Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.052097

ARTICLE

An Enhanced GAN for Image Generation

Chunwei Tian[1,2,3,4], **Haoyang Gao**[2,3], **Pengwei Wang**[2] **and Bob Zhang**[1,*]

[1]PAMI Research Group, University of Macau, Macau, 999078, China

[2]School of Software, Northwestern Polytechnical University, Xi'an, 710129, China

[3]Yangtze River Delta Research Institute, Northwestern Polytechnical University, Taicang, 215400, China

[4]Research & Development Institute, Northwestern Polytechnical University, Shenzhen, 518057, China

*Corresponding Author: Bob Zhang. Email: bobzhang@um.edu.mo

Received: 22 March 2024 Accepted: 20 June 2024 Published: 18 July 2024

ABSTRACT

Generative adversarial networks (GANs) with gaming abilities have been widely applied in image generation. However, gamistic generators and discriminators may reduce the robustness of the obtained GANs in image generation under varying scenes. Enhancing the relation of hierarchical information in a generation network and enlarging differences of different network architectures can facilitate more structural information to improve the generation effect for image generation. In this paper, we propose an enhanced GAN via improving a generator for image generation (EIGGAN). EIGGAN applies a spatial attention to a generator to extract salient information to enhance the truthfulness of the generated images. Taking into relation the context account, parallel residual operations are fused into a generation network to extract more structural information from the different layers. Finally, a mixed loss function in a GAN is exploited to make a tradeoff between speed and accuracy to generate more realistic images. Experimental results show that the proposed method is superior to popular methods, i.e., Wasserstein GAN with gradient penalty (WGAN-GP) in terms of many indexes, i.e., Frechet Inception Distance, Learned Perceptual Image Patch Similarity, Multi-Scale Structural Similarity Index Measure, Kernel Inception Distance, Number of Statistically-Different Bins, Inception Score and some visual images for image generation.

KEYWORDS

Generative adversarial networks; spatial attention; mixed loss; image generation

1 Introduction

Due to the development of image vision techniques, image generation techniques have been applied in many fields, i.e., person privacy protection [1] and entertainment [2]. That is, digital devices can use generated face rather than captured unauthorized faces to address personal privacy protection questions [1]. Image generation techniques [2] can generate virtual persons rather than spokesmen in entertainment to save costs. Traditional image generation techniques can use multiple two-dimensional images to recover three-dimensional structures to achieve a transform of multiple images to one aim image [3]. For instance, three-dimensional morphable model uses principal component analysis to decrease texture and facial shape features in low-dimensional space to generate more real face images

[3]. However, it requires a lot of varying illuminations, postures, and expressions, which may cause high data collection costs and low of the proposed method. GANs with generating high-quality and diverse images have become popular in image generation [4]. Designed two different encoders in an unsupervised way can learn potential distribution to address attribute entanglement, where output distribution can use adversarial strategy learning to maintain characteristic features of GAN to improve the effect of image generation [5]. Using ResNet feature pyramid as an encoder network can extract style from three features with different scales, and then a mapping network is used to extract learned styles from corresponding images in image generation [6]. Alternatively, an iterative feedback mechanism is used to improve the generation quality of face images to keep a balance between image fidelity and editing ability. To improve the quality of image generation, a residual learning is used in a transfer process to improve the iterative feedback mechanism [7]. Using compute similarity between potential vectors and images to design an adaptive similarity encoder to generate high-fidelity images, which can use existing encoders into different GANs to generate images [8]. To improve performance of image generation without increasing computational costs, multi-layer losses of ID and facial analysis are referred to generate more detailed information for image generation [9]. Alternatively, Xu et al. [10] used a novel hierarchical encoder to extract hierarchical features via input images to improve the effects of generated images. Although mentioned gametic GANs may improve the effects of generated images, they may suffer from challenges from varying scenes.

In this paper, we present an enhanced GAN via improving a generator for image generation termed as EIGGAN. EIGGAN utilizes a spatial attention mechanism to improve the generator in order to extract salient information that can enhance the correctness of the predicted images. To improve the generation effects, parallel residual operations are gathered into a generation network to extract more structural information from the different layers in terms of their relation to context. To make a tradeoff between speed and accuracy in image generation, a mixed loss function is used in a GAN to generate more realistic images. Experiments illustrate that the proposed EIGGAN is competitive in terms of the metrics: Frechet Inception Distance (FID) [11], Learned Perceptual Image Patch Similarity (LPIPS) [12], Multi-Scale Structural Similarity Index Measure (MS-SSIM) [13], Kernel Inception Distance (KID) [14], Number of statistically-Different Bins (NDB) [15] and Inception Score (IS) [16] for image generation.

The contributions of this paper can be summarized as follows:

1. A spatial attention mechanism is employed to improve the generation to facilitate more salient information, enhancing the truthfulness of the generated images.
2. Parallel residual operations are used to extract more complementary and structural information with a spatial attention mechanism to improve the effects of image generation, according to its relation to context.
3. A mixed loss is applied to generate more realistic images.

The remaining parts of this paper are conducted as follows. Section 2 presents related work. Section 3 gives the proposed method. Section 4 provides experimental analysis and results. Section 5 summarizes the whole paper.

