Can Forensic Detectors Identify GAN Generated Images?

Haodong Li*, Han Chen*, Bin Li*, Shunquan Tan†
* Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security,
National Engineering Laboratory for Big Data System Computing Technology,
College of Information Engineering, Shenzhen University, Shenzhen 518060, China
E-mail: lihaodong@szu.edu.cn, 2016130205@email.szu.edu.cn, libin@szu.edu.cn Tel/Fax: +86-755-22673509
† College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
E-mail: tansq@szu.edu.cn

***Abstract*—Generative adversarial network (GAN) has shown its powerful capability in generating photorealistic images. Although the generated images can fool human eyes, it is not clear whether they can evade the detection of forensic detectors, which aim to identify the originality and authenticity of images. In this paper, we investigate how forensic detectors perform in differentiating between GAN generated images and real images. We consider two kinds of approaches, one is intrusive and the other is non-intrusive, based on whether the GAN architecture is needed for performing detection. We have conducted extensive experiments on a celebrity face image dataset to evaluate the effectiveness of different approaches. The results and analyses show that the intrusive approach can detect GAN generated images but with a relatively high false alarm rate. The non-intrusive approach with features extracted from a VGG network is very effective for detecting GAN generated images when the training data is sufficient, but it still faces challenge when the training data and testing data are mismatched.**

I. Introduction

In recent years, generative models based on GAN [1] have attracted more and more attention in many applications, such as speech synthesis [2], image super-resolution [3], image translation [4], [5], and image inpainting [6]. The generative models are trained to generate samples that reproduce the same distribution of the training data. Ideally, the generative models are expected to create any plausible samples that are exactly similar to those coming from real world by improving the model designs and increasing the training data. Up to now, it has been reported that the training of GAN is more stable by employing improved designs of network architecture [7] or better distance metrics [8]–[11]. Because of such improvements, it is now possible to generate images with high quality and sufficient variation with GAN [12].

Since GANs can produce photorealistic images, it may lead to some potential security issues. For example, one can use the GAN generated images to counterfeit some personal information in social networks and cheat others, one can also employ the generated images as materials to falsify images or videos and spread fake messages. As the GANs are becoming more powerful, the generated images will be more similar to real ones, resulting in more serious problems. Therefore, it is important to differentiate between GAN generated images and real images. Although the generated images can fool human eyes, it is not clear whether they can fool forensic detectors. The main purpose of this paper is to study if there is any possible approach to detect GAN generated images.

In this paper, we consider two categories of approaches for the detection of GAN generated images. The first category is called intrusive approach. It means that the GAN architecture is available when constructing a detector. In this case, some modules of the GAN can be employed as a detector for identifying generated images. In contrast to the intrusive approach, the other category is called non-intrusive approach. In this case, we are unable to obtain any modules of GAN and have to design the detector ourself. Three possible non-intrusive approaches are investigated in this paper, which are based on face quality assessment [13], inception scores [14], and latent features from a trained VGG-16 network [15], respectively. The experimental results on the CelebA [16] image dataset show that the intrusive approach can effectively detect the GAN generated images from the same (or earlier) epoch as the training data. However, it would misclassify the real images and the generated images from later epochs. Among the non-intrusive approaches, the one based on VGG features is superior. It can accurately detect GAN generated images when the training data is sufficient. However, its performance is degraded when the training data and the testing data are mismatched.

The rest of this paper is organized as follows. Section II introduces the basic knowledge of GAN. Section III presents the intrusive detection approach and the corresponding results. Section IV reports and discusses the non-intrusive detection approaches. Finally, the concluding remarks are given in Section V.

This work was supported in part by the NSFC (61572329, 61772349, U1636202), Shenzhen R&D Program (JCYJ20160328144421330). This work was also supported by Alibaba Group through Alibaba Innovative Research (AIR) Program. *(Corresponding author: Bin Li)*

TABLE I
DETECTION ACCURACIES FOR DISCRIMINATORS OF DIFFERENT EPOCHS.

