training efficiency is also accelerated. Inspired by PatchGAN [15], the feature map(it can also be called feature matrix) is treated as the discriminator's output for subsequent filtering of low-quality areas. Different from PatchGAN with small size of the feature map, the feature map with a larger size is adopted in our paper, so that the measured receptive field is smaller and more refined. According to the feature map, the quality measure of each area is obtained in an image. Therefore the local bilevel optimization model can be regarded as a reliable guiding strategy for generator optimization by accurately capturing the low-quality areas of the sample firstly and then only optimizing the captured areas by updating the generator parameters. The local optimization model realizes the refinement of the generated region inside the image, thus greatly improving the generated performance. To stabilize training, we investigate the spectral normalization and apply the local-norm to our model.

In addition, we conduct the adaptive global and local bilevel optimization method(Ada-OP) based GL-GAN which coordinately optimizes the whole and local of the image. In the paper, the standard deviation between elements within the feature matrix is used as an index to measure local quality balance nature in an image(we call it the local standard deviation). We argue the mean of elements within a feature matrix as the whole quality measure. Then the quality standard deviation between images is regarded as global quality merit(we call it the global standard deviation). When the global standard deviation is greater than a certain value (this indicates that there is a wide range unbalance area), the global optimization is performed, otherwise, the local bilevel optimization is enforced. And the region scope of local optimization is divided into three levels according to the specific value. The approach is inspired by the painter's painting skills, i.e., a rough description of the whole is presented firstly, and then the local refinement is finished. Through Ada-OP, high-quality images are generated in a complementary and promoting way and the computational efficiency is greatly improved, which provides an approach to realize the balance between image quality and training efficiency.

Extensive experiments have been conducted on CelebA, CelebA-HQ (see Fig 2) and LSUN datasets, which validate the effectiveness of GL-GAN in the generation of high-quality images and the improvement of training efficiency. During applying Ada-OP, local and global optimization is implemented adaptively to perfect local and whole quality of images respectively, which improves the speed of convergence and reduces the cost of training. The ablation experiment is carried out on CelebA dataset for our model. Source code can be found at https://github.com/summar6/GL-GAN.

1.3. Organization

The remainder of this paper is organized as follows. In Section 2 briefly introduces the related works. Then, in Section 3, we describe the GL-GAN model. The experiments are shown in Section 4. Finally, Section 5 draws the conclusions.

2. Related work

2.1. Generative adversarial networks

GANs [1] have become a central part of generation tasks compared with other models. In the early development of GAN, most researchers explored various types of methods to further improve the generation quality and training efficiency. In terms of net architecture, DCGAN [4] converts the full connection layer into the convolution layer, which greatly improves the generation performance. VAE/GAN [5] combines the Variational Autoencoder with the generative adversarial net, in which the learning characteristic representation in the GAN discriminator is used as the basis for VAE reconstruction goal. In loss function, F-GAN [17] constructs various loss function by the general F divergence. LS-GAN [18] adapts the least square loss function as object to optimize model. CGAN [11] and InfoGAN [19] both use conditional information as input to realize accurate learning. Iizuka et al. [20] use global and local context discriminators to train the image completion network for inpainting. Loss-Sensitive GAN [21] trains a generator to create more realistic images by minimizing the boundaries between the real and fake samples.

In the aspect of maintaining stability, the main purpose is to stabilize training by ensuring models' Lipschitz continuity, which motivates the development of weight clipping [7], gradient penalty [8] and spectral normalization [9]. Meanwhile, WGAN-QC [22] proposes an optimal transport regulator (OTR) based on the theory of secondary transport cost to stabilize training. Through the analysis of Dirac-GAN, [23] shows the necessity of absolute continuity for convergence. Al-Dujaili et al. analyse the effect of co-evolution for understanding and improving the gradient-based learning dynamics [24]. CR-GAN [25] introduces a consistency regularization techniques, which augments images with semantic-preserving augmentations and penalizes the sensitivity of the discriminator. Furthermore, ICR-GAN [26] applies forms of consistency regularization to the generated images, the latent vector space, and the generator to achieve further performance gains. The above model is great in training efficiency, but there is still much margin for improvement in image generation quality.

Recently, some models greatly improve the image quality with the cost of computing source. The Pix2pixHD model [27] gen-

Fig. 2. Generated images on CelebA-HQ256 dataset by GL-GAN.

erates high-resolution photo-realistic from semantic label maps using conditional GANs. The model includes a new multi-scale generator and discriminator architectures with a novel adversarial loss. BigGAN [16], which adapts the orthogonal regularization method and timely truncates the input prior distribution, greatly improves GAN's generating performance. Progressive GAN [13] and Style-GAN [12] adapt progressive growth method to train GAN through layer-by-layer to generate high-resolution images. MSG-GAN [28] maintains the overlap in the supports of the real and fake distributions by allowing the flow of gradients from the discriminator to the generator at multiple scales, which provides a stable approach for high resolution image synthesis. Jinlin Liu et al. manage to generate small codes in latent space instead of large images and adopt two-stage way to generate high-resolution image [29]. IP-GAN [30] employs a cascading rejection (CR) module for discriminator so as guiding the generator effectively. StyleALAE [31] proposes adversarial latent autoencoder to address entanglement in a adversarial way and achieve good performance. In images completion task, Qiang Wang et al. [32] incorporate deep generative adversarial networks with a Laplacian pyramid mechanism to recover the spatial information of missing face regions in a coarse-to-fine manner. Uras Mutlu et al. [33] present an encoder working in the inverse direction of the generator to provide auxiliary reconstruction losses as hints for a better generator. CG-GAN [34] applies generative and evolutionary computation to allow casual users to interactively breed and edit faces. Guo et al. investigate the perturbation on the generator's input and develop a smooth generator to generate stable and high-quality images [35]. MGO-GAN [36] learns a mapping function parameterized by multiple generators to overcome mode collapse.