2 Related Work

GANs with strong generative abilities are used for image generation in generation [17]. To improve generation effects, designing novel network architectures can improve GANs as the popular way of image generation [17]. It can be summarized into two kinds: improved common GANs and

StyleGANs. The first method usually improves a generator or discriminator to improve the learning abilities of GANs for image generation. For improving a generator, asymmetric network architectures between a generator and discriminator are designed to address image content differences between input and outputs to promote the effects and performance of image-to-image translation [17]. To address native relation effect of semantic and latent space, a combination of a pretrain GAN and prior knowledge is used to extract latent semantic from architecture attributes in shallow layers and in apparent deep layers to improve the quality of generation images [18]. Alternatively, zooming areas of training data in image representation space in a discriminator can easily train a GAN to generate images [19]. To address dense visual alignment questions, a spatial Transformer is used to map random samples to a joint object mode for image generation [20]. To edit more attributions of generated images, the second method based on StyleGANs is proposed.

That is, a StyleGAN compares a learnable intermediate latent space W and standard Gaussian latent space to the reflected distribution of training data and they can effectively code rich semantic information to change the attribution of generated images, i.e., expressions and illuminations in image generation [21]. To better address image generation tasks, the second method uses fine-tuning styleGANs to enhance the quality of generation images [22]. To address inverse mapping questions of high-quality reconstruction, editability and fast reference, hypernetworks are proposed [16]. A two-phase mechanism was conducted as follows. The first phase was used to train an encoder to map an input image to a latent space. The second phase utilized a hypernetwork to recover lost information from the first phase to image editing [22]. A hypernetwork can be used to tune the weights of StyleGAN to better express given images in editable areas from the latent space for image editing [23]. To address domain transfer of image generation, a simple feature match loss is gathered into a StypleGAN to improve generation quality with less computational costs [24]. To avoid the collapse of latent variants in StyleGAN, a class embedding enhancement mechanism is referred to a self-supervised learning based on a latent space to reduce the relation of latent variants to improve the results of image generation [25]. Although these methods can improve the effects of image generation, StyleGANs refer to more training time and higher computational costs. To better generate images, a Progressive Growing of Generative Adversarial Networks (PGGAN) uses a progressively large model to reduce training difficulty to generate smoother and continuous images [26]. Inspired by that, we use a PGGAN to improve a GAN for image generation in this paper.

3 Proposal Method

3.1 Network Architecture

To better generate high-quality images, we designed an enhanced GAN to improve a generator in image generation (EIGGAN) in Fig. 1. EIGGAN mainly uses a generator and a discriminator to game to generate high-quality images. To enhance the robustness of gamistic generator and discriminator in varying scenes, an improved generator is developed. To generate more real images, three operations are affected on a generation network. The first operation is that a parallel residual learning operation is used to extract more structural information of different layers in terms of relation of context. To enhance the truthfulness of generating images, a spatial attention mechanism is gathered into this generation network to extract more salient information to improve the effect of image generation. The third operation utilizes a mixed loss function to update the parameters of a GAN to balance speed and accuracy in image generation. More detailed information on the generation network can be shown as follows.

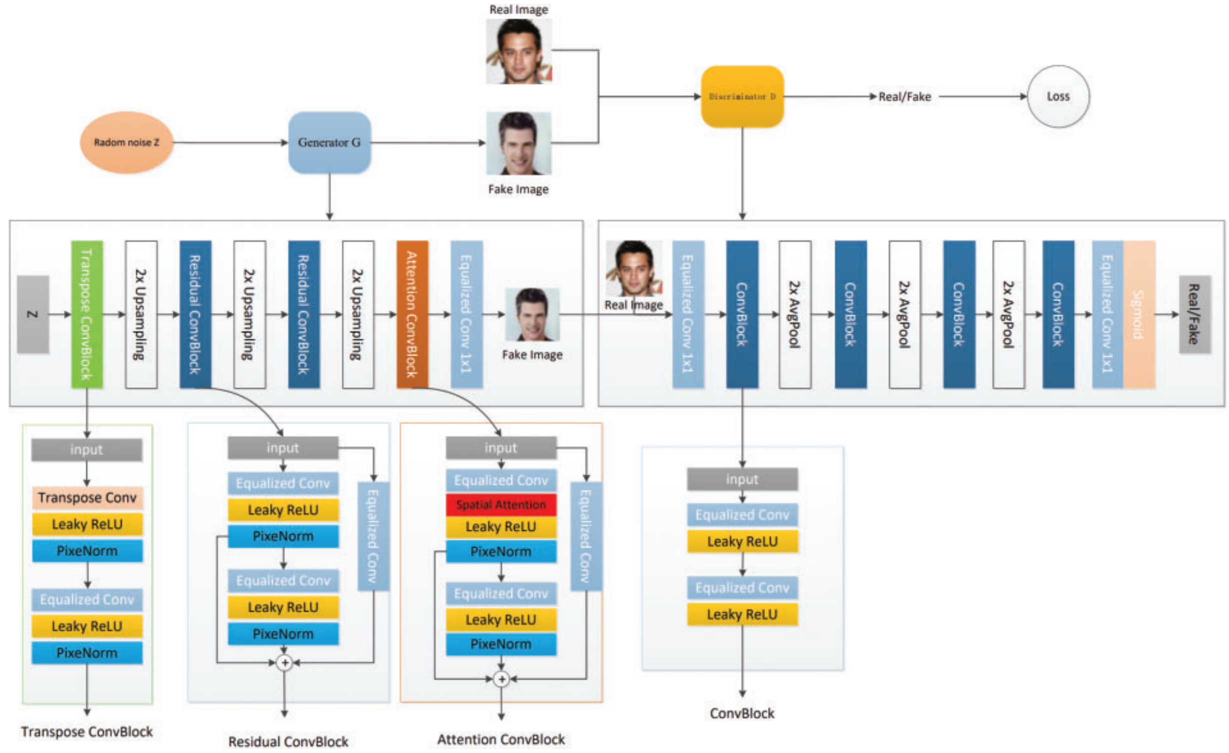


Figure 1: Network architecture of EIGGAN

The generation network is composed of five components, i.e., Transpose ConvBlock, Residual ConvBlock, 2X Upsampling, Attention ConvBlock, Equalized Conv operation. The Transpose ConvBlock [26] is set to the first layer, which can convert a noisy vector to a matrix. 2X Upsampling is set to the second, fourth and sixth layers to enlarge obtained feature mapping to capture more context information. Enhanced Residual ConvBlock is set to the third and fifth layers to extract more structural information from different layers in terms of relation to context. An enhanced attention ConvBlock is set to the sixth layer to extract more salient information for image generation. An Equalized Conv [26] is used as the last layer to normalize obtained features to accelerate training speed. To clearly express the mentioned process, the following equation can be conducted.