Discriminator	Φ_{15}	Φ_{17}	Φ_{19}	Φ_{21}	Φ_{23}	Φ_{25}
Real ACC	0.5500	0.7568	0.5333	0.6001	0.8015	0.6744
Fake ACC	0.9677	0.8115	0.9799	0.9551	0.8568	0.9404
Avg. ACC	0.7589	0.7842	0.7566	0.7776	0.8292	0.8074

II. GENERATIVE MODEL BASED ON GAN

GAN was first proposed by Goodfellow et al. [1]. There are a generator network and a discriminator network in a GAN. The generator tries to create samples that make the discriminator impossible to distinguish them from real ones, while the discriminator tries to classify generated samples and real samples. During the training stage, a game is played between the generator and the discriminator. When the discriminator is able to differentiate the generated samples and real samples, the generator adjusts its parameters to produce samples that are more similar to the real ones. And then, the discriminator adjusts its parameters to tell apart the two classes again. Theoretically, the generator eventually reproduces the distribution of real data, and the discriminator is deteriorated into random guesses.

In order to mitigate the difficulties of training GANs, many effective methods have been proposed, such as [7]–[11]. In this paper, we adopt DCGAN [7] and WGAN [10] as examples for study. DCGAN replaces the pooling layers with convolutional layers, employs batch normalization [17], removes the fully connected layers, and carefully sets the activation functions, making the GAN more stable to be trained than the vanilla GAN [1] and thus produce images with better visual quality. WGAN employs Wasserstein distance in the discriminator (critic) and improves the loss function, which helps the GAN to be more stable and avoids the collapse mode.

III. INTRUSIVE APPROACH

In this section, we introduce the intrusive approach for the detection of GAN generated images and show the corresponding experimental results.

An approach is regarded as intrusive if we can access to the GAN architecture that produces generated images. Since the GAN is available in this case, an intuitive way is to examine whether the discriminator of GAN can differentiate between generated images and real images. Ideally, the discriminator will fail to distinguish the two classes of images when the GAN is well trained. However, it is known that the discriminator is usually the dominant side in training process, meaning that it is still able to differentiate between generated images and real images even when the training is completed. Therefore, we evaluate the detection performance by using the discriminator of GAN as a forensic detector.

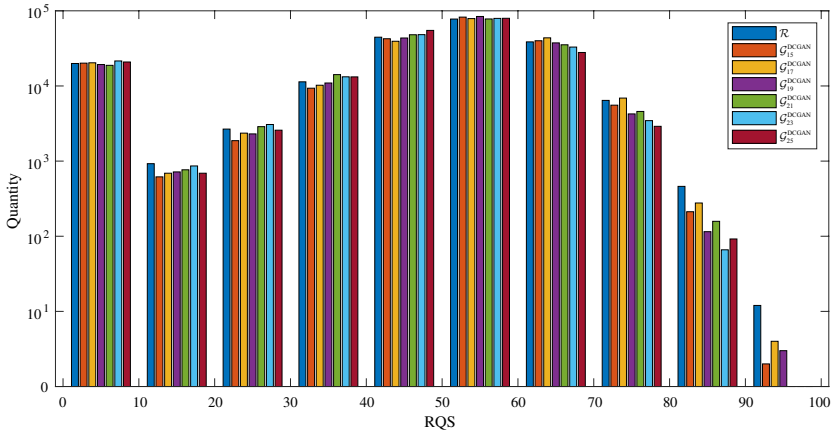
We conducted the experiments on the CelebFaces Attributes dataset (CelebA) [16]. This dataset consists of 202599 celebrity face images. We used the align&cropped PNG version of this dataset, and cropped the facial region from each image to remove the background and resized the cropped region into 128×128. The resulting image set was treated as real in our experiments, which is denoted as $\mathcal{R}$ for short. With the real image set $\mathcal{R}$, we trained a DCGAN [7] to create generated images. We used the same network architecture as introduced in [7], and adopted Adam optimizer and the learning rate of 0.0002 for training. The GAN was trained 25 epochs in total. After each epoch, we saved the weights of the GAN model, and used the model to produce a set of generated images with same number of images in CelebA. For simplicity, we denote the discriminator in the i-th epoch as Φ_i, and denote the generated image set generated in the i-th epoch as $\mathcal{G}_i^{\text{DCGAN}}$.

