Learning to Generate Realistic Scene Chinese Character Images by Multitask Coupled GAN

Qingxiang Lin[1], Lingyu Liang[1], Yaoxiong Huang[1], and Lianwen Jin[1,2(✉)]

[1] School of Electronic and Information Engineering,
South China University of Technology, Guangzhou, China
lhlqx2014@gmail.com, lianglysky@gmail.com, hwang.yaoxiong@gmail.com,
lianwen.jin@gmail.com
[2] SCUT-Zhuhai Institute of Modern Industrial Innovation,
South China University of Technology, Zhuhai, China

Abstract. Scene text recognition, is challenging due to the large appearance variances of the scene character. Recently, deep learning technique has shown its power for scene text recognition, but it requires enormous annotated data for training and it is time-consuming to manually obtain abundant data for all the categories of characters. This paper proposes a new architecture, called multitask coupled generative adversarial network (MtC-GAN), for scene Chinese character recognition (SCCR). The MtC-GAN consists of coupled GAN networks for scene character style transfer and classifier networks trained by the style-transferred data generated by the coupled GAN. To make the generated data be realistic enough for SCCR, we train the multitask networks using a new loss function that combines the constrains of encoders, generators and classifiers simultaneously. Experiments show that the proposed MtC-GAN framework is general and flexible to improve the accuracy for SCCR.

Keywords: Scene Chinese character recognition
Generative adversarial networks · Multitask training

1 Introduction

Scene text recognition (STR) has been drawing ever-increasing research interests in recent years given its potential for many applications, such as autonomous driving [1,2], license plate recognition [3,4] and industrial automation [5,6]. Although traditional optical character recognition has been extensively studied, naively adapting the technique to STR may fail to perform well, especially for scene Chinese character recognition (SCCR). The main challenge of SCCR lies in the large appearance variances of the scene character caused by style, font, resolution, illumination, projection transformation or partially occluded.

Recently, deep learning technique has been introduced into the field of STR [7–9]. The deep neural networks (DNN) consists of hierarchical nonlinear transformation, and is allowed to learn the feature and classifier with great invariant

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11258, pp. 41–51, 2018.
https://doi.org/10.1007/978-3-030-03338-5_4

and discriminate properties. The developed system with DNN structure obtains the state-of-the-art performance for SCCR. However, it requires enormous annotated data to train and fine-tune the DNN-based system. Although large-scale benchmark databases have been constructed for STR and SCCR [10], it is still time-consuming to obtain abundant labels, and the large categories of SCCR may also suffer from data imbalance. For instance, in the recently proposed CTW dataset [10], Chinese character samples of common categories can exceed the 17000 entries, whereas some rare categories contain only one sample. Therefore, it would be significant to generate scene Chinese character images for SCCR using DNN architecture.

The generation of scene Chinese character images can be divided into rule-based and learning-based methods. For the rule-based scheme, Campos et al. [11] generated English characters to train a character-level English scene text classifier; Jaderberg et al. [12] create a synthetic word data generator through physical rendering process to train a whole-word-based English scene text classifier; Gupta et al. [13] proposed a fast and scalable engine to generate synthetic images of text in clutter which further consider the local 3D scene geometry, and then train a text localisation network. The abovementioned methods which are limited by their rule-based nature seems to hardly simulate all the important variances in the real-world. For example, the work of [13] is limited by the segmentation and depth prediction of background images.

The learning-based method is mostly motivated by the GAN architecture [14], which can estimate the target distribution, and then generate similar images to the real ones. Although the previous X-GAN framework can have many advantages, it can't be ensured that each samples generated by GAN methods can preserve annotation information, and the naively synthetic data generated by GAN method may fail to improve the prediction performance due to these bad samples.

To tackle this problem, we propose a multitask coupled GAN framework for scene Chinese character recognition, which generates realistic scene Chinese character and improves the classification accuracy by the generated data simultaneously. The MtC-GAN consists of coupled GAN networks for scene character style transfer and classifier networks trained by the style-transferred data generated by the coupled GAN. To make the generated data be realistic enough for scene Chinese character recognition, we propose a new loss that combines the constrains of encoders, generators and classifiers simultaneously. Experiments show that the synthetic data by our method have great visual consistency to the realistic data. Furthermore, classifiers with different deep structures, like ResNet18 [15], ResNet34 [15] or VGG16 [16], can obtain apparent performance improvement, which indicate that the proposed multitask coupled GAN framework is general and flexible to improve the accuracy for SCCR.

The contributions of our work can be summarized as follows:

– A multitask coupled GAN learning framework for SCCR, which is general and flexible to generate realistic data and improve the accuracy of the classifier by generated data simultaneously without extra human annotation efforts;

– A new loss that combines the constrains of encoders, generators and classifiers to regularize the learning of the multitask coupled GAN.
– We qualitatively and quantitatively assess the classifier performance to demonstrate the effectiveness of the proposed method.

2 Related Works

Scene text image generation is a challenging task given the presence of complex background and font diversity. Many researchers have proposed the generation of realistic scene text images. Campos et al. [11] generated English character images to train a character-level English scene text classifier. Jaderberg et al. [12] create a synthetic word data generator through physical rendering process to train a whole-word-based English scene text classifier. Gupta et al. [13] proposed a fast and scalable engine to generate clutter-text synthetic images considering local 3D scene geometry, and then train a text localisation network. However, these methods are limited by their rule-based nature. For instance, the method in [13] is limited by the segmentation and depth prediction of background images. Unlike the abovementioned methods, we propose a learning-based method to generate realistic scene Chinese character images and further improve the recognition performance.

As one of the most considerable improvements on the research of deep generative models [17,18], GANs [14] are being intensively studied by the deep learning and computer vision communities alike. A GAN basically consists of generator and discriminator networks, where the former generates samples to increase the discriminator error rate, and the latter aims to distinguish real from synthetic images. This adversarial training allows the generator to estimate the target distribution and then generate similar images to the real ones. Mathematically, the standard GAN training aims to solve the following optimization problem:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

To extend the abilities of GANs, Mirza et al. [19] proposed a conditional GAN to direct data generation by conditioning both the generator and discriminator on additional information. This type of GAN has been successfully used in plenty of applications, such as image super-resolution [20,21], image style transfer [22–25], domain adaptation [26], etc.

Furthermore, conditional GANs are suitable for image-to-image translation, which has been applied for different purposes including the generation of maps from aerial photos and colorization of grayscale images. Conditional GAN is well suited for this task and many researchers have achieved great success based on it. Likewise, Isola et al. [22] proposed the pix2pix model to learn the mapping from input to output images using paired images. Zhu et al. proposed CycleGAN [23] based on a cycle consistency loss to break the limit of training with paired images. Liu et al. [25] proposed an unsupervised image-to-image translation (UNIT) network assuming a shared latent space. Azadi et al. [27] proposed

the multi-content GAN(MCGAN) for few-shot font style transfer. Shrivastava et al. [28] proposed a simulated and unsupervised SimGAN to enhance the realism of an image simulator while preserving annotation data and demonstrated a high performance with no labeled real data. Zhao et al. [29] proposed a dual-agent GAN(DA-GAN) to enhance the realism of a face simulator output by using unlabeled real-face images while preserving identity information. Our proposed multitask coupled GAN combines the advantages of the UNIT network [25] and DA-GAN [29] to improve the quality of synthetic images and consequent classifier performance.

3 Multitask Coupled GAN

3.1 Source Data

We first propose a synthetic character generator that retrieves simple Chinese character images through font rendering, affine transformation, and perspective transformation. We denote the synthetic data generated in this way as source data $\mathbf{x_s}$. By using diverse TrueType and OpenType font files obtained from the Internet, we generate plenty of simple Chinese character images with annotation information. In addition, we use real image dataset published by Yuan et al. [10] and denote it as $\mathbf{x_t}$. We aim to simultaneously reduce the difference between $\mathbf{x_s}$ and $\mathbf{x_t}$ and improve the performance of a scene Chinese character classifier.

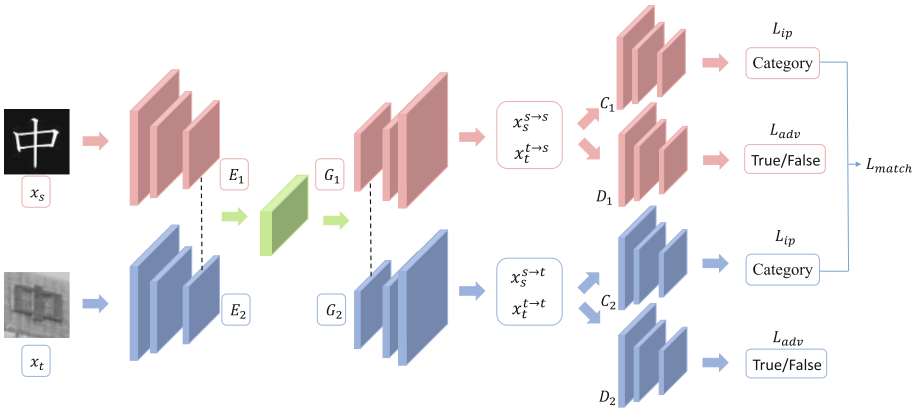

Fig. 1. Diagram of the proposed multitask coupled GAN architecture. E_1 and E_2 are two encoding functions that map images to latent codes. G_1 and G_2 are generation functions that map latent codes to images. D_1 and D_2 are adversarial discriminators for the respective domains. C_1 and C_2 are classifiers for the respective domains. L_{ip}, L_{adv} and L_{match} are the identity perception, adversarial, and matching losses, respectively. The dash lines denote weight sharing.

3.2 Coupled Generator

The same Chinese characters can present appearance variations in natural images arising from complex backgrounds and writing styles. Still, humans can easily recognize these characters, suggesting that the same characters written with different styles might share high-level semantic characteristics in the human brain. This semantic similarity can be represented by a map from characters with different styles into the same latent space, and an inverse map from a latent space into different domain images. Consequently, if the same characters with different styles are mapped into a latent space, we can generate corresponding images in two domains using autoencoders. To this end, we use concepts of coupled GAN [30] and UNIT network [25] to establish a shared latent-space assumption through a weight-sharing constraint. The architecture of the proposed MtC-GAN model is illustrated in Fig. 1 and relies on a UNIT network, where generator loss L_{unit} is formulated as:

$$\begin{aligned} L_{unit} = & L_{VAE_1}(E_1, G_1) + L_{GAN_1}(E_1, G_1, D_1) + L_{CC_1}(E_1, G_1, E_2, G_2) + \\ & L_{VAE_2}(E_2, G_2) + L_{GAN_2}(E_2, G_2, D_2) + L_{CC_2}(E_2, G_2, E_1, G_1) \end{aligned} \quad (2)$$

where L_{VAE} denotes the variational autoencoder loss, L_{CC} denotes the cycle-consistent loss [23], L_{GAN} denotes the standard adversarial loss [14]. and D, G, and E denote adversarial discriminators, generators and encoders, respectively. More details on the loss functions can be found in [25]. The loss constraint can only add realism to synthesized images in appearance, but hardly preserves annotation information well. However, to use the synthesized data for improving classification performance, the synthesized images should preserve annotation information. Therefore, we include identity perception loss L_{ip} that is a multi-class cross-entropy loss to preserve annotation information. Then, we update the generator parameters by minimizing the following loss:

$$L_G = L_{unit} + \lambda_1 L_{ip} \quad (3)$$

where hyperparameter λ_1 control the weights of the objective terms. This combined loss both enhances the realism of synthetic images and preserves annotation data.

3.3 Multitask Discriminator

The discriminator aims to distinguish real from synthesized images. Its loss is given by:

$$\begin{aligned} L_{adv} = & \log D_1(x_s) + \log(1 - D_1(G_1(E_2(x_t)))) + \\ & \log D_2(x_t) + \log(1 - D_2(G_2(E_1(x_s)))) \end{aligned} \quad (4)$$

In addition, we train a classifier to preserve label information of the generated data using identity perception loss L_{ip} defined as:

$$L_{ip} = \sum_{n} -Y_s \log D_{c_1}(x_s) + \sum_{n} -Y_t \log D_{c_1}(G_1(E_2(x_t))) + \sum_{n} -Y_t \log D_{c_2}(x_t) + \sum_