$$O_g = G(z) = EC(EAC(Up(ERC(Up(ERC(Up(TC(z)))))))) \tag{1}$$

where z is random noise, G is a function of a generation network, O_g is an output of the generation network. TC denotes a function of a Transpose ConvBlock, ERC denotes a function of an enhanced residual ConvBlock. Up denotes a 2X upsampling operation. EC denotes a function of an Equalized Conv. Its parameters can be updated by a mixed loss function in Section 3.2.

3.2 Loss Function

To make a tradeoff between performance and speed, a mixed loss is conducted. That is, a mixed loss is composed of normal loss-based GAN [27], non-saturating loss [28] and a combination of R1 [29] and R2 regularization [29]. Non-saturation loss can make a generator more stable in the training process [28]. R1 regularization loss can make a discriminator easier to converge via penalizing gradients of real images to obtain more reliable images [28]. R2 regularization loss can penalty gradients of generating images to easier converge for image generation [29]. Mentioned a loss function is shown as follows.

$$L_D = -\frac{1}{m}\sum_{i=1}^{m}\left[\log D\left(x^{(i)}\right)\right] + \left[\log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right)\right] + \frac{\gamma}{2}E_{p\vartheta+p\theta}\left[\nabla \left\| D_{\psi}\left(x\right)\right\|^2\right] D\left(G\left(z\right)\right) \tag{2}$$

$$L_G = \frac{1}{m}\sum_{i=1}^{m}\left[\log\left(G\left(z^{(i)}\right)\right)\right] \tag{3}$$

where L_D is a loss of a discriminator. $D(x)$ denotes probability of real images, where x is a given real image. $D(G(z))$ denotes probability of generating images, where z is noise. m is a sample minibatch of noise. $p\vartheta$ is distribution of true images. $p\theta$ is distribution of a generator. L_G is a loss of a generator.

3.3 Enhanced Residual ConvBlock

Enhanced Residual ConvBlock is used to extract more structural information of different layers for image generation in terms of relation of context as shown in Fig. 1. It is composed of three Equalized Conv, two Leaky ReLU and two PixelNorm. Two combinations of Equalized Conv, Leaky ReLU and PixeNorm are stacked to extract more structural information to improve the effect of image generation. To extract more complementary information, an extra Equalized Conv, an output of the first stacked Equalized Conv, Leaky ReLU and PixelNorm and an output of two stacked Equalized Conv, Leaky ReLU and PixelNorm in a parallel way are gathered to extract more structural information, where a residual learning operation denotes a fusion way. It can be expressed as the following equation:

$$O_{ERC} = ERC\left(I_t\right) = PN\left(LR\left(EC\left(I_t\right)\right)\right) + PN\left(LR\left(EC\left(PN\left(LR\left(EC\left(I_t\right)\right)\right)\right)\right)\right) + EC\left(I_t\right) \tag{4}$$

where I_t is an input of an enhanced residual ConvBlock. O_{ERC} is output of an enhanced residual ConvBlock. LR is a function of Leaky ReLU. PN is a function of PixelNorm. + is a residual learning operation.

3.4 Enhanced Attention ConvBlock

An enhanced attention ConvBlock can extract more salient information for image generation. It is composed of four components, i.e., Equalized Conv, Spatial Attention, Leaky ReLU and PixelNorm. It is different from ERresidual Block that spatial attention [30] is set between the first Equalized Conv and Leaky ReLU to extract salient information to improve effect of image generation. This process can be shown as follows:

$$\begin{aligned} O_{EAC} = EAC\left(I_{EA}\right) &= PN\left(LR\left(SA\left(EC\left(I_{EA}\right)\right)\right)\right) \\ &+ \mathrm{PN}\left(\mathrm{LR}\left(\mathrm{EC}\left(\mathrm{PN}\left(\mathrm{LR}\left(\mathrm{SA}\left(\mathrm{EC}\left(\mathrm{I}_{EA}\right)\right)\right)\right)\right)\right)\right) \\ &+ \mathrm{EC}\left(\mathrm{I}_{EA}\right) \end{aligned} \tag{5}$$

where I_{EA} is an input of EAttention ConvBlock. O_{EAC} is an output of EResidual ConvBlock. SA denotes a function of spatial attention.

4 Experimental Analysis and Results

4.1 Datasets

Real image datasets are composed of 202,599 face images from CelebA [31] and 60,000 images with ten different categories, i.e., planes, cars, birds, cats, deer, dogs, frogs, horses, ships and trucks from CIFAR-10 [32]. Each image is 32 × 32. Generating image datasets: Each EIGGAN model can generate 10,000 images.

