Learning Fast Converging, Effective Conditional Generative Adversarial Networks with a Mirrored Auxiliary Classifier

Zi Wang
The University of Tennessee, Knoxville
zwang84@vols.utk.edu

Abstract

Training conditional generative adversarial networks (GANs) has been remaining as a challenging task, though standard GANs have developed substantially and gained huge successes in recent years. In this paper, we propose a novel conditional GAN architecture with a mirrored auxiliary classifier (MAC-GAN) in its discriminator for the purpose of label conditioning. Unlike existing works, our mirrored auxiliary classifier contains both a real and a fake node for each specific class to distinguish real samples from generated samples that are assigned into the same category by previous models. Comparing with previous auxiliary classifier-based conditional GANs, our MAC-GAN learns a fast converging model for high-quality image generation, taking benefits from its robust, newly designed auxiliary classifier. Experiments on multiple benchmark datasets illustrate that our proposed model improves the quality of image synthesis compared with state-of-the-art approaches. Moreover, much better classification performance can be achieved with the mirrored auxiliary classifier, which can in turn promote the use of MAC-GAN in various transfer learning tasks.

1. Introduction

Generative adversarial networks (GANs) [8, 20, 25, 38, 23, 21], as one of the most interesting topics of deep learning [14, 15, 30, 34, 10, 41, 43, 6], have gained tremendous attention and developed substantially in recent years. Unlike other kinds of generative models with explicitly probability density functions such as PixelCNN and PixelRNN [39, 26, 32] and Variational Autoencoders (VAE) [12, 35], GANs can generate high-quality samples with implicit densities. Generally, GANs synthesize images via a two-player (a generator and a discriminator) min-max game: the generator takes as input latent variables and generates samples that can fool the discriminator to the largest extent; the discriminator learns to distinguish fake samples generated by the generator from real samples. The two players continuously compete and evolve until an equilibrium is reached in the end.

In general, standard GANs are only capable of generating images that fall into random classes because the relation between the input latent variables and image labels is intractable. Therefore, the conditional generative adversarial network [20, 25, 44], which incorporates the label information during training, is designed for allowing the generated images to have a certain type. The auxiliary classifier GAN (AC-GAN) [25] is one of the most successful variants of conditional GAN. Besides predicting the image source (real or fake), it introduces an extra classifier in the discriminator to predict class labels as well (Fig. 1(left)). It has been validated that AC-GAN is capable of generating high-quality images with large resolution. Moreover, another important advantage of AC-GAN is that the trained auxiliary classifier can be further used in classification transfer learning tasks to acquire better performance.

However, training AC-GAN is quite challenging because new constraints are introduced by the auxiliary classifier. When training the discriminator, both real and fake samples are fed into the neural network in each iteration. Usually, there exist large gaps between the generator distribution (fake samples) and the target distribution (real samples), especially at early training stages. This issue makes AC-GAN difficult to train and converge because both real and fake samples are assigned to the same class and confuse the classifier. To mitigate this problem, fast converging GAN (FC-GAN) [17] proposes to introduce an extra 'fake' class in the auxiliary classifier and all generated images are categorized into this class rather than assigning them to a specific class label (Fig. 1(middle)). By doing so, the convergence of the auxiliary classifier is boosted. However, there is still room for improvement of FC-GAN. As the neural networks evolve, the generator distribution and the target distribution get closer and closer. The 'fake class' that handles all fake samples, which are usually with multimodal distributions, becomes a barrier that prevents the auxiliary classifier from getting converged well (see Section 5 for detailed analysis).

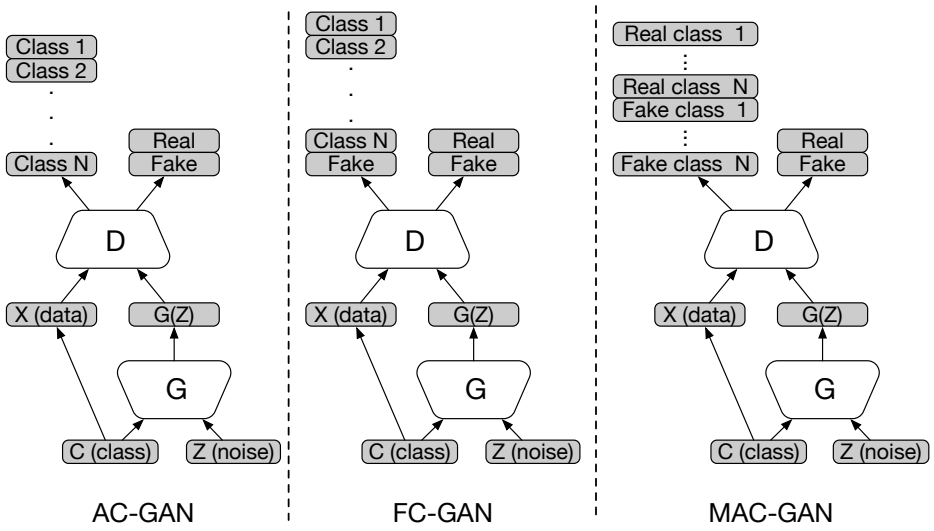


Figure 1. The architectures of AC-GAN (left), FC-GAN (middle), and the proposed MAC-GAN (right).

The deterioration of the classification performance further affects the generated image quality.