2.2. Feature representation in GAN

Feature map has been a very concerned concept that can capture a specific feature(e.g.,style, figure, location) in an image. Pix2Pix [15] proposes the PatchGAN, which uses feature map to measure quality of region within a image. Several early models mainly apply the feature map information to achieve style transfer, which is realized by the loss function [37,38]. StarGAN [14] and DRPAN [39] both take the feature map information as an area's quality measure in a sample, which helps with generating high-quality details. SAGAN [10] learns to efficiently find global, long-range dependencies within the feature map of images through the self-attention mechanism. Nevertheless, there are some limits in models, including applying the feature map of small size and large computational cost. In this work, we regard the local information within the feature map as a basis to process the adaptive global and local bilevel optimization.

3. Adaptive global and local bilevel optimization GAN

Most GAN-based models aim at measuring the whole image quality by the global optimization method, which is mainly implemented by the output probability of discriminator. Global optimization mode roughly focuses on the quality of the overall area by a single value, thus it is nontrivial for some local details within an image to fine modify.

In this section, we propose an adaptive global and local bilevel optimization model, which can guide the generator to reasonably modify parameters based on the max-min two-player game of GAN. Although the proposed method inherits from the idea of PatchGAN, GL-GAN adopts a larger size feature map instead of the small size, so that the measured receptive field area is smaller and more refined. Moreover, we employ global and local bilevel optimization to training the model. The approach can balance the global and local distribution of the target.

3.1. Feature map

By analyzing some samples generated by GANs models, we observe that there are always some low-quality areas within an image. To explicitly illustrate this phenomenon, the hot maps about low-quality areas are presented (see Fig 1), where the red areas denote the low-quality regions. We can observe that the red areas contain artifacts, distorted and poor quality regions of images. And the other areas is perfect in performance. Therefore, we argue that the quality distribution within a synthetic image is unbalanced.

Considering the single output of discriminator as the whole image quality metric in some models, we regard that using the output obtained by a patch model to represent the receptive field level quality metric within an image is also reasonable. The patch model inherits from the idea of PatchGAN in [15]. To capture the small range low-quality details, we establish a discriminator model, whose output is a feature matrix:

$$y_{h\times w} = D_{\varphi}(x) \tag{1}$$

Where D_{φ} represents the discriminator with parameter φ and x is a sample. The output of discriminator $y_{h\times w}\in R^{h\times w}$ is a matrix, where every element in the matrix corresponds to a receptive field of an image. Specifically, $y_{i,j}$, which is an element in matrix $y_{h\times w}$, denotes the quality evaluation of the i-th row and the j-th column receptive field. Different from the PatchGAN with small size feature map, our output size is larger, which benefits measuring the receptive field in a smaller area and making the optimization more refined.

The adaptive global and local bilevel optimization model is to optimize the generator parameters in the global and local perspectives. In order to carry out the method, we first build the assessment model of receptive field level for discriminator:

$$\max_{\varphi} E_{x\sim P_d}[f(D_{\varphi}(x))] - E_{G_{\theta}(z)\sim P_g}[f(D_{\varphi}(G_{\theta}(z)))] \tag{2}$$

Where P_d, P_g respectively denote the distribution of real samples and generated samples. G_{θ} denotes the generator with parameter θ. $f: R^d \longrightarrow R$ is an operation function to the output of the discriminator, such as the linear function and the nonlinear function, which is utilized to get various types of loss functions, where

we choose the hinge-loss among all the optimization formulas. The model target is the same as the original model, which is to distinguish the generated images from the real images. Hinge-loss enables the generated image to be as separate from the real image as possible. There is an interval between the two. Ideally, the real image and the generated image are located on either side of the interval to facilitate image separation and avoid gradient crash due to the large values. According to the above object (2), hinge-loss is represented by:

$$\min_{\varphi} E_{x_{\sim P_d}}[max(0, 1 - D_\varphi(x))] + E_{G_\theta(z)\sim P_g}[max(0, 1 + D_\varphi(G_\theta(z)))] \tag{3}$$

During training discriminator, the value $D_\varphi(x)$ should be greater than or equal to 1 as far as possible when $x \sim P_d$. When $x \sim P_g$, the value $D_\varphi(x)$ should be smaller than or equal to -1. So that the loss as small as possible, and model can effectively distinguish the generated image and the real image.

3.2. Local bilevel optimization model

Based on the feature map mentioned in Section 3.1, we construct a local bilevel optimization model, which optimizes the local low-quality areas of the generated image by the bilevel method. The local bilevel optimization model first selects the low-quality region and then optimizes the generator parameters about the low quality region. The defective regions are first captured by dot multiplying the output of discriminator with a mask matrix m^*, where the mask matrix(it is composed of 0 and 1) is obtained by inner level optimization (see Fig 4). Then the generator (G_θ) optimizes the low-quality areas by using gradient descent algorithm.

$$\begin{aligned} Object1 = &\min_{\theta} E_{z\sim P_z}[max(0, 1 - m^* \odot (D_\varphi(G_\theta(z))))] \\ &s.t.\ m^* \in arg\max_{m\in M} \sum_{h,w} (m \odot (\alpha - D_\varphi(G_\theta(z)))) \end{aligned} \tag{4}$$

The objective of inner layer in (4) is to select an optimal mask matrix m^* when other parameters θ, φ are fixed. To simplify the choice of mask matrix m, we empirically design a constant α as the criteria for evaluating quality (ideally, when the values from the output are lower than the criteria, the values in mask matrix corresponding to the same position are 1, and vice versa the values are 0). The constant α is determined by the standard deviation of output, it depends on the data distribution of $D_\varphi(G_\theta(z))$ and the scope of the low quality area. When the range of low quality area is large, α should be a larger value, so the output values $D_\varphi(G_\theta(z))_{h,w}, 0 \le h, w < N$ which are smaller than α are selected as low-quality areas. Where $M = \{m_1, m_2, \cdots, m_n, \cdots\}$ is a matrix set, in which each matrix has the same size with the discriminator's output. The $\odot$ denotes the dot product.