4.2 Experimental Settings

Parameters of training a EIGGAN in image generation are listed as follows. Slope parameter of LeakyReLU is 0.2. λ of regularization parameter is 1 on CelebA and it is 0.1 on CIFAR-10. Batch size is 64. Epoch number on CelebA is 255. Epoch number on CIFAR-10 is 266. Learning rate on CelebA is 0.002, learning rate on CIFAR-10 is 0.001. $\beta 1 = 0$ and $\beta 2 = 0.99$. Also, parameters can be optimized by Adam [33]. All the codes implemented by PyTorch 1.13.1, and Python of 3.9.13 run on Ubuntu 20.04.3 with AMD EPYC 7502/3.35 GHz, Central Processing Unit (CPU) of 32 cores, 128 RAM. Also, a Graphics Processing Unit (GPU) of Nvidia GeForce GTX 3090 and Nvidia CUDA 12.2 can be used to improve training speed.

4.3 Experimental Analysis

To improve robustness of GANs in image generation, an enhanced GAN for image generation is proposed. To improve the quality of generating images, an enhanced generation network is conducted. Deep networks rely on a deep network architecture to extract more structural information [34]. To quickly extract key information, a spatial attention mechanism is proposed [30]. It can use different dimensional channels to extract salient information [30]. Inspired by that, a spatial attention mechanism is fused into the third EResidual ConvBlock from a generation network to extract salient information to improve effects in image generation, where its information can be shown in Section 3.4 Its effectiveness is given in Table 1. That is, a combination of PGGAN and Spatial attention mechanism has obtained a lower FID than that of PGGAN in Table 1. Although deep networks can extract more accurate information to pursue better performance in vision tasks, they may ignore the importance of hierarchical information to limit their better performance and robustness [34]. To overcome the drawbacks, a parallel residual learning operations are set in the third EResidual ConvBlock as shown in Fig. 1 to extract more structural information from different layers. Its competitive results can be proved by comparing 'A combination of PGGAN, Spatial attention and Parallel residual learning operations' and 'A combination of PGGAN and Spatial attention in Table 1. A mixed loss function to update parameters of a GAN to balance speed and accuracy in image generation. Non-saturation loss can make a generator more stable in training process. R1 regularization loss can make a discriminator easier to converge via penalizing gradients of real images to obtain more reliable images. R2 regularization loss can penalty gradients of generating images to easier converge for image generation. Effects of R1 for EIGGAN for image generation can be tested by using 'A combination of PGGAN, Spatial attention, Parallel residual learning operations, Non-saturation and R1' and 'A combination of PGGAN, Spatial attention and Parallel residual learning operations' in terms of FID in Table 1. It has an improvement of 2.76 in terms of FID in Table 1, which shows effectiveness of R1 for image generation. Positive effects of a mixed loss can verified by comparing 'A combination of PGGAN, Spatial attention, Parallel residual learning operations, Non-saturation and R1' and 'A combination of PGGAN, Spatial attention, Parallel residual learning operations and Mixed loss' in Table 1. Effectiveness of proposed key techniques can be verified by comparing 'A combination of PGGAN, Spatial attention, Parallel residual learning operations and Mixed loss' and 'PGGAN' in Table 1.

Table 1: FID values of different methods in image generation

Methods	FID
PGGAN [26]	28.13
A combination of PGGAN and spatial attention	26.01
A combination of PGGAN, spatial attention and parallel residual learning operations	24.33
A combination of PGGAN, spatial attention, parallel residual learning operations, non-saturation and R1	21.57
A combination of PGGAN, spatial attention, parallel residual learning operations and Mixed loss	21.36

4.4 Experimental Results

To test the effectiveness of our method, quantitative and qualitative analysis are used to conduct. Quantitative analysis that uses a Deep Convolution GAN (DCGAN) [35], WGAN-GP [36], PGGAN [26], NCSN [37] as comparative methods on CIFAR-10 and CelebA to test performance of the proposed EIGGAN in image generation. We have chosen six indicators, i.e., Frechet Inception Distance (FID) [11], Learned Perceptual Image Patch Similarity (LPIPS) [12], Multi-Scale Structural Similarity Index Measure (MS-SSIM) [13], Kernel Inception Distance (KID) [14], Number of statistically-Different Bins (NDB) [15] and Inception Score (IS) [16] to demonstrate the superiority of our method. All formulas of FID [11], LPIPS [12], MS-SSIM [13], KID [14] and NDB [15] can be given at https://github.com/hellloxiaotian/EIGGAN/blob/main/equation (accessed on 19/03/2024).

As shown in Table 2, we can see that the proposed EIGGAN has obtained the best results in terms of FID, LPIPS, MS-SSIM and KID on CIFAR-10. Also, it has obtained the second result in terms of NDB on CIFAR-10 in Table 2. To verify its robustness, we conduct some extended experiments on CelebA dataset. That is, our EIGGAN has obtained the best performance in terms of FID and LPIPS on CelebA in Table 3. Also, it has obtained the second results in terms of MS-SSIM, KID and NDB in Table 3. According to mentioned illustrations, we can see that the proposed EIGGAN is effective for image generation.