To resolve the above issues, we propose a novel conditional GAN architecture with a mirrored auxiliary classifier (MAC-GAN), which in its discriminator contains both a real and a fake class for each category to distinguish real images from generated images that fall into the same category (Fig. 1(right)). In this way, real samples and generated samples of the same category, which are usually with different distributions, are categorized into two separate classes in the classifier. Therefore, the discriminator can converge fast and well. We evaluate the proposed MAC-GAN on three widely used benchmark datasets, namely, MNIST [16], CIFAR-10 [13], and ILSVRC-2012 (ImageNet) [7]. Extensive experimental results and analyses demonstrate that MAC-GAN can generate high-quality, large resolution images compared with recent state-of-the-art works. Furthermore, we believe that the resulting robust classifier with high classification accuracy can also be used in transfer learning tasks.

The rest of this paper is organized as follows. A brief literature review is presented in Section 2. The details of the proposed MAC-GAN is described in Section 3. Experimental configurations and results are presented and analyzed in Section 4. Section 5 discusses the effectiveness of the proposed model and points out some future directions that can be extended from this study. Finally, Section 6 concludes the paper.

2. Related works

2.1. Generative adversarial networks.

GAN was first introduced by Ian Goodfellow in [8] and has become the most popular generative framework in recent years. Although the vanilla GAN is hard to train and can only model distributions of samples from simple datasets and generate low-resolution images in the early years, a huge amount of improvements on GAN's architectures [28], loss functions [2, 36], and training strategies [24] have been proposed. DCGAN [28] extends the vanilla GAN with convolutional neural networks and provides a series of detailed hyperparameters and achieves remarkable performance with the LSUN [45] and ImageNet-1k datasets. In [2], the wasserstein distance is used as the metric for the measurement of the similarity between the generated and the real samples and the performance is significantly improved. GAN has also been widely implemented in various areas of applications, including image synthesis [25, 17], text to image generation [47], domain adaptation [33, 3], biomedical imaging [27, 40], and high-resolution face generation [49, 18, 48].

2.2. Conditional extension of GANs.

Conditional GANs [20, 29, 25, 17] are built on the basis of GANs, which incorporates class labels for class conditional image generation [22]. CGAN [20] is the first proposed conditional GAN model in which class labels are embedded to the input latent variables or the feature maps of some middle layers of the generator. Therefore, images with certain types can be generated.

Rather than concatenating class labels with latent variables, AC-GAN [25, 11] introduces an auxiliary classifier in the discriminator to predict class labels, allowing the generated images to have a certain type. Although AC-GAN can generate high-quality images with large resolution, the auxiliary classifier makes AC-GAN difficult to train. This is because both the real and fake data of a certain type, which are usually with different distributions, are assigned to the

same class and confuse the discriminator. To deal with this problem, fast-converging GAN (FC-GAN) [17] proposes to add an extra 'fake' class in the auxiliary classifier. All generated samples are assigned to the 'fake' class rather than a specific label. By doing so, the discriminator's loss can converge much faster. However, as the generator distribution and target distribution get closer and closer, FC-GAN usually has difficulties in converging to the equilibrium. This is because the generated samples exhibit multimodal distributions as the sample quality get better but are assigned to the same 'fake' class. To resolve the above issues, our MAC-GAN introduces a real class and a fake class for each label in the auxiliary classifier. As a result, all samples (real and fake) can be properly handled by the discriminator. High-quality images can be generated, benefiting from the fast converging discriminator with a robust classifier.

Several other types of conditional GAN framework are also proposed recently. For example, [22] improves cGAN with an inner projection module in the discriminator. SAGAN [46] introduces a self-attention mechanism for the conditional image generation. Since we focus on improving the performance of auxiliary classifier-based conditional GANs, these works are beyond the scope of our study.

3. Methodology

In this section, we first briefly introduce necessary preliminaries of GANs and then describe the rationale of our proposed MAC-GAN. The architecture design of the generator and the discriminator is presented in the experiment setup section.

3.1. Preliminaries

A standard GAN consists of two neural networks, a generator G and a discriminator D, which are parameterized by θ_g and θ_d, respectively. The training process of a standard GAN can be considered as a two-player min-max game. The generator takes a latent variable $\mathbf{z}$ as the input and generates an image $G(\mathbf{z})$ that tries to mislead the discriminator into recognizing it as a real sample. The discriminator is trained to distinguish real samples from generated samples. GAN's objective function can be formulated as Eq. (1).

$$\mathcal{L}_{GAN} = \mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})} \log\left[D(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})} \log\left[1 - D(G(\mathbf{z})\right], \tag{1}$$

where $p_r(\mathbf{x})$ and $p_z(\mathbf{z})$ are the prior distributions of real samples $\mathbf{x}$ and latent variables $\mathbf{z}$, respectively. In the rest of the paper, we omit the subscripts of $\mathbb{E}$ for notational simplicity.

cGAN [20] is the first conditional extension of GAN, which naively concatenating class label information with both the inputs of the generator (latent variables) and the discriminator (images). Therefore, samples conditioned by class labels are generated without any change of GAN's objective function. Rather than embedding labels as part of the inputs, AC-GAN [25, 17] introduces an auxiliary classifier for the purpose of class prediction. As a result, the discriminator outputs two probabilities over class and source (real or fake). The objective function is correspondingly modified by combining the classification loss (also known as auxiliary classifier loss) and source loss (also known as adversarial loss). FC-GAN works in similar way, except that an extra 'fake' class is introduced in the auxiliary classifier to boost the convergence of the discriminator.