TABLE II
DETECTION ACCURACIES OF Φ_{21} ON IMAGES GENERATED FROM DIFFERENT EPOCHS.

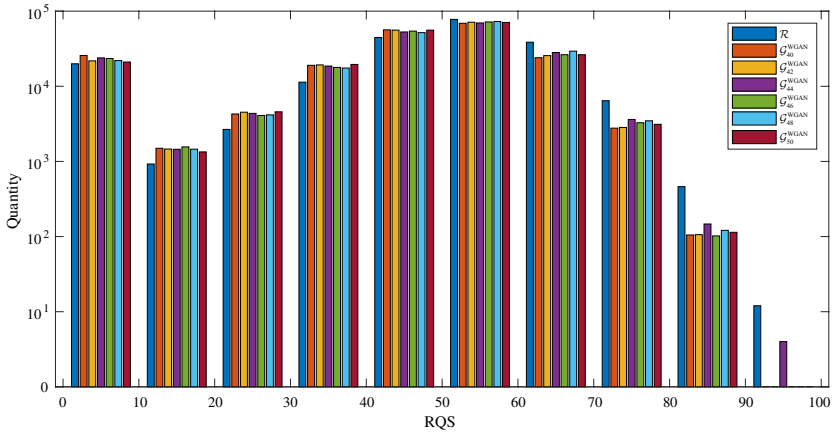
Testing Set	$\mathcal{G}_{17}^{\text{DCGAN}}$	$\mathcal{G}_{19}^{\text{DCGAN}}$	$\mathcal{G}_{21}^{\text{DCGAN}}$	$\mathcal{G}_{23}^{\text{DCGAN}}$	$\mathcal{G}_{25}^{\text{DCGAN}}$
ACC	0.9850	0.9903	0.9551	0.7781	0.7787

We first tested the performance by using Φ_i to classify images in $\mathcal{G}_i^{\text{DCGAN}}$ and $\mathcal{R}$. Since the images generated from early epochs usually contain visually unnatural artifacts, they can be easily identified by human eyes, thus we only conducted experiments for those later epochs. Specifically, $i \in \{15, 17, 19, 21, 23, 25\}$ were considered. Table I shows the detection results for discriminators in different epochs, where "Real ACC" means the testing accuracy for real images and "Fake ACC" means the testing accuracy for generated images. From this table, we can observe that the discriminator is bias towards the generated images, which are much more easier to be detected than real images. The accuracies for the 15th, 19th, 21st, 25th epochs are around 95%, while accuracies for the 17th, 23rd epochs are lower but still larger than 81%. For the real images, however, the discriminators failed to achieve satisfactory performance. In most of the cases, the detection accuracies for real images are around 50%-70%. The highest accuracy for real images is just over 80%, obtained by Φ_{23}. Such results indicate that the discriminator was not sufficiently optimized during the training of GAN, and thus it would make many false alarms if it is used to detect GAN generated images.

Although the discriminators can effectively detect generated images in the corresponding epoch, it is not so practical since we probably have no idea about which epoch the generated images come from. Therefore, it is necessary to assess the performance of discriminators for detecting generated images from different epochs. To this end, we used the discriminator Φ_{21} to detect the images generated in 17th, 19th, 21st, 23rd,



(a) DCGAN



(b) WGAN

Fig. 1. The histograms of RQS for different images sets.

and 25th epochs. The detection accuracies are shown in Table II. It is observed that the accuracies are over 98% for $\mathcal{G}_{17}^{\text{DCGAN}}$ and $\mathcal{G}_{19}^{\text{DCGAN}}$, while the accuracies drop to about 78% for $\mathcal{G}_{23}^{\text{DCGAN}}$ and $\mathcal{G}_{25}^{\text{DCGAN}}$. It means that a discriminator can achieve good performance when detecting generated images from earlier epochs, but tends to misclassify some generated images when they are from later epochs.

Based on above results, we can conclude that the discriminator of GAN can effectively detect the generated images from the same or earlier epochs, but is unable to accurately identify the real images and the generated images from later epochs, which would limit the applications of the intrusive approach. Hence, we need to study on more practical approaches.

IV. Non-intrusive Approach

In this section, we introduce non-intrusive approaches for detecting GAN generated images. For a non-intrusive approach, the GAN model is unavailable. Consequently, we need to seek for other measures or features to perform the detection. In the following subsections, we will discuss three non-intrusive approaches, which leverage face quality assessment, inception scores, and VGG based features, respectively. In addition to DCGAN, WGAN was also included in the following experiments regarding non-intrusive approaches. We used the same network architecture as that of DCGAN, but replaced the loss function and added weight clipping as described in [10]. Instead of the Adam optimizer, we used the RMSProp optimizer as it did in [10]. Similarly, the generated image set generated in the j-th epoch[1] of WGAN is denoted as $\mathcal{G}_j^{\text{WGAN}}$.

A. Detection based on Face Quality Assessment

Although GANs try to produce high quality images, as we known, there are still some disparities in image quality

[1]During the training, the parameter n_{critic} of WGAN was set as 5. Therefore, in every epoch the number of iterations of the discriminator is 5 times to that of the generator.

TABLE III
MEANS OF INCEPTION SCORE FOR REAL IMAGES AND GENERATED IMAGE.

Image set	$\mathcal{R}$	$\mathcal{G}_{15}^{\text{DCGAN}}$	$\mathcal{G}_{17}^{\text{DCGAN}}$	$\mathcal{G}_{19}^{\text{DCGAN}}$	$\mathcal{G}_{21}^{\text{DCGAN}}$	$\mathcal{G}_{23}^{\text{DCGAN}}$	$\mathcal{G}_{25}^{\text{DCGAN}}$
Inception score	3.3346	2.3314	2.3700	2.3979	2.4461	2.4270	2.4141
Image set	–	$\mathcal{G}_{40}^{\text{WGAN}}$	$\mathcal{G}_{42}^{\text{WGAN}}$	$\mathcal{G}_{44}^{\text{WGAN}}$	$\mathcal{G}_{46}^{\text{WGAN}}$	$\mathcal{G}_{48}^{\text{WGAN}}$	$\mathcal{G}_{50}^{\text{WGAN}}$
Inception score	–	2.3681	2.2807	2.3939	2.3864	2.3513	2.3902

between the GAN generated images and the real images. Therefore, it is possible to detect the GAN generated images by measuring the image quality. Due to face images are considered in this paper, we adopt the face quality assessment method proposed in [13] to measure image quality. The method in [13] is based on learning to rank [18], and it assigns a *rank based quality score* (RQS) for a face image. The higher the RQS, the better quality of the face image.

For the images in $\mathcal{R}$, $\mathcal{G}_i^{\text{DCGAN}}$, and $\mathcal{G}_j^{\text{WGAN}}$ (we investigated $i \in \{15, 17, 19, 21, 23, 25\}$ and $j \in \{40, 42, 44, 46, 48, 50\}$ in experiments), we used the method [13] to compute the RQS for each image, and then counted the histogram of RQS for each image set. Fig. 1 shows the obtained histograms. By comparing the histograms of GAN generated images to the histogram of real images, it is observed that there are only minor differences between the histograms, meaning that it is difficult to detect generated images based on the RQS. It may be due to the fact that the RQS only considers the image contents, i.e., the appearance of face, where the differences between GAN generated images and real images are not distinct enough.

B. Detection based on Inception Score

To assess the quality and variation of GAN generated images, many existing works adopted the *Inception score* [14] as a measurement. Inception score is computed by feeding a set of images into the Inception model [19]. For an input image $\mathbf{x}$, its softmax output of the Inception model is denoted as $p(y|\mathbf{x})$; for all the input images, the mean of their softmax output is denoted as $p^*(y)$. The Inception score is given by $\exp(\mathbb{E}_{\mathbf{x}}\text{KL}(p(y|\mathbf{x})||p^*(y)))$, where $\text{KL}(\cdot)$ is the KL divergence, and $\mathbb{E}_{\mathbf{x}}$ means the calculation of mathematical expectation for all $\mathbf{x}$. It was reported that the Inception score is correlated with human evaluation [14]. For a GAN model, if it can generate images with higher Inception score, then it will be considered to have the ability to produce images with better quality and variation.

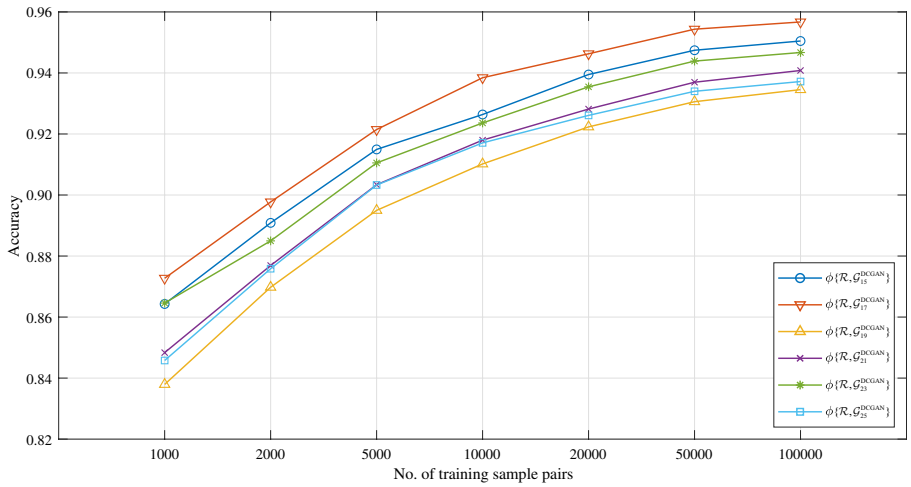
In order to test the effectiveness of Inception score for detecting GAN generated images, we fed the image sets $\mathcal{R}$, $\mathcal{G}_i^{\text{DCGAN}}$ ($i \in \{15, 17, 19, 21, 23, 25\}$), and $\mathcal{G}_j^{\text{WGAN}}$ ($j \in \{40, 42, 44, 46, 48, 50\}$) into the Inception model, and calculated the Inception score for each batch of images. The means of Inception score for different image sets are listed in Table III. From Table III, we observe that the Inception score for real images is significantly larger than those for the generated images, meaning that the generated images are different from real ones in Inception score. Besides, it is observed that there is trend of increase of the Inception score from earlier epochs to later epochs. For example, the Inception scores for the images generated by DCGAN monotonically increase from the 15th epoch to the 21st epoch, and the Inception scores for the images generated by WGAN at the 44th, 46th, 50th epochs are larger than those at the 40th, 42nd epochs. Such a phenomenon implies that the quality of generated images would be enahnced during the training process.

It is noted that Inception score should be calculated on a large enough number of samples to ensure its reliability. This shortcoming would limit the application of Inception score for detecting a single image, although a set of GAN generated images and a set of real images exhibit different Inception scores. Therefore, we need to seek for more effective detection method.

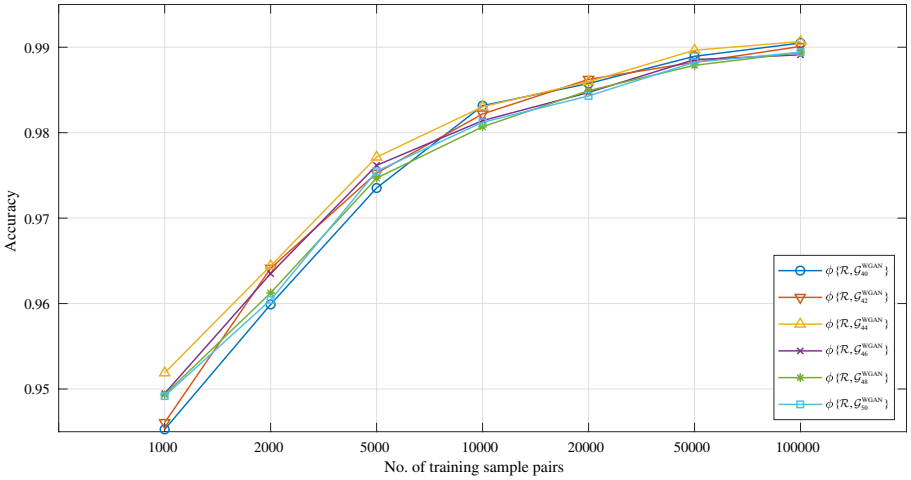
C. Detection based on VGG features

As the two non-intrusive approaches discussed above lack the ability to identify whether a specific image is generated or not, we further try to construct detectors based on discriminative features. At this point, it is critical to design an effective feature set to capture the disparities between GAN generated images and real images. In many image classification applications, it has shown that the convolutional outputs of CNN can serve as powerful features. In this paper, we adopt the outputs of the last convolutional layer in VGG-16 [15] as features for the detection of GAN generated images.

To extract features from the images, we fed each image to the VGG-16 network that was pre-trained on imagenet and then concatenated the convolutional outputs before the fully connected layers as a feature vector. As the size of images is 128×128, the obtained feature set is of 8192-D. With the obtained features, we trained FLD (fisher linear discriminant) ensemble classifiers [20] to classify the generated images and real images. In the training stage, we randomly selected different numbers of training sample pairs and obtained different classifiers. The trained classifiers are denoted as $\phi\{\mathcal{R}, \mathcal{G}_i^*\}$, where $\mathcal{G}_i^*$ is the generated image set used for training. In the testing stage, we evaluated $\phi\{\mathcal{R}, \mathcal{G}_i^*\}$ by computing the average accuracy on testing sets of real images and generated images corresponding to $\mathcal{G}_i^*$. The testing accuracies are shown in Fig. 2. It can be observed that the accuracies increase with the growing of training samples. When the training sample pairs are over 50000, the detection accuracies become stable, meaning that the training data is sufficient. For the images generated by DCGAN, the VGG features can achieve good performance with accuracy over 90% when the training sample pairs are larger than 10000, and the average accuracy is over



(a) DCGAN



(b) WGAN

Fig. 2. Testing accuracies with the VGG16 based features.

94% when the training sample pairs reach 100000. For the images generated by WGAN, the detection accuracies are larger than 94% in all the cases. Especially, when the training sample pairs reach 100000, the obtained accuracies are very close to 99%. These experimental results indicate that the features derived from CNN for image classification are very effective for detecting GAN generated images when there are sufficient training samples.

The above experiment was conducted in an ideal situation that the training and testing images were generated at the same epoch. In practice, however, the training and testing images are probably mismatched. To evaluate the performance for practical situations, we used the classifiers $\phi\{\mathcal{R}, \mathcal{G}_{21}^{\text{DCGAN}}\}$ and $\phi\{\mathcal{R}, \mathcal{G}_{46}^{\text{WGAN}}\}$ to test the images generated by different GAN architectures and from different epochs. The experimental results are shown in Table IV. From this table we observe that the detection performance is degraded when the training and testing data are mismatched. For the detector $\phi\{\mathcal{R}, \mathcal{G}_{21}^{\text{DCGAN}}\}$, the degradation of accuracy is about 10% in the mismatched cases. For the detector $\phi\{\mathcal{R}, \mathcal{G}_{46}^{\text{WGAN}}\}$, the performance degradation for images generated by WGAN but from different epochs is slight. However, its detection accuracy for images generated by DCGAN is dropped to just about 60%, which is quite poor. This experiment implies that it is challenging to detect generated images when there is a lack of information about exact sources the testing images.

TABLE IV
DETECTION ACCURACIES FOR THE CASES THAT THE TRAINING DATA AND TESTING DATA ARE MATCHED (SHOWN IN BOLD) AND MISMATCHED

Detector	$\phi\{\mathcal{R}, \mathcal{G}_{21}^{\text{DCGAN}}\}$					
Testing data	$\{\mathcal{R}, \mathcal{G}_{15}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{17}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{19}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{21}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{23}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{25}^{\text{DCGAN}}\}$
Avg. ACC	0.8098	0.8581	0.8073	**0.9401**	0.8300	0.8144
Detector	$\phi\{\mathcal{R}, \mathcal{G}_{21}^{\text{DCGAN}}\}$					
Testing data	$\{\mathcal{R}, \mathcal{G}_{40}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{42}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{44}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{46}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{48}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{50}^{\text{WGAN}}\}$
Avg. ACC	0.8779	0.8667	0.8711	0.8704	0.8777	0.0.8742
Detector	$\phi\{\mathcal{R}, \mathcal{G}_{46}^{\text{WGAN}}\}$					
Testing data	$\{\mathcal{R}, \mathcal{G}_{15}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{17}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{19}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{21}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{23}^{\text{DCGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{25}^{\text{DCGAN}}\}$
Avg. ACC	0.5945	0.6133	0.5716	0.5915	0.5817	0.5812
Detector	$\phi\{\mathcal{R}, \mathcal{G}_{46}^{\text{WGAN}}\}$					
Testing data	$\{\mathcal{R}, \mathcal{G}_{40}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{42}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{44}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{46}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{48}^{\text{WGAN}}\}$	$\{\mathcal{R}, \mathcal{G}_{50}^{\text{WGAN}}\}$
Avg. ACC	0.9871	0.9865	0.9881	**0.9891**	0.9865	0.9872

V. CONCLUSIONS

In this paper, we discuss how to detect the GAN generated images. We consider two kinds of approaches. The first kind of approach is intrusive, which employs the discriminators in GAN to detect the generated images. The second kind of approach is non-intrusive. Three non-intrusive approaches are evaluated in this paper, which are based on face quality assessment, Inception score, and VGG features, respectively. Although the current GAN based models can generate realistic images, the experimental results show that both the intrusive and non-intrusive approaches are able to detect GAN generated images. Among the non-intrusive approaches, the last one achieves the most satisfactory performance. However, there are still some challenges for detecting GAN generated images, especially considering the fact that the false alarm rate of intrusive approach is not so satisfactory and the performance degradation of VGG features based non-intrusive approach when training and testing data are mismatched.

In the future, we will further analyze the disparities between GAN generated images and real images, and study more effective and practical approaches to perform the detection. On the other hand, we will try to improve the GANs to produce more realistic images.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Conf. Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[2] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4681–4690.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2223–2232.

[5] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Conf. Neural Information Processing Systems (NIPS)*, 2017, pp. 700–708.

[6] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 107:1–107:14, 2017.

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2016.

[8] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2813–2821.

[9] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

[10] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Conf. Neural Information Processing Systems (NIPS)*, 2017, pp. 5769–5779.

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[13] J. Chen, Y. Deng, G. Bai, and G. Su, "Face image quality assessment based on learning to rank," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 90–94, Jan 2015.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Conf. Neural Information Processing Systems (NIPS)*, 2016, pp. 2234–2242.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 3730–3738.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int.Conf. Machine Learning (ICML)*, 2015, pp. 448–456.

[18] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. ACM SIGKDD Int. Conf. Knowledge discovery and data mining*, 2002, pp. 133–142.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[20] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, Apr. 2012.