Table 2: Results of some methods on CIFAR-10 for image generation

Methods	FID↓	LPIPS↑	MS-SSIM↓	KID↓	NBD↓	IS↑
DCGAN [35]	35.05	0.207	0.12147	0.0238	33	6.37
WGAN-GP [36]	29.30	0.217	0.10598	0.0145	**18**	7.86
PGGAN [26]	28.13	0.214	0.11058	0.0133	68	**8.80**
NCSN [37]	25.32	0.195	0.11225	0.0111	81	8.37
EEIGGAN (ours)	**21.03**	**0.229**	**0.10501**	**0.0098**	33	8.56

To verify the superiority of our EIGGAN for image generation, we choose newest image generation methods, i.e., StyleGAN2 [38], TransGAN [39], ViTGAN [40], D2WMGAN [41], PFGAN

[42] on CIFAR-10 in terms of Information System (IS) to conduct experiments. As mentioned earlier, the FID value is used to evaluate the quality of generated images. In addition, since the CIFAR-10 dataset contains 10 different types of object images, we also tested the IS value to evaluate the type diversity of generated images. To fairly compare the performance of our proposed method for image generation, we conduct some experiments on CIFAR-10. Due to some methods, i.e., StyleGAN2, TransGAN, ViTGAN, D2WMGAN and PFGAN didn't release codes, public index, i.e., FID and IS from related GANs [38–42] can be obtained. As shown in Table 4, we can see that StyleGAN2 is superior to our EIGGAN. Also, our method has fewer parameters. Also, our method has obtained better effects than that of other methods for image generation in Table 5. In summary, our EIGGAN is comparative for image generation.

Table 3: Results of some methods on CelebA for image generation

Methods	FID↓	LPIPS↑	MS-SSIM↓	KID↓	NBD↓
DCGAN [35]	32.99	0.230	0.35644	0.0117	83
WGAN-GP [36]	31.17	0.292	**0.26625**	0.0073	**28**
PGGAN [26]	12.88	0.283	0.28567	**0.0010**	46
NCSN [37]	42.59	0.118	0.31952	0.0191	86
EEIGGAN (Ours)	**10.62**	**0.309**	0.28250	0.0019	45

Table 4: Comparing FID and IS results of some image generation methods on CIFAR-10

Methods	FID↓	IS↑
StyleGAN2 [38]	**8.41**	**9.16**
TransGAN [39]	22.53	8.26
ViTGAN [40]	30.72	8.30
D2WMGAN [41]	34.94	7.31
PFGAN [42]	47.32	7.97
EEIGGAN (ours)	21.03	8.56

Table 5: Comparison of parameters for some methods

Methods	Parameters
DCGAN [35]	12.14 M
WGAN-GP [36]	12.14 M
PGGAN [26]	17.77 M
NCSN [37]	30.18 M
StyleGAN2 [38]	28.27 M
ViTGAN [40]	45.37 M
PFGAN [42]	22.19 M
EIGGAN (Ours)	17.90 M

Qualitative analysis is composed of two parts: visual generation images and contrastive visual generation images. The first part contains two parts, i.e., ten objects (regarded as planes, cars, birds, cats, deer, dogs, frogs, horses, ships and trucks) and different face images. Also, each object is generated from the CIFAR-10 and each object has eight images with different directions in Fig. 2. Different face images from the CelebA are generated in Fig. 3. These show that our EIGGAN is effective in visual images in image generation.