3.2. The proposed MAC-GAN

As mentioned previously, both AC-GAN and FC-GAN have difficulties in their loss convergence. We propose MAC-GAN, an improved version of AC-GAN and FC-GAN, to solve this problem. The architecture of the proposed MAC-GAN is presented in Fig. 1(right). The auxiliary classifier in MAC-GAN represents a probability distribution over $2N$ class labels, where N is the number of classes of the original dataset. Specifically, the classifier contains both a fake node and a real node for each class. In this way, samples with the same class label but from different sources (i.e., real or fake) are categorized into two different classes. Let $\mathbf{x}^r$ and $\mathbf{x}^f$ denote the real samples from the original dataset and the fake samples from the generator, $P(s|\mathbf{x})$ and $P(c|\mathbf{x})$ denote the distributions over sources and class labels, respectively. We define the objective functions of MAC-GAN as follows.

Source loss. Similarly as the standard GAN, for the source loss, the generator is optimized to generate images that fool the discriminator (Eq. (2)).

$$\mathcal{L}_s^G = -\mathbb{E}\left[\log P(s = real|\mathbf{x}^f)\right]. \tag{2}$$

The discriminator aims to distinguish real images from generated images, which is defined as Eq. (3).

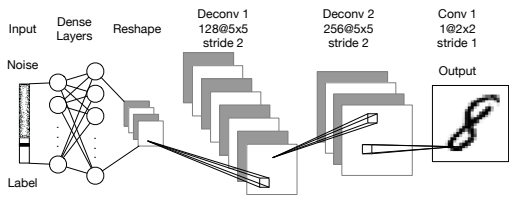
$$\mathcal{L}_s^D = -\mathbb{E}\left[\log P(s = real|\mathbf{x}^r)\right] - \mathbb{E}\left[\log P(s = fake|\mathbf{x}^f)\right]. \tag{3}$$

Classification loss. Suppose the auxiliary classifier represents a distribution over $2N$ nodes:

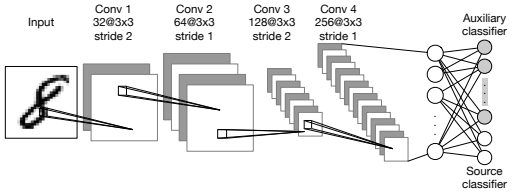
$$\left[\mathbf{C}^r, \mathbf{C}^f\right] = \left[c_1^r, c_2^r, \cdots, c_N^r, c_1^f, c_2^f, \cdots, c_N^f\right],$$

where N is the number of classes, $c_i^{r/f} (i = 1, 2, \cdots, N)$ represent class labels for the real and fake samples. For the classification loss, the discriminator is trained to assign each real sample to the correct real class and each generated sample to the correct fake class, respectively, which is defined as Eq. (4).

$$\mathcal{L}_c^D = -\mathbb{E}\left[\log P(C = c_i^r|\mathbf{x}^r)\right] - \mathbb{E}\left[\log P(C = c_i^f|\mathbf{x}^f)\right]. \tag{4}$$



(a) Generator



(b) Discriminator

Figure 2. The architectures of the proposed MAC-GAN's generator and discriminator.

On the other hand, the generator is trained to maximize the probability to assign each generated image to the correct real class rather than its corresponding fake class. The loss function of the generator is formulated as Eq. (5).

$$\mathcal{L}_c^G = -\mathbb{E}\left[\log P(C = c_i^r|\mathbf{x}^f)\right]. \tag{5}$$

Finally, we combine the source loss and classification loss of both the generator and the discriminator to get MAC-GAN's objective functions, i.e., $\mathcal{L}^G = \mathcal{L}_s^G + \mathcal{L}_c^G$, $\mathcal{L}^D = \mathcal{L}_s^D + \mathcal{L}_c^D$.

4. Experimental results

We first describe our implementation details of the proposed MAC-GAN, including the datasets used for training, network architecture design and hyperparameter selection. Then we thoroughly evaluate the performance of our proposed MAC-GAN by comparing it with other well-known auxiliary classifier-based conditional GAN models, i.e., AC-GAN and FC-GAN from several perspectives.

4.1. Setup

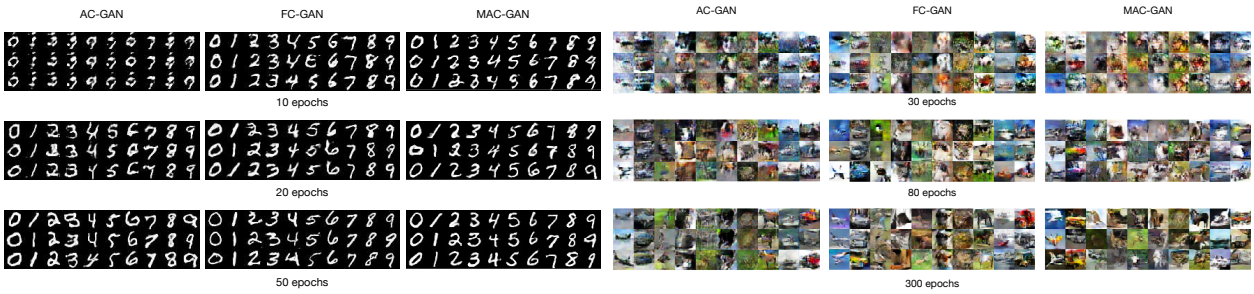
Dataset. We use three widely used benchmark dataset for image classification to evaluate the performance of our proposed MAC-GAN, i.e., MNIST [16], CIFAR-10 [13], and ILSVRC-2012 (ImageNet) [7]. The MNIST dataset contains 60,000 training samples and 10,000 test samples of handwritten digits from 0 to 9. All the images has a fixed resolution of 28x28. The CIFAR-10 has a training set of 50,000 images and a test set of 10,000 images of 10 different classes. The resolution of the images is 32x32. The ILSVRC-2012 (ImageNet) dataset is a large-scale dataset with more than a million training samples of 1,000 classes. The validation set contains 50,000 images with 50 samples for each class. In our experiments, all the samples are cropped to a fixed image size of 224x224 before feeding into the neural network. More specifically, training samples are randomly cropped for the purpose of data augmentation, and test samples are center cropped.