The objective of the outer layer is directly to optimize the generator parameters θ by using gradient descent algorithm within the defective receptive fields, which is selected by dot multiplying the output of discriminator with the optimal mask matrix m^*.

3.3. Adaptive global and local optimization method(Ada-OP)

Adaptive global and local optimization method generates high-quality images by adaptive conducting global optimization (which focuses on the whole image as the optimizing objective) and local optimization (which only optimizes the low-quality areas within an image) during training generator. The local optimization model has been shown in previous section. To clarify the method, we first build the global optimization model about the generator.

$$Object2 = \min_{\theta} E_{z\sim P_z}[max(0, 1 - D_\varphi(G_\theta(z)))] \tag{5}$$

In the training process, different extent quality differences between receptive fields and the quality differences between images both will affect the selection of optimization mode. So it is necessary to define some metrics that be able to measure global or local differences in quality. We choose the mean(σ_l)(in a batch) of the quality standard deviation between different receptive fields within an image to measure the local differences. And the quality standard deviation(σ_g) between images is selected to measure the global difference, where we regard the quality mean(μ_k) between different receptive fields within an image as the whole image quality evaluation.

The global standard deviation σ_g can be used as the standard to measure whether it needs to conduct the global optimization or local optimization. Because if σ_g is high, there is a large range of quality differences between images, so global optimization should be carried out. Otherwise the local optimization is performed. The intuition that we do this is: rough image first is generated (the overall quality should be basically the same), then the details are optimized. The evaluation metric for global optimization is as follows:

$$\begin{aligned} &\mu_k = \frac{\sum_{i,j} y_{i,j}}{h \cdot w}, \mu = \frac{\sum_{k=1}^{K} \mu_k}{K} \\ &\sigma_g = \sqrt{\frac{\sum_{k=1}^{K} (\mu_k - \mu)^2}{K}} \end{aligned} \tag{6}$$

Where K is the batch-size of images and μ_k represents the quality of the k-th image. h, w denote the height and width of feature matrix. μ denotes the quality mean of all K images. σ_g is the global standard deviation of all k images. Where β is a constant, we get it through the statistic of σ_g. According to the threshold selection of σ_g, the global standard deviation σ_g of input image for each epoch is recorded during global optimization training, and then the value of $P_{value} = 0.7$ (set as β) is selected as the threshold. When $\sigma_g \ge \beta$, global optimization is executed, otherwise local optimization is carried out.

When implementing local optimization, the choice of the mask matrix depends on the size of unrealistic regions. Thus we divide the level of local bilevel optimization mode into I, II and III, and the corresponding constant α in formula (4) are respectively $\alpha_1, \alpha_2, \alpha_3$. The higher level ($III$) defines the larger local optimization scope, so the constant α in formula (4) is bigger. The specific local standard deviation is as follows:

$$\sigma_k = \frac{\sum_{i,j} (y_{i,j} - \mu_k)^2}{h \cdot w}, \sigma_l = \sqrt{\frac{\sum_{k=1}^{K} \sigma_k}{K}} \tag{7}$$

Where σ_k is the quality standard deviation of all receptive fields in the k-th image. σ_k is utilized for measuring whether the quality of local areas in a single image is balanced. For K images, the mean of the quality standard deviation between the receptive fields is used as the local variance σ_l. When the local variance σ_l is large, it indicates that there is a great difference in quality within a single image, so the optimization area should be larger, and vice versa. Owing to the mean of the different standard deviation corresponding to different local scope, the larger σ_l means larger internal difference, so the level is higher.

The threshold selection of local standard deviation σ_l is similar to the global, and the local standard deviation of each epoch data during global optimization training is recorded, and then the value of $P_{value} = 0.4$ (set as δ_1) and $P_{value} = 0.7$ (set as δ_2) are selected as the threshold.

On the basis of the above definition, the specific local and global optimization method can be given as follows:

$$Obj = \begin{cases} Object2 & \text{if } \sigma_g \ge \beta \\ Object1, where\ \alpha = \alpha_1 & \text{if } \sigma_l \le \delta_1 \\ Object1, where\ \alpha = \alpha_2 & \text{if } \delta_1 < \sigma_l < \delta_2 \\ Object1, where\ \alpha = \alpha_3 & \text{if } \sigma_l \ge \delta_2 \end{cases} \tag{8}$$