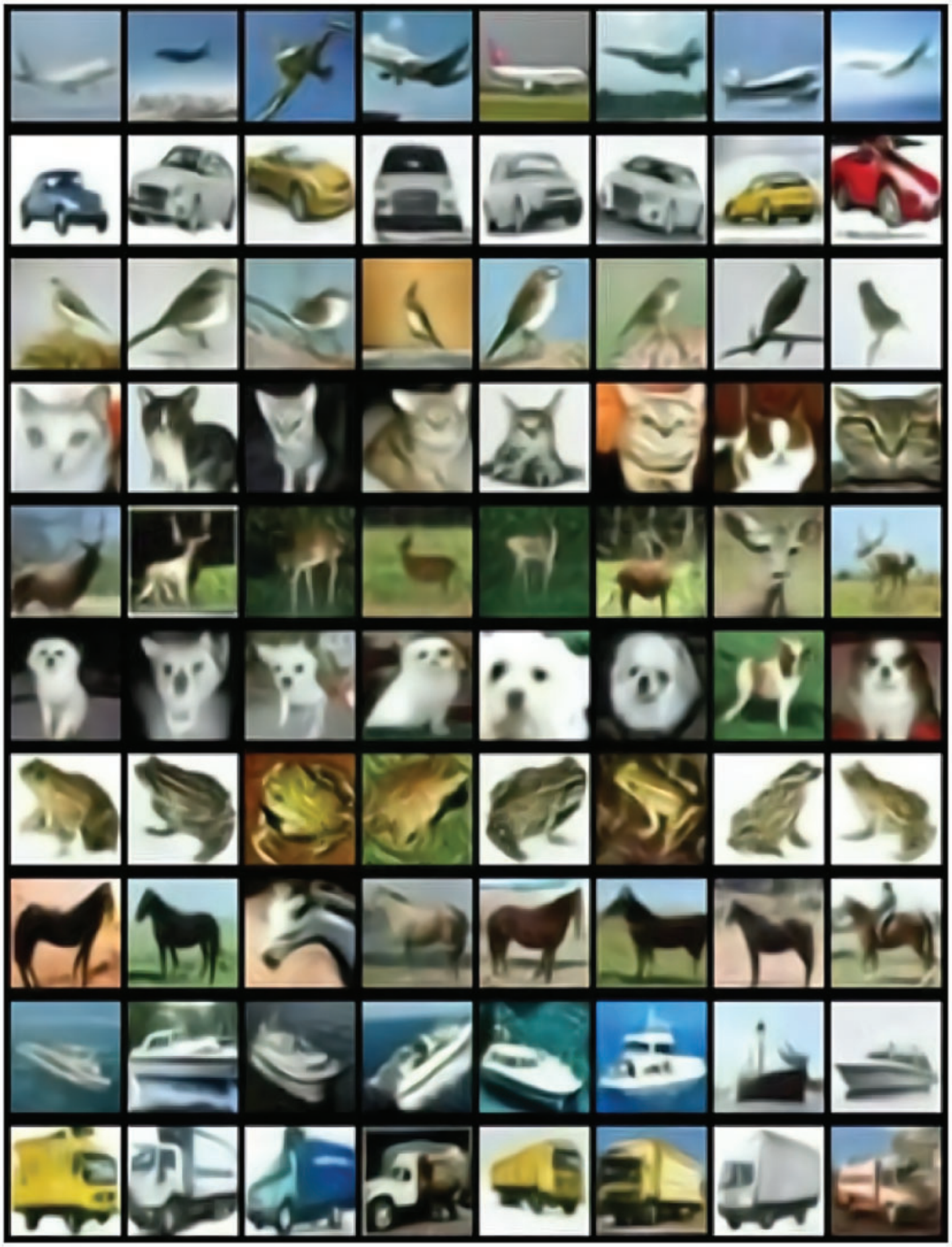


Figure 2: Generated images on the CIFAR-10 from our EIGGAN

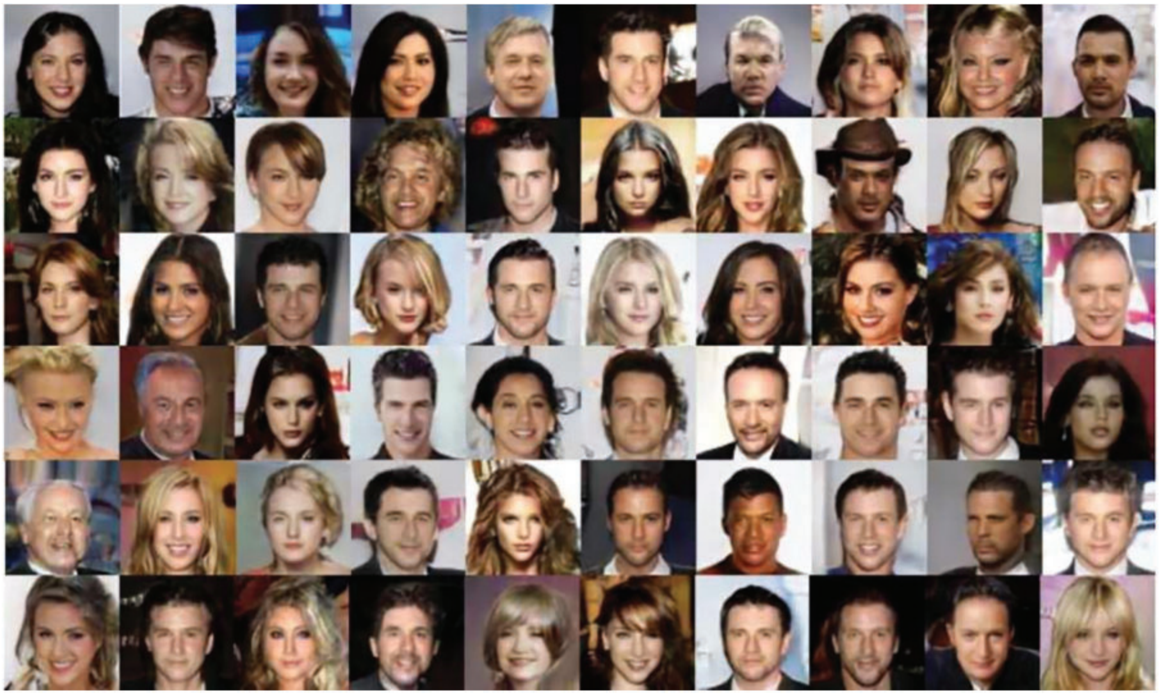


Figure 3: Generated face images from our EIGGAN

The second part conducted some comparative visual images of popular methods and our method for image generation to test the generation ability of our EIGGAN. As shown in Fig. 4, we can see that our EIGGAN is clearer than other methods. As listed in Fig. 5, we can see that our EIGGAN has more detailed information than that of other methods, i.e., DCGAN, WGAN-GP, PGGAN and NCSN. For instance, DCGAN may cause distorted faces. WGGAN-GP may generate blurred faces. NCSN may generate images of poor colour and luster. According to mentioned illustrations, our method is useful for image generation in terms of qualitative and quantitative analysis.

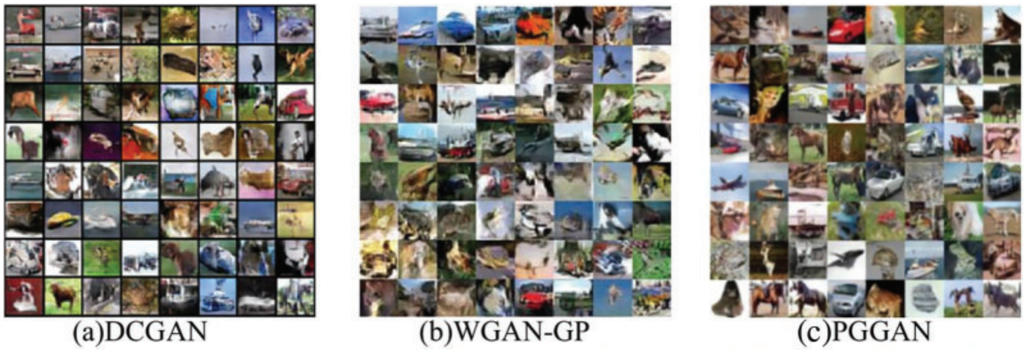

(a)DCGAN
(b)WGAN-GP
(c)PGGAN

Figure 4: (Continued)