For MNIST and CIFAR-10 we use all the training samples to train a single model and use the whole test set to test the performance of the auxiliary classifier. For ILSVRC-2012, we follow the strategy in [25], i.e., training multiple models on images of 10 classes and ensemble them together to reduce the variability and improve the performance.

Network architecture. In our study, we focus on the lightweight implementations of conditional GANs, which have benefits such as reductions of power consumption and fast inference [9, 42]. The detailed architecture of the proposed MAC-GAN is shown in Fig. 2. For the MNIST and CIFAR-10 dataset, we implement a discriminator with four convolutional layers, followed by a fully connected layer. A leaky ReLU (with a negative slope of 0.3) activation and a dropout (with a probability of 0.3) layer are used after all of the convolutional layers. For the generator, we first sample the latent variable $\mathbf{z}$ with a dimension of 100 from a uniform distribution $[-1, 1]$ and embed the label information c with the latent variable. Two dense layers are first implemented to extend the dimensions and the output are reshaped into a 4-dimensional tensor followed by two deconvolutional layers and a convolutional layer. ReLU is used as the activation function except for the last layer after which the tanh activation is used. For the ILSVRC-2012 dataset, the architecture is similar except that we add two more convolutional and deconvolutional layers in the discriminator and generator, respectively. This is because the resolution of the samples from the ILSVRC-2012 dataset is much larger than those in the MNIST and CIFAR-10 datasets. To achieve fair comparison, we use the same architectures as described above for AC-GAN and FC-GAN, except for the output layer of the discriminator.

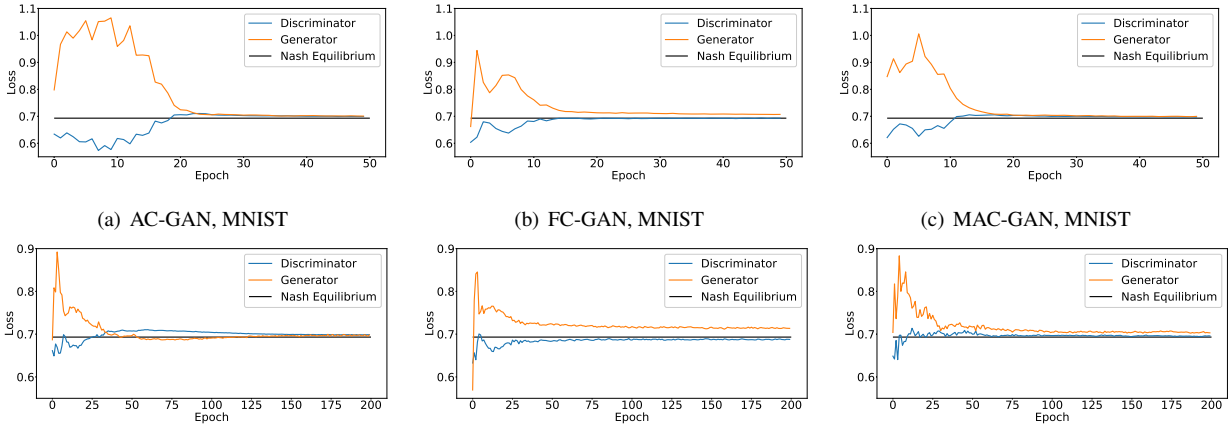
Training details. We train both the generator and the discriminator for 100 epochs for the MNIST dataset, and 600 epochs for the CIFAR-10 and ILSVRC-2012 dataset, with a batch size of 100, respectively. We use the Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) for both the generator and the discriminator, with a learning rate of 0.00002 for all three datasets. The Python deep learning library Keras [5] with Tensorflow [1] as the backend is used to conduct the experiments.

Evaluation strategy. We compare MAC-GAN with other state-of-the-art auxiliary classifier-based conditional GAN implementations, namely, AC-GAN and FC-GAN. The performance is evaluated and compared qualitatively from the perspective of visual fidelity, and quantitatively from the perspective of parzen window estimates, inception score, and Frechet Inception Distance (FID) score. Furthermore, we also evaluate the classification performance with the auxiliary classifiers of all the above models.



(a) MNIST (b) CIFAR-10

Figure 3. Visual fidelity of generated images. It takes much less training iterations for FC-GAN and MAC-GAN to generate recognizable samples than AC-GAN. The sample quality of MAC-GAN is the best among all three models.