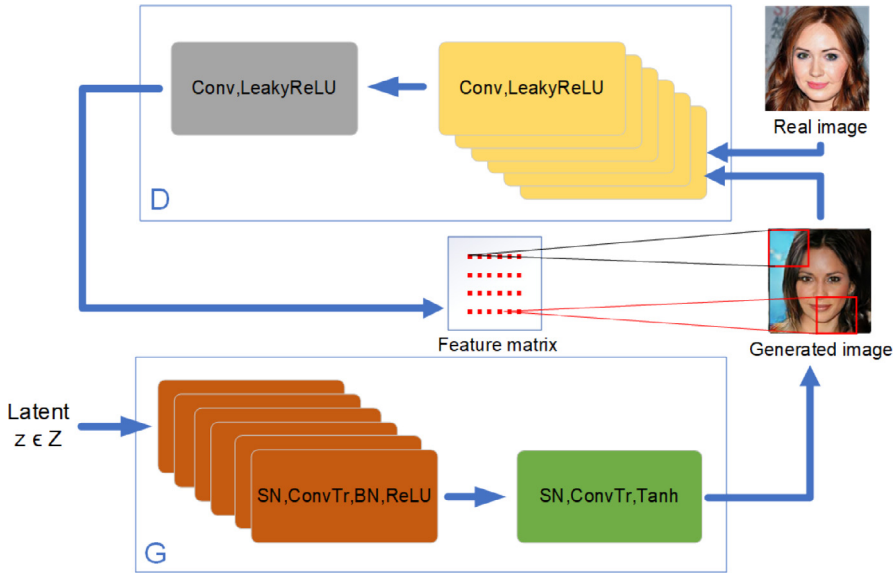


Fig. 3. Architecture of GL-GAN, where the red box is the receptive field corresponding to the element in feature matrix. The network on the top is the generator G, and the network on the down is the discriminator D. Where SN, ConvTr, BN, Conv denote spectral normalization, ConvTranspose, Batch Normalization, Convolution respectively. The generator network includes six layers(SN, ConvTr, BN, ReLu) and one layer(SN, ConvTr, Tanh). The discriminator network includes six layers(Conv, LeakyReLu) and one layer(Conv, LeakyReLu) which is employed to obtain the feature matrix($4 \times 4, 8 \times 8$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Where δ_1, δ_2 are the threshold of σ_l and both constants to divide the local standard deviation into different limits. In the paper, local optimization is divided into three local optimization modes, I: when $\sigma_l \leq \delta_1$, the local standard deviation is low, and the value of $P_{value} = 0.2$ in the feature matrix data is selected as α, which is set as α_1. II: when $\delta_1 < \sigma_l < \delta_2$, the value of $P_{value} = 0.5$ in the feature matrix data is selected as α, which is set as α_2. III: when $\sigma_l \geq \delta_2$, the local standard deviation is large, and the value of $P_{value} = 0.8$ in the feature matrix data is selected as α, which is set as α_3. Based on the method, the generator performs adaptive global and local bilevel optimization. Finally, we give the Algorithm 1 to clarify the training process.

Algorithm 1 GL-GAN.

Input: Real data X, batch-size m, epoch n, k_G. Adam parameters, α, β_1, β_2
Output: G_θ, D_φ

for $i = 0$ to n **do**
  Sample $\{x_i\}_{i \in I} \sim P_d$ for real data.
  Sample $\{z_j\}_{j \in J} \sim P_z$ for random noise.
  Let $y_j = G_\theta(z_j), \forall j \in J$.
  $g_\varphi \leftarrow$ *the gradient of* (2).
  $\varphi \leftarrow Adam(g_\varphi, \varphi, \alpha, \beta_1, \beta_2)$.
  for $t = 0$ to k_G **do**
    $output = D_\varphi(G_\theta(z_j))$.
    Compute σ_g, σ_l according to Eq. (6),(7).
    Applying the global and local bilevel optimization model according to Eq. (8)
    $g_\theta \leftarrow$ *the gradient of* (8).
    $\theta \leftarrow Adam(g_\theta, \theta, \alpha, \beta_1, \beta_2)$.
  end for
end for

4. Experiments

To evaluate the effectiveness of the proposed GL-GAN, extensive experiments on CelebA, Oxford Flowers, CelebA-HQ and LSUN church datasets are conducted. In this section, we firstly present experimental datasets and the evaluative criteria. Then implementation details consisting of model structure and some parameter settings are shown. Next we present the results of GL-GAN numerically and visually and the comparison to state-of-the-arts. Finally, the ablation experiments investigate the role of feature map, spectral normalization and the adaptive global and local bilevel optimal method(Ada-OP).

4.1. Datasets and evaluative criteria

We evaluate the proposed model using face datasets. Scenario datasets are also used to demonstrate the wide applicability of the GL-GAN.

CelebA CelebA [40] is a large face properties dataset with 202,599 celebrity images. In the dataset, each image has 40 attribute annotations and 5 landmark locations, which can be used for attribute editing and face detection. And the size of the images is 178×218. In the paper, the size of images is resized to 128×128 for training and 64×64 for comparison.

Oxford Flowers Oxford Flowers [41] is a flowers dataset which has approx 8K images of 102 different categories. In the dataset, each class consists of between 40 and 250 images. In the paper, the size of images is resized to 256×256 for training and comparison.

Celeba-HQ Celeba-HQ [13] is a high-resolution face dataset which is obtained based on CelebA. In our model, we choose the size of 256×256 and 512×512 face images as training set. Each resolution has $30K$ images.

LSUN LSUN [42] is a high-resolution images dataset of 10 scenarios, which includes bedroom, bridge, church, living-room scenes et al. The church dataset is used for training in our method. There were 7,907 images in the dataset (the sum of training set and testing set), which is cropped and resized to 256×256 by bicubic interpolation.

Evaluative Criteria Frechet Inception Distance is chosen (FID) [43] for quantitative evaluation. The distance between the real images and the synthetic images at the feature level is calculated as a measure. So the smaller FID means the higher quality. FID shows the more consistent with human evaluation in the realism and variation of the generated images.