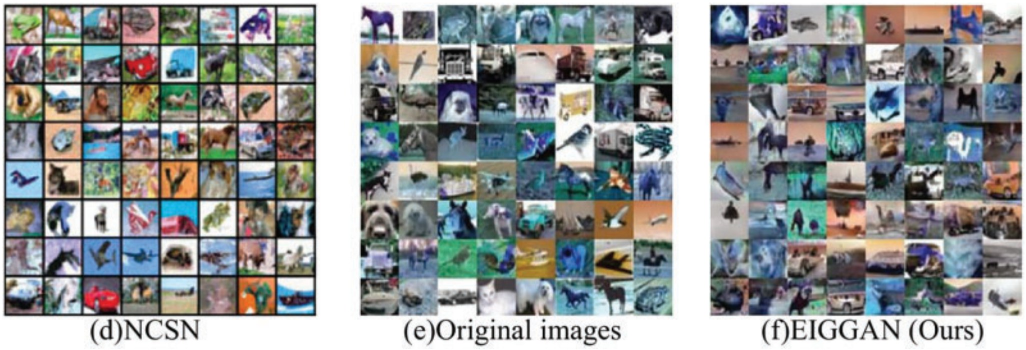
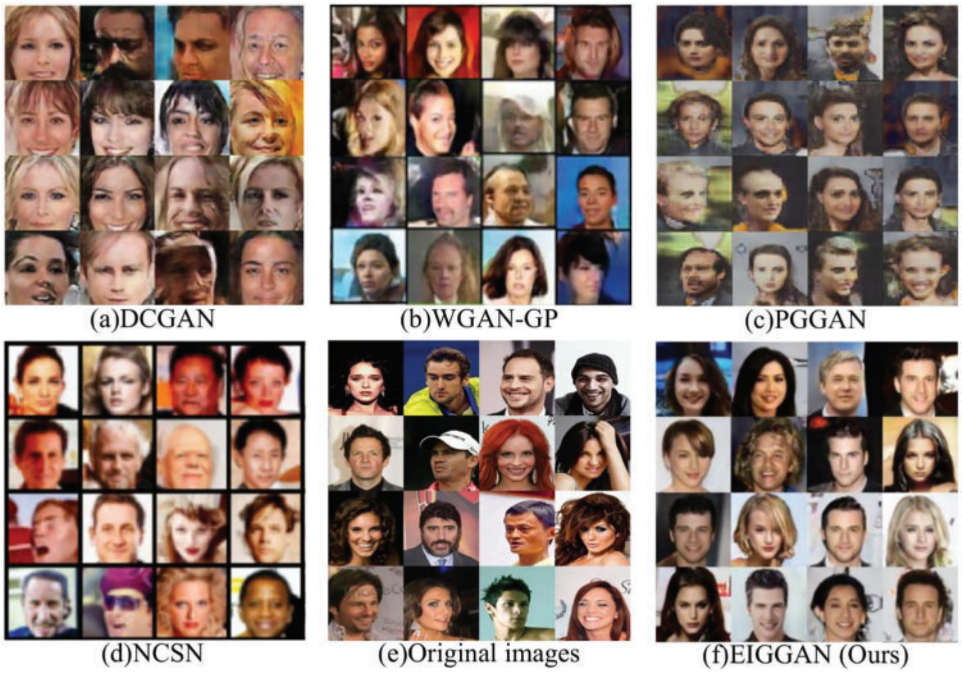

(d)NCSN (e)Original images (f)EIGGAN (Ours)

Figure 4: Generated images of different methods on CIFAR-10


(a)DCGAN (b)WGAN-GP (c)PGGAN

(d)NCSN (e)Original images (f)EIGGAN (Ours)

Figure 5: Generated images of different methods on CelebA

5 Conclusion

In this paper, we improve a generator to enhance GAN for image generation. To enrich the effect of a generator, spatial attention is applied to a generation network to extract salient information, generating more information that is detailed. To enhance the relation of context, parallel residual operations are used to increase more structural information from different layers in image generation. Taking into consideration the speed and accuracy, a mixed loss function is merged into the generation network to produce images that are more realistic. Overall, the proposed method (EIGGAN) has obtained good performance in image generation. As part of our future work, we will improve the performance of image generation according to the image attributes.

Acknowledgement: This work was supported in part by the Science and Technology Development Fund, in part by Fundamental Research Funds for the Central Universities.

Funding Statement: This work was supported in part by the Science and Technology Development Fund, Macao S.A.R (FDCT) 0028/2023/RIA1, in part by Leading Talents in Gusu Innovation and Entrepreneurship Grant ZXL2023170, in part by the TCL Science and Technology Innovation Fund under Grant D5140240118, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110079.

Author Contributions: The first author Chunwei Tian gives the main conception and writes of this paper. The second author Haoyang Gao conducts experiments, visual figures and writes part of this paper. The third author Pengwei Wang conducts part experiments. The fourth author Bob Zhang gives key comments. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Sun, Q. Wang, and Z. Liu, “Human action image generation with differential privacy,” in *2020 IEEE Int. Conf. Multimedia and Expo (ICME)*, London, UK, Jul. 6–10, 2020, pp. 1–6.

[2] W. Gao, S. Shan, X. Chai, and X. Fu, “Virtual face image generation for illumination and pose insensitive face recognition,” in *2003 Int. Conf. Multimedia and Expo. ICME’03*, Hong Kong, China, IEEE, Apr. 6–10, 2003, vol. 3.

[3] V. Blanz and T. Vetter, “Face recognition based on fitting a 3D morphable model,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003. doi: 10.1109/TPAMI.2003.1227983.

[4] I. Goodfellow *et al.*, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020. doi: 10.1145/3422622.