(a) AC-GAN, MNIST (b) FC-GAN, MNIST (c) MAC-GAN, MNIST

(d) AC-GAN, CIFAR-10 (e) FC-GAN, CIFAR-10 (f) MAC-GAN, CIFAR-10

Figure 4. Loss curves over epochs of the generators and discriminators of AC, FC, and MAC-GAN with the MNIST and CIFAR-10 datasets.

4.2. Performance with MNIST and CIFAR-10

4.2.1 Visual fidelity and loss curves

Visual fidelity is the most intuitive metric to evaluate the performance of image generation. In Fig. 3 we present the generated images at different training stages of AC-GAN, FC-GAN, and MAC-GAN, respectively. We observe that both FC-GAN and MAC-GAN generate recognizable samples much earlier than AC-GAN. For example, after 10 epochs training with MNIST, both MAC-GAN and FC-GAN can generate relatively clear images while AC-GAN cannot. After the training is done, the samples generated by MAC-GAN present the best visual fidelity with both the MNIST and the CIFAR-10 datasets compared with the other two models. Samples from MAC-GAN are clear and with substantial diversity, but those from the other two models are more or less blur, with artifacts.

We further analyze the performance by plotting the loss curves of the generator and the discriminator over training epochs. As shown in Fig. 4, our proposed MAC-GAN converges better and faster as compared with AC-GAN and FC-GAN. More specifically, for the MNIST dataset, MAC-GAN's generator and discriminator converges to the Nash equilibrium after around 15 epochs. On the other hand, it takes about 25 epochs for AC-GAN to get converged. With the datasets of MNIST and CIFAR-10, both MAC-GAN and AC-GAN converge closer enough to the Nash equilibrium. However, FC-GAN cannot achieve comparable performance on loss convergence, which is consistent with our analysis in the previous sections.

These results qualitatively indicate that the proposed MAC-GAN achieves comparable performance of convergence speed as FC-GAN, and the best performance on visual fidelity among all three models. The effectiveness of MAC-GAN is because it takes both the advantage of faster convergence from FC-GAN and better convergence from AC-GAN.

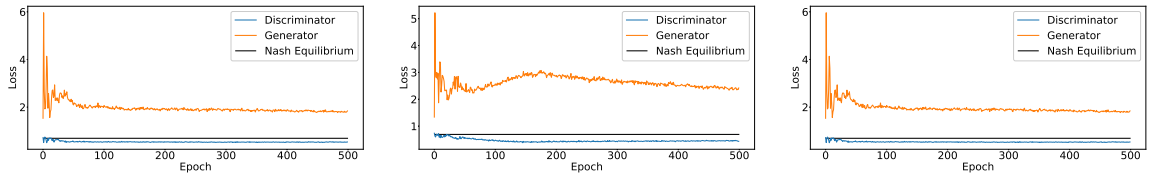
4.2.2 Quantitative measurement of performance

We use three popular quantitative metrics to evaluate the image generation performance of MAC-GAN, i.e., parzen window estimate, inception score and Frechet Inception Distance (FID) score.

Parzen window estimate is used as an alternative to approximate GAN's exact likelihood, which is originally intractable [4]. We follow the procedure introduced by the

Dataset	Criterion	Epoch	Methods			
			AC-GAN	FC-GAN	MAC-GAN	Validation set
MNIST	Parzen window	10	-109.9 ± 1.81	$\mathbf{85.3 \pm 1.73}$	55.8 ± 1.86	238.8 ± 2.14
		20	75.6 ± 1.74	156.1 ± 1.54	$\mathbf{159.3 \pm 1.44}$	
		50	141.8 ± 1.51	165.2 ± 1.48	$\mathbf{166.3 \pm 1.43}$	
		100	154.2 ± 1.44	159.7 ± 1.46	$\mathbf{165.2 \pm 1.41}$	
	Inception score	10	1.17 ± 0.09	1.68 ± 0.15	$\mathbf{1.87 \pm 0.15}$	2.14 ± 0.03
		20	1.79 ± 0.21	1.74 ± 0.13	$\mathbf{1.91 \pm 0.10}$	
		50	1.84 ± 0.16	1.82 ± 0.14	$\mathbf{1.97 \pm 0.13}$	
		100	1.97 ± 0.13	1.99 ± 0.15	$\mathbf{2.06 \pm 0.16}$	
CIFAR-10	Parzen window	50	571.0 ± 5.30	$\mathbf{576.6 \pm 5.39}$	510.3 ± 5.32	752.9 ± 4.83
		100	580.4 ± 5.26	624.6 ± 5.43	$\mathbf{654.8 \pm 4.98}$	
		300	615.1 ± 5.56	615.4 ± 5.38	$\mathbf{684.9 \pm 4.96}$	
		600	611.9 ± 5.49	640.0 ± 5.26	$\mathbf{693.4 \pm 4.98}$	
	Inception score	50	2.76 ± 0.09	$\mathbf{4.21 \pm 0.07}$	2.58 ± 0.08	11.05 ± 0.34
		100	3.10 ± 0.08	$\mathbf{5.01 \pm 0.15}$	4.05 ± 0.12	
		300	3.69 ± 0.10	$\mathbf{5.29 \pm 0.08}$	5.19 ± 0.19	
		600	4.22 ± 0.12	5.57 ± 0.22	$\mathbf{5.81 \pm 0.18}$	
	FID	50	92.86	**70.66**	100.60	7.96
		100	83.22	62.41	**58.33**	
		300	67.84	47.26	**43.47**	
		600	64.37	39.56	**35.50**	

Table 1. Quantitative measurements of the generated images by AC-GAN, FC-GAN, and the proposed MAC-GAN with the MNIST and CIFAR-10 datasets.



(a) AC-GAN, ILSVRC-2012 (b) FC-GAN, ILSVRC-2012 (c) MAC-GAN, ILSVRC-2012

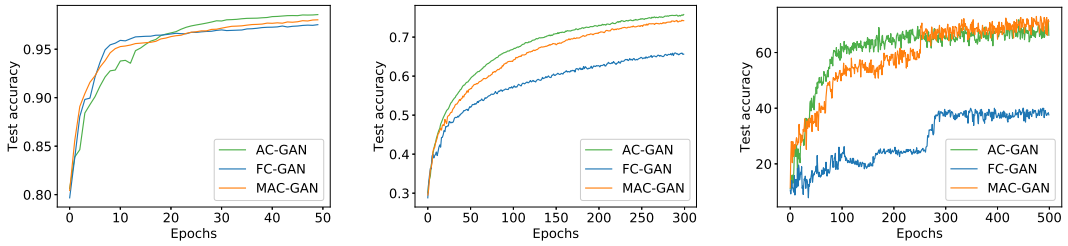
Figure 5. Loss curves over epochs of the generators and discriminators of AC, FC, and MAC-GAN with the ILSVRC-2012 dataset.

initial implementation of GAN [8] for parzen window estimates. We first fit a Gaussian parzen window to the samples from the validation set to obtain the σ parameter of the Gaussian distribution, with 10-fold cross validation. Then we use the selected σ to obtain the parzen window estimates with randomly generated samples from the generator. The results are listed in Table 1. Besides the measurements of the generated samples, we also provide the scores calculated with the test/validation set as a reference, which can be considered as the performance upper bounds. During the training process on both datasets, MAC-GAN generates images with higher scores compared with AC-GAN and FC-GAN. For example, the parzen window estimate of MAC-GAN with MNIST is 165.2 after convergence, while the scores of AC-GAN and FC-GAN are only 154.2 and 159.7, respectively. After training for 300 epochs on the CIFAR-10 dataset, the parzen window estimate of MAC-GAN gets quickly saturated to 684.9, which is 11.3% higher than AC-GAN and FC-GAN.

Inception score [37], which is with a high correlation with human's subjective judgment of image quality [31], is another widely used approach to measure GAN's performance. Here we follow the procedure in [31, 25] for the calculation of inception score. From the results presented in Table 1 we observe that for the MNIST dataset, MAC-GAN performs better at all stages of the training process. After the training process is completed, the inception score of the images generated by MAC-GAN is 2.06, which is only 3.7% less than the score obtained with the validation set. On the other hand, AC-GAN and FC-GAN only achieve inception scores that are 7.0% and 7.9% less than the score obtained with the validation set, respectively. For the CIFAR-10 dataset, FC-GAN performs slightly better than MAC-GAN at the early stages. The reason behind this phenomenon may be that samples from the CIFAR-10 dataset are with much more variability, it is easier for FC-GAN's auxiliary classifier to get converged at early stages because all of the low quality, similar-looking, fake samples are assigned to a single fake class. However, as the quality of the generated images get improved gradually, MAC-GAN performs the best finally, which is more essential compared with the results at early stages. After training

Dataset	Criterion	Epoch	Methods			
			AC-GAN	FC-GAN	MAC-GAN	Validation set
ILSVRC-2012	Parzen window	100	284.9 ± 8.09	377.6 ± 8.57	$\mathbf{440.9 \pm 8.67}$	879.4 ± 6.98
		200	332.6 ± 9.88	501.2 ± 9.17	$\mathbf{520.7 \pm 8.65}$	
		500	397.5 ± 8.37	523.7 ± 6.08	$\mathbf{531.1 \pm 8.48}$	
	Inception score	100	3.80 ± 0.13	3.83 ± 0.10	$\mathbf{3.93 \pm 0.07}$	15.38 ± 0.59
		200	$\mathbf{4.72 \pm 0.14}$	4.10 ± 0.12	4.41 ± 0.10	
		500	10.50 ± 0.27	10.94 ± 0.23	$\mathbf{11.09 \pm 0.23}$	

Table 2. Quantitative measurements (parzen window estimate and inception score) of the generated images by AC-GAN, FC-GAN, and the proposed MAC-GAN with the ILSVRC-2012 dataset.



(a) MNIST (b) CIFAR-10 (c) ILSVRC-2012

Figure 6. Training accuracy with the auxiliary classifiers in the discriminators.

all the models for 600 epochs, the inception score obtained with MAC-GAN (5.81) significantly outperforms those obtained with AC-GAN (4.22) and FC-GAN (5.57).

To further evaluate the generated sample quality, we introduce the FID score for the comparison on CIFAR-10. FID is another popular metric that computes the distance between feature vectors calculated for real and fake samples. Similar to other measurements, it can be observed that the FID scores of the samples generated by MAC-GAN are lower than the samples obtained with AC and FC-GAN, which indicates that MAC-GAN generates images with smaller distances compared with the real images.

4.3. Performance comparison with large scale dataset (ILSVRC-2012)

To further evaluate the performance of MAC-GAN, we conduct experiments with the large scale dataset ILSVRC-2012. Fig. 5 shows that MAC-GAN achieves better convergence compared with AC-GAN and FC-GAN for both the generator and the discriminator. It is worth mentioning that compared with the MNIST and CIFAR-10 datasets, all three models cannot completely converge to the Nash equilibrium with ILSVRC-2012. This is because ILSVRC-2012 contains millions of training samples with much larger resolution and diversity, and training a GAN with ILSVRC-2012 is a much more challenging task. Even though, the loss of MAC-GAN still reduces to a relatively lower value in a shorter period of time compared with AC and FC-GAN.

We also quantitatively measure the parzen window estimate and the inception score of MAC-GAN at different training stages with ILSVRC-2012 (Table 2). For parzen window estimate, MAC-GAN obtains a score of 440.9 after training for only 100 epochs, while the scores of AC-GAN and FC-GAN are only 284.9 and 377.6, respectively. After 500 epochs, the parzen window estimate of MAC-GAN is 531.1, which is the best among all three models. For inception score, MAC-GAN achieves a score of 11.09 after training the model for 500 epochs, while the scores of the other two models are less than 11. These results validate the effectiveness of MAC-GAN on ILSVRC-2012.

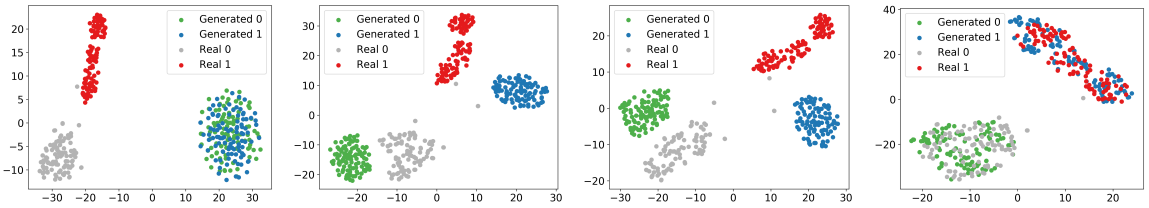
Dataset	Methods		
	AC-GAN	FC-GAN	MAC-GAN
MNIST	99.04%	98.19%	98.60%
CIFAR-10	78.50%	69.32%	77.60%
ILSVRC-2012	71.60%	38.40%	71.80%

Table 3. Test accuracy of the auxiliary classifiers.

4.4. Performance of the auxiliary classifiers

Since our proposed MAC-GAN focuses on the improvement of the auxiliary classifier, we further analyze the classifier's performance by plotting the classification accuracy during the training process. The results are shown in Fig. 6. We observe that the classification performance of FC-GAN is better at the early stages. During this time, for AC-GAN, the distribution of the images from each class is highly disturbed by the low quality generated images. Therefore, it does not work well. As the generator's performance gets better and better, the generated samples become to benefit AC-GAN's auxiliary classifier but perturb FC-GAN's classifier, especially the extra 'fake' class. MAC-GAN takes advantage of both AC-GAN and FC-GAN at different training stages and therefore achieves outstanding performance.

We further evaluate the trained auxiliary classifier with the test (validation) sets of the three datasets and the results



(a) Epoch 1 (b) Epoch 15 (c) Epoch 30 (d) Epoch 50

Figure 7. Distribution of generated and real MNIST samples with t-SNE visualization.

are presented in Table 3. We can see that for the MNIST and CIFAR-10 datasets, MAC-GAN's classifier achieves comparable performance with AC-GAN, which is much better than FC-GAN. For the ILSVRC-2012 dataset, a test accuracy of 71.80% is achieved by MAC-GAN's classifier, which is the best among all three models. These results validate the effectiveness of our modified auxiliary classifier. We believe better performance can be obtained in transfer learning tasks by using MAC-GAN's auxiliary classifier.

5. Discussion and future work

To justify the effectiveness of MAC-GAN for the mitigation of the multi-modal issue during the training of auxiliary-based conditional GANs, we visualize both the real training samples and generated samples at different training stages via t-SNE [19] (Fig. 7, using the samples of digits 0 and 1 in MNIST for illustration). It can be observed that at the beginning, fake samples from different classes are clustered together, which indicates that FC-GAN will work better at early training stages (Fig. 7(a)). However, during most of the time, fake and real samples from different categories are separated from each other with clear margins (Fig. 7(b,c)), which explains that MAC-GAN outperforms other approaches by assigning real and fake samples of each class to an individual node, respectively. By assigning either all the fake samples to a single 'fake' class (FC-GAN), or the real and fake samples that belong to the same category to a single class (AC-GAN) harms the performance of the auxiliary classifier. At the end of the training, real and fake samples from the same class mixed with each other, indicating AC-GAN's loss is more reasonable. However, since the generator has almost converged at this time, there is little impact on the sample quality.

We use several fake and real samples of from MNIST to have a further understanding of the effectiveness of MAC-GAN (Fig. 8). For each image, we present the top-5 predictions of the auxiliary classifier. We see that MAC-GAN and AC-GAN always make correct predictions while FC-GAN tends to mix up the images' true classes with the extra 'fake' class. This is because generated samples with a multimodal distribution are assigned to a single class, and prevent FC-GAN's classification loss from getting converged. In contrast, MAC-GAN resolves this issue by adding an fake node

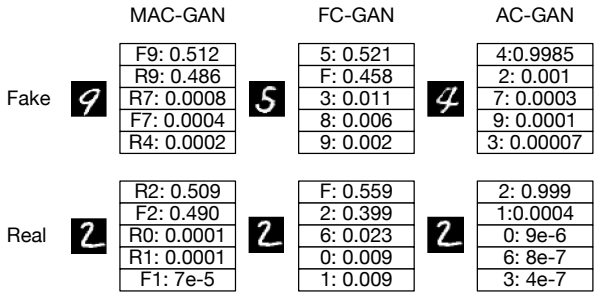


Figure 8. Top 5 predictions of AC, FC, and MAC-GAN's auxiliary classifier on real and fake samples. R: real class, F: fake class.

for each class and achieves comparable performance as AC-GAN, but with a much faster convergence speed.

Our future work will focus on the following aspects. (1) From theoretical analysis and experiment results we find that AC-GAN, FC-GAN, and our proposed MAC-GAN exhibit their advantage at different training stages. FC-GAN works better than the other two at the early stages, but MAC-GAN catches up later. Therefore, an intuitive assumption is that, if we can build a conditional GAN that fuses the objective functions of these three models, a faster and more robust generative model would be achieved. (2) Since MAC-GAN's auxiliary classifier contains both a real node and a fake node for each class label, which functionally includes the other part of the discriminator, i.e., the real/fake prediction. There is reason to believe that the adversarial loss can be removed while maintaining the performance of the discriminator, which could help to stabilize the GAN's training. (3) We will also focus on using MAC-GAN for other challenging computer vision tasks.

6. Conclusion

In this paper, we proposed mirrored auxiliary classifier GAN (MAC-GAN), which is a conditional GAN model with a newly designed auxiliary classifier. The classifier contains both a real and a fake class for each label to properly handle all the real and fake samples. By doing so, the generative model can get converged fast and well, which benefits greatly from the robust mirrored auxiliary classifier. Extensive experiment results and evaluations validated that MAC-GAN can achieve state-of-the-art image generation performance.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017.

[4] Olivier Breuleux, Yoshua Bengio, and Pascal Vincent. Quickly generating representative samples from an rbm-derived process. *Neural computation*, 23(8):2058–2073, 2011.

[5] François Chollet et al. Keras. `https://keras.io`, 2015.

[6] Eric M Christiansen, Samuel J Yang, D Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'Neil, Kevan Shah, Alicia K Lee, et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell*, 173(3):792–803, 2018.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[11] Takuhiro Kaneko, Yoshitaka Ushiku, and Tatsuya Harada. Label-noise robust generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2476, 2019.

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] Chengcheng Li, Zi Wang, and Hairong Qi. Fast-converging conditional generative adversarial networks for image synthesis. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2132–2136. IEEE, 2018.

[18] Jianshu Li, Jian Zhao, Yunpeng Chen, Sujoy Roy, Shuicheng Yan, Jiashi Feng, and Terence Sim. Multi-human parsing machines. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 45–53, 2018.

[19] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[21] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[22] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.

[23] Gonçalo Mordido, Haojin Yang, and Christoph Meinel. microbatchgan: Stimulating diversity with multi-adversarial discrimination. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3061–3070, 2020.

[24] Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. Dual discriminator generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2670–2680, 2017.

[25] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017.

[26] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

[27] Anton Osokin, Anatole Chessel, Rafael E Carazo Salas, and Federico Vaggi. Gans for biological image synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2233–2242, 2017.

[28] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[29] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[32] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[33] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[34] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[35] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, pages 3738–3746, 2016.

[36] Yuxin Su, Shenglin Zhao, Xixian Chen, Irwin King, and Michael Lyu. Parallel wasserstein generative adversarial nets with multiple discriminators. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3483–3489. AAAI Press, 2019.

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[38] Quan Hoang Trung Le, Hung Vu, Tu Dinh Nguyen, Hung Bui, and Dinh Phung. Learning generative adversarial networks from multiple data sources. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

[39] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[40] Dali Wang, Zheng Lu, Yichi Xu, Zi Wang, Chengcheng Li, Anthony Santella, and Zhirong Bao. Cellular structure image classification with small targeted training samples. *bioRxiv*, page 544130, 2019.

[41] Zi Wang, Chengcheng Li, Huiru Shao, and Jiande Sun. Eye recognition with mixed convolutional and residual network (micore-net). *IEEE Access*, 6:17905–17912, 2018.

[42] Zi Wang, Chengcheng Li, Xiangyang Wang, and Dali Wang. Towards efficient convolutional neural networks through low-error filter saliency estimation. In *Pacific Rim International Conference on Artificial Intelligence*, pages 255–267. Springer, 2019.

[43] Zi Wang, Dali Wang, Chengcheng Li, Yichi Xu, Husheng Li, and Zhirong Bao. Deep reinforcement learning of cell movement in the early stage of c. elegans embryogenesis. *Bioinformatics*, 34(18):3169–3177, 2018.

[44] Han Xu, Pengwei Liang, Wei Yu, Junjun Jiang, and Jiayi Ma. Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators. In *Proc. Int. Joint Conf. Artif. Intell.*, pages 3954–3960, 2019.

[45] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[46] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.

[48] Jian Zhao, Junliang Xing, Lin Xiong, Shuicheng Yan, and Jiashi Feng. Recognizing profile faces by imagining frontal view. *International Journal of Computer Vision*, 128(2):460–478, 2020.

[49] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *Advances in neural information processing systems*, pages 66–76, 2017.