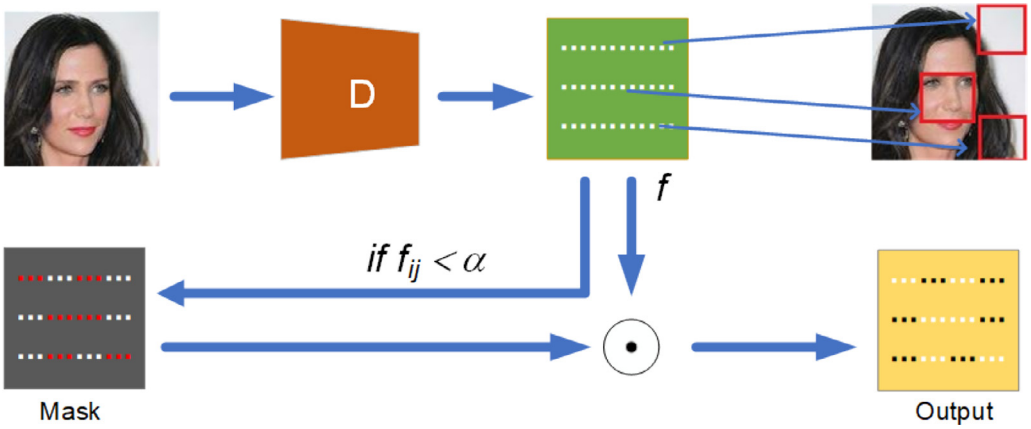


Fig. 4. The process of local bilevel optimization method, where f is the feature matrix of discriminator output. The red point denotes the element is smaller than α in feature matrix, so the value is 1 and the black point is 0. The local optimized matrix is obtained by dot multiplying the Mask with the feature map f, where the white dot represented the retained measurement value of the low-quality area, and the black dot is 0. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 5. Randomly generated high-resolution images by GL-GAN method on LSUN church dataset.

4.2. Implementation details

The model of **GL-GAN** is put forward on the basis of **DCGAN**. The model is effective to boost convergence and improve image performance with a lightweight network. Generator(G) is composed of a series of basic unit, whose architecture is ConvTranspose-BatchNorm-ReLU and 7 units in total for the 256×256 generated images. The input is a random vector of size 100×1 from the standard normal distribution. Except for the ConvTranspose of first layer with parameters (4,1,1) for the kernel size, stride, pooling, all other layers are (4,2,1). Discriminator(D) is the same as the normal classifier with Convolution-LeakyReLU as a basic unit layer, with (4,2,1) for the kernel size, stride, pool of convolution layer in units. The input is generated image and real image and there is 6 units in total and the last layer(Convolution with (1,3,1,1) for output channel,kernel size,strid,pooling) is employed to obtain the feature matrix($4 \times 4, 8 \times 8$). The specific network architecture can be referred to Fig 3. Meanwhile, the parameters of the generator and discriminator respectively are carried out whole and local spectral normalization to stabilize training.

Except for the WGAN-QP(64×64) model (which is trained on the NVIDIA TITAN Xp and has higher operating efficiency than NVIDIA TENSLA V100), CR-GAN and ICR-GAN (the data from the original article), all experiments are conducted on the same NVIDIA TESLA V100. GL-GAN adapts the Hinge-loss function and Adam optimization method with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. By default, the learning rate for generator is 0.0004 and for discriminator is 0.0001 in CelebA. The learning rate is kept unchanged during training with 1:1 balanced updates for the discriminator and generator. Except the CelebA dataset's batch-size is 64, the rest is all of 16. In addition to 30 epochs trained on CelebA, the rest are trained with 70 epochs.

Table 1
Running time(where hrs denotes hours) and FID comparison between different models in CelebA datasets about 128×128 and 64×64 resolution.

Method	FID(128)	time(hrs)	FID(64)	time(hrs)
WGAN-GP	30.37	47.52	25.47	15.12
SNDCGAN	28.53	18.96	12.28	7.92
SAGAN	38.51	33.84	46.87	12.96
WGAN-QC	16.33	22.08	12.9	14.4
CR-GAN	16.97	-	-	-
ICR-GAN	15.43	-	-	-
IP-GAN	-	-	10.15	-
OURS	**12.37**	**12.24**	**8.3**	**5.04**

4.3. Comparison to state-of-the-arts

GL-GAN is applied to CelebA, Oxford Flowers, CelebA-HQ and LSUN church datasets to show the effectiveness of our method. The images randomly generated with the model on CelebA, Oxford Flowers and CelebA-HQ datasets are shown in Fig 2, Fig 7, Fig 6, and Fig 5 presents the generated images on LSUN church. From the perspective of qualitative analysis, it can be seen that the GL-GAN can achieve a relatively outstanding effect on high-resolution images whether the resolution is 64×64 or 512×512. Although there are still some minor flaws in the image, our method is the first to generate high-resolution images via a lightweight model.

In addition, the implementation of GL-GAN is beneficial to improving the convergence speed of the model and reducing the running time. The Table 1 lists the FID score and training times for the CelebA dataset in 128×128 and 64×64 resolution. The comparison models about CelebA dataset include WGAN-

Fig. 6. Randomly generated high-resolution images by GL-GAN method on Oxford Flowers(256×256) dataset.

Fig. 7. Randomly generated high-resolution images by GL-GAN method on Celeba-HQ512 dataset.

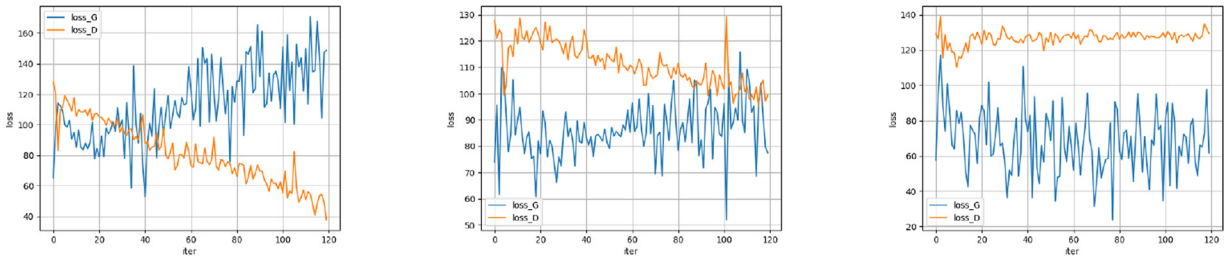


Fig. 8. The loss curves about applying spectrum normalization of the different degree on the CelebA dataset.

Table 2
Running time (where hrs denotes hours) and FID comparison between different models in Oxford Flowers datasets about 256×256 resolution. The results of ProGANs and StyleGAN come from the article [28].

Method	Real Images	GPU used	Training Time	FID(↓)
WGAN-GP	50K	1 V100-32GB	23.3hrs	60.86
ProGANs	10M	1 V100-32GB	104hrs	60.40
StyleGAN	7.2M	2 V100-32GB	33 hrs	64.70
GL-GAN	50k	1 V100-32GB	16.3hrs	51.30

Table 3
FID comparison between different models in CelebA-HQ datasets where MD and HQ denotes model and CelebA-HQ256 dataset. FID is based on 50,000 generated samples compared to training samples.

MD	SNGAN	AFHR	IP-GAN	StyleALAE	OURS
HQ	24.46	20.78	20.93	19.21	**19.09**

GP [8], SNDCGAN [9], SAGAN [10], CR-GAN[1] [25], ICR-GAN [26], IP-GAN [30],WGAN-QC [22] and GL-GAN. In a shorter time, our model can generate higher quality images, whether at 64×64 or 128×128 resolution(see Table 1) in CelebA. Compared with WGAN-QC, the training time of our method is reduced by almost half while maintaining high quality images. We improve the state-of-the-art FID from 15.43 to 12.37 in 128×128 resolution and from 12.28 to 8.3 in 64×64 resolution about CelebA dataset, which achieves state-of-the-art image synthesis result.

Table 2 shows the results of the proposed method on Oxford Flowers dataset. The baseline models include WGAN-GP [8], ProGANs [13] and StyleGAN [12]. In Table 2, the row of real images means the number of images which are used to calculate FID. GL-GAN can synthesis the highest quality images(FID:51.3) with the lowest training time. The comparison models about CelebA-HQ include SNGAN [9], AFHR [29], IP-GAN [30], StyleALAE [31] and our model in Table 3. In CelebA-HQ, our method also achieves relatively great performance (Fid:19.09). Although our results are slightly superior to StyleALAE (Fid:19.21), we can achieve almost the same effect using a simple network structure in a shorter time, comparing with StyleALAE which owns complex network architecture and takes much time to train. The aforementioned results show that our method boosts the convergence and yields high-quality images owing to the adoption of global and local optimization method.

4.4. Ablation study

In this section, we analyze the validity of some methods used in GL-GAN, including the effectiveness of local spectral normalization

[1] The codes of CR-GAN, ICR-GAN, IP-GAN, AFHR and StyleALAE aren't open source, so we just use the relevant results in the original article for reference.

(a) baseline+patch8×8

(b) baseline+patch8×8+Ada-OP

Fig. 9. Generated images in different training modes about CelebA dataset.

Table 4
FID scores about CelebA(128) datasets with different degree of spectrum norm on the base+patch8 method.

Method	no-norm	local-norm	global-norm
FID	40.01	14.09	171.02

Table 5
The FID on CelebA(128), CelebA-HQ and LSUN dataset under different cases. FID is based on 10,000 generated samples and training samples.

Method	CelebA	CelebA-256	LSUN
base	21.24	35.75	56.59
base+path4	16.67	20.63	51.29
base+path8	14.09	28.96	49.70
base+path4+Ada-OP	**12.37**	21.61	**38.66**
base+path8+Ada-OP	12.86	**20.35**	41.14

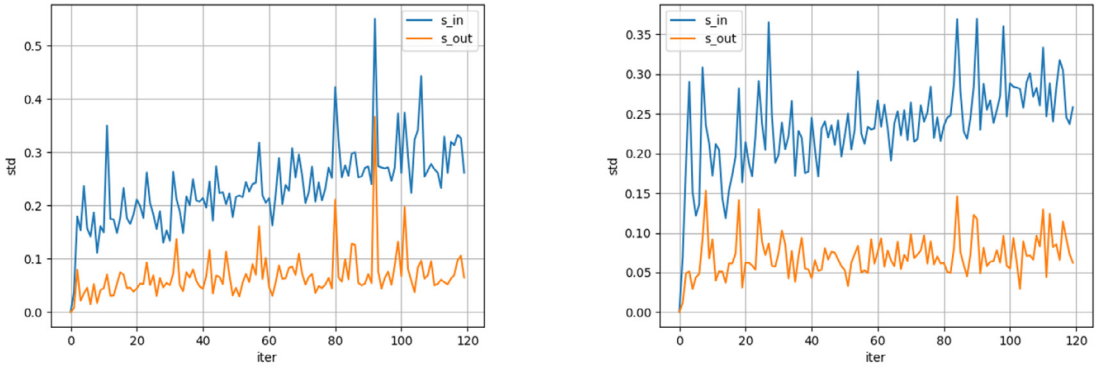
(local-norm), feature map method and Ada-OP method in improving the stability, focusing out the low-quality regions and generating high-quality images.

Spectral Normalization In terms of stability, the method of executing global and local spectrum norm to generator and discriminator respectively(called local-norm) is more effective. In Fig 8 and Table 4, we both present the results by applying different degree spectral norm with model, where no-norm, local-norm and global-norm respectively represent no, local and global implementation of spectrum normalization. As can be seen from the Fig 8(a), the loss curve about D has great change and about G is fairly unstable when without spectral normalization. The reason could be that it is not enough to stabilize the training only by regularization techniques. In contrast, the loss curve about D hardly changes in the condition of the global spectral norm(see Fig 8(c)). It indicates that the global spectral(about G and D) norm will lead to the strict parameter limitation, which has a negative effort to convergence speed.

Compared with no-norm and global-norm method, the curve of the local-norm implies that the model realizes the mutual progress between generator and discriminator with stable loss change. One possible explanation is that local-norm method has a certain limitation on parameters and is beneficial to the stable training. Besides, we also show the FID of images about implementing the three methods in Table 4, which can see that the FID score (14.09) obtained by local-norm is the lowest. These results both demonstrate that applying local-norm with model is effective to stabilize training and further improve the generation quality. The baseline(base) mentioned later in the paper refers to the method of adapting local-norm on the original model.

Feature Map Owing to the appearance that some local regions have poor generation quality within some generated images compared with other regions, the feature map, which is as the output of discriminator, is applied to the model to further select the low-quality regions. It can be seen from the Fig 1 that the output's feature map can accurately find the low-quality regions (i.e., the red region), which implies that it is reasonable to using feature map as a basis for detail modification. We experiment on CelebA, CelebA-HQ and LSUN datasets under different conditions respectively(i.e., base model, base+patch4 and base+patch8). As can be seen from the Table 5, the base+patch(14.09 for CelebA, 20.63 for CelebA-hq256 and 49.70 for LSUN-church) method has a lower FID score compared with the base method(21.24 for CelebA, 35.75 for CelebA-hq256 and 56.59 for LSUN) in all the three datasets, where base, patch4, patch8 and Ada-OP respectively denote using the local-norm method in original models, using the size of 4×4 feature map, using the size of 8×8 feature map and using the adaptive global and local bilevel optimization method. The low FID scores of the feature map method on three datasets show that it is general to improve the synthetic image performance.

Ada-OP Based on the feature map obtained by discriminator, we carry out adaptive global and local bilevel optimization method(Ada-OP) according to local and global standard deviation of the feature map. To test the performance of the method, we show the change curves of standard deviation under the situation of applying or not the method. In Fig 10, s_in and s_out denote local and global standard deviation respectively. Compared to Fig 10(a) (without the method), the change ranges of two standard deviations are both smaller and more stable with Ada-OP (Fig 10(b)). In addition, we can observe that the generated images look more realistic with fewer artifacts in Fig 9(b) than the images in Fig 9(a) (the artifact images marked with red boxes). From Table 5, it achieves the optimal FID with base+patch+Ada-OP method in all three datasets(12.37 for CelebA, 20.35 for CelebA-HQ256 and 38.66 for LSUN). Compared to the baseline, FID scores respectively decrease by 15 points and 17 points on the CelebA-HQ256 and LSUN by using base+patch+Ada-OP method in Table 5. These results both demonstrate that Ada-OP is effective to reduce the inner difference and accurately modify low-quality regions.



(a) base+patch (b) base+patch+Ada-OP

Fig. 10. Standard deviation changes curves in different training modes about CelebA. s_in,s_out denote local and global standard deviation respectively. The horizontal axis is the number of iterations and the vertical axis is the standard deviation.

5. Conclusion

Owing to the imbalance of quality distribution within a generated image where some poor-quality areas appear compared with other regions and low training efficiency, we proposes an adaptive global and local bilevel optimization model (GL-GAN). The model adaptively optimizes the image from both global and local aspects. Aiming at local optimization, a local bilevel optimization model is proposed. Based on the feature matrix from the discriminator's output in which each element measures the quality of a receptive field of the image, the local bilevel optimization model firstly finds out the region with poor generation quality, and then optimizes only this region to guide the update of generator parameters. GL-GAN is allowed to effectively avoid the nature of generated images' imbalance by local bilevel optimization model.

Furthermore, we conduct the adaptive global and local bilevel optimization method(Ada-OP) based GL-GAN theory. On account of the quality feature matrix from the discriminator's output, Ada-OP adopts the local and global quality standard deviation as the optimized mode measure. High-quality images are generated in a complementary and promoting way, where global optimization is to optimize the whole images and the local is only to optimize the low-quality areas. The computational efficiency is greatly improved through Ada-OP, which provides an idea for the low-efficiency models. To stabilize training, we investigate the spectral normalization and apply the local-norm to our model. On the dataset CelebA, Oxford Flowers, CelebA-HQ and LSUN, the convergence speed and the quality of image both have an excellent improvement.

We note that our method could easily only select the low-quality rectangular receptive field. And there may be an overlap between rectangular receptive fields. Further research is needed in this regard. Another, as an interactive optimization method, Ada-OP can achieve adaptive optimization to a certain extent by using quality standard deviation of feature matrix as an evaluation index, but it also has some limitations, which needs to be further studied. We will figure out the more appropriate local and global optimization method. In the field of application, interesting work is extending GL-GAN to some lightweight applications, such as edge computing and mobile devices. Another application of the proposed method is to transfer the frame of the adaptive local and global optimization to other models.In any case, GL-GAN provides an inspiration for the following work in improving the imbalance of generated image quality and low training efficiency.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work.

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted, the paper title "GL-GAN: Adaptive Global and Local Bilevel Optimization model of Image Generation".

Acknowledgement

This work was primarily supported by National Key R&D Program of China (NO.2018YFC1604000) and Fundamental Research Funds for the Central Universities(2662021JC008).

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[2] D.P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).
[3] L. Dinh, D. Krueger, Y. Bengio, Nice: non-linear independent components estimation, arXiv preprint arXiv:1410.8516 abs/1410.8516 (2015).
[4] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: International Conference on Learning Representations, 2016, pp. 1–15.
[5] A.B.L. Larsen, S.K. Sønderby, H. Larochelle, O. Winther, Autoencoding beyond pixels using a learned similarity metric, in: International Conference on Machine Learning, PMLR, 2016, pp. 1558–1566.
[6] G. Perarnau, J. Van De Weijer, B. Raducanu, J.M. Álvarez, Invertible conditional gans for image editing, in: Advances in Neural Information Processing Systems, 2016, pp. 1–9.
[7] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 214–223.
[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.
[9] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, in: International Conference on Learning Representations, 2018, pp. 1–26.
[10] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 7354–7363.
[11] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).
[12] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.
[13] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018.
[14] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[15] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5967–5976.
[16] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis, in: International Conference on Learning Representations, 2019.
[17] S. Nowozin, B. Cseke, R. Tomioka, f-gan: Training generative neural samplers using variational divergence minimization, in: Advances in Neural Information Processing Systems, 2016, pp. 271–279.
[18] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S.P. Smolley, On the effectiveness of least squares generative adversarial networks, IEEE Trans Pattern Anal Mach Intell 41 (12) (2019) 2947–2960, doi:10.1109/TPAMI.2018.2872043.
[19] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems, 2016, pp. 2172–2180.
[20] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Transactions on Graphics (ToG) 36 (4) (2017) 1–14.
[21] G.-J. Qi, Loss-sensitive generative adversarial networks on lipschitz densities, Int J Comput Vis 128 (5) (2020) 1118–1140.
[22] H. Liu, X. Gu, D. Samaras, Wasserstein gan with quadratic transport cost, in: IEEE International Conference on Computer Vision, 2019, pp. 4832–4841.
[23] L. Mescheder, A. Geiger, S. Nowozin, Which training methods for gans do actually converge? in: International Conference on Machine Learning, PMLR, 2018, pp. 3481–3490.
[24] A. Al-Dujaili, T. Schmiedlechner, U.-M. O'Reilly, et al., Towards distributed coevolutionary gans, AAAI Conference on Artificial Intelligence (2018) 1–6.
[25] H. Zhang, Z. Zhang, A. Odena, H. Lee, Consistency regularization for generative adversarial networks, in: International Conference on Learning Representations, 2020.
[26] Z. Zhao, S. Singh, H. Lee, Z. Zhang, A. Odena, H. Zhang, Improved consistency regularization for gans, arXiv preprint arXiv:2002.04724 (2020).
[27] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8798–8807.
[28] A. Karnewar, O. Wang, Msg-gan: Multi-scale gradients for generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 7796–7805.
[29] J. Liu, Y. Yao, J. Ren, An acceleration framework for high resolution image synthesis, arXiv preprint arXiv:1909.03611 (2019).
[30] Y.-J. Yeo, Y.-G. Shin, S. Park, S.-J. Ko, A simple yet effective way for improving the performance of gans, IEEE Trans Neural Netw Learn Syst (2021) 1–8.
[31] S. Pidhorskyi, D. Adjeroh, G. Doretto, Adversarial latent autoencoders, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 14092–14101.
[32] Q. Wang, H. Fan, G. Sun, Y. Cong, Y. Tang, Laplacian pyramid adversarial network for face completion, Pattern Recognit 88 (2019) 493–505.
[33] U. Mutlu, E. Alpaydin, Training bidirectional generative adversarial networks with hints, Pattern Recognit 103 (2020) 107320.
[34] N. Zaltron, L. Zurlo, S. Risi, Cg-gan: An interactive evolutionary gan-based approach for facial composite generation, in: AAAI Conference on Artificial Intelligence, 2020, pp. 2544–2551.

[35] T. Guo, C. Xu, B. Shi, C. Xu, D. Tao, Smooth deep image generator from noises, in: AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 3731–3738.
[36] W. Li, L. Fan, Z. Wang, C. Ma, X. Cui, Tackling mode collapse in multi-generator gans with orthogonal vectors, Pattern Recognit 110 (2021) 107646.
[37] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[38] C. Li, M. Wand, Combining markov random fields and convolutional neural networks for image synthesis, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2479–2486.
[39] W. Tu, W. Hu, X. Liu, J. He, Drpan: A novel adversarial network approach for retinal vessel segmentation, in: IEEE Conference on Industrial Electronics and Applications, IEEE, 2019, pp. 228–232.
[40] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: International Conference on Computer Vision, 2015, pp. 3730–3738.
[41] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Indian Conference on Computer Vision, Graphics and Image Processing, 2008.
[42] F. Yu, Y. Zhang, S. Song, A. Seff, J. Xiao, Lsun: construction of a large-scale image dataset using deep learning with humans in the loop, arXiv preprint arXiv:1506.03365 (2015).
[43] D. Dowson, B. Landau, The fréchet distance between multivariate normal distributions, J Multivar Anal 12 (3) (1982) 450–455.

Jinhai Xiang received the M.E. degree in Computer Science from China University of Geoscience (Wuhan) in 2003, and the Ph.D. degree in Computer Architecture from Huazhong University of cience and Technology (HUST) in 2014. He is now an Associate Professor in the College of Informatics, Huazhong Agricultural University, Wuhan, China. His main interests include computer vision and machine learning.