[5] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto, “Adversarial latent autoencoders,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 13–19, 2020, pp. 14104–14113.

[6] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, “Designing an encoder for stylegan image manipulation,” in *ACM Trans. Graph.*, New York, NY, USA, 2021, vol. 40 no. 4, pp. 1–14.

[7] Y. Alaluf, O. Patashnik, and D. Cohen-Or, “ReStyle: A residual-based stylegan encoder via iterative refinement,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 10–17, 2021, pp. 6711–6720.

[8] C. Yu and W. Wang, “Diverse similarity encoder for deep GAN inversion,” arXiv preprint arXiv: 2108.10201, 2021.

[9] T. Wei *et al.*, “E2Style: Improve the efficiency and effectiveness of StyleGAN inversion,” arXiv preprint arXiv:2104.07661, 2022.

[10] Y. Xu, Y. Shen, J. Zhu, C. Yang, and B. Zhou, “Generative hierarchical features from synthesizing images,” arXiv preprint arXiv:2007.10379, 2021.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *NIPS’17: Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, California, USA, Dec. 4–9, 2017, vol. 30, pp. 6629–6640.

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 18–23, 2018, pp. 586–595.

[13] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thrity-Seventh Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, IEEE, Nov. 9–12, 2003, vol. 2, pp. 1398–1402.

[14] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," arXiv preprint arXiv, abs/1801.01401, 2018.

[15] E. Richardson and Y. Weiss, "On GANs and GMMs," arXiv preprint arXiv:1805.12462, 2018.

[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, "Improved techniques for training GANs," in *NIPS'16: Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 5–10, 2016, pp. 2234–2242.

[17] H. Tang, D. Xu, H. Liu, and N. Sebe, "Asymmetric generative adversarial networks for image-to-image translation," arXiv preprint arXiv:1912.06931, 2019.

[18] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan and S. He, "High-resolution face swapping via latent semantics disentanglement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, Jun. 18–24, 2022, pp. 7642–7651.

[19] H. Liu *et al.*, "Improving GAN training via feature space shrinkage," in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 17–24, 2023, pp. 16219–16229.

[20] W. Peebles, J. Y. Zhu, R. Zhang, A. Torralba, A. A. Efros and E. Shechtman, "Gan-supervised dense visual alignment," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 18–24, 2022, pp. 13470–13481.

[21] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 15–20, 2019, pp. 4401–4410.

[22] T. M. Dinh, A. T. Tran, R. Nguyen, and B. S. Hua, "Hyperinverter: Improving stylegan inversion via hypernetwork," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 18–24, 2022, pp. 11389–11398.

[23] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. Bermano, "Hyperstyle: Stylegan inversion with hypernetworks for real image editing," in *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 18–24, 2022, pp. 18490–18500.

[24] D. Lee, J. Y. Lee, D. Kim, J. Choi, and J. Kim, "Fix the noise: Disentangling source feature for transfer learning of StyleGAN," arXiv preprint arXiv:2204.14079, 2022.

[25] H. Rangwani, L. Bansal, K. Sharma, T. Karmali, V. Jampani and R. V. Babu, "Noisytwins: Class-consistent and diverse image generation through stylegans," in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 17–24, 2023, pp. 5987–5996.

[26] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *The Six Int. Conf. Learn. Represent.*, Vancouver, BC, Canada, Vancouver Convention Center, Apr. 30–May 3, 2018.

[27] I. Goodfellow *et al.*, "Generative adversarial nets," in *NIPS'14: Proc. 27th Int. Conf. on Neural Inf. Process. Syst.*, Montreal, Canada, Dec. 8–13, 2014, vol. 2, pp. 2672–2680.

[28] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2016.

[29] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?," in *Proc. 35th Int. Conf. Mach. Learn. Stockholmsmässan*, Stockholm Sweden, Jul. 10–15, 2018, vol. 80, pp. 3481–3490.

[30] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *ECCV 2018:15th Eur. Conf.*, Munich, Germany, Sep. 8–14, 2018, pp. 3–19.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 7–13, 2015, pp. 3730–3738.

[32] A. Krizhevsk and G. Hinton, "Learning multiple layers of features from tiny images," *Handbook Syst. Autoimmu. Dis.*, vol. 1, no. 4, 2009.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent. 2014*, Banff, Canada, Apr. 14–16, 2014.

[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Int. Conf. Learn. Represent. 2015*, The Hilton San Diego Resort & Spa, May 7–9, 2015.

[36] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *NIPS'17:Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, California, USA, Dec. 4–9, 2017, vol. 30, pp. 5769–5779.

[37] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *NIPS'19: Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 8–14, 2019, vol. 32, pp. 11918–11930.

[38] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 13–19, 2020, pp. 8110–8119.

[39] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong gan, and that can scale up," in *NIPS'21:Proc. 35th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Dec. 6–14, 2021, vol. 34, pp. 14745–14758.

[40] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu and C. Liu, "ViTGAN: Training gans with vision transformers," in *2021 The Ninth Int. Conf. Learn. Represent.*, Virtual Only Conference, May 3–7, 2021.

[41] B. Liu, L. Wang, J. Wang, and J. Zhang, "Dual discriminator weighted mixture generative adversarial network for image generation," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 8, pp. 10013–10025, 2023. doi: 10.1007/s12652-021-03667-y.

[42] T. Zhang, W. Zhang, Z. Zhang, and Y. Gan, "PFGAN: Fast transformers for image synthesis," *Pattern Recognit. Lett.*, vol. 170, pp. 106–112, 2023. doi: 10.1016/j.patrec.2023.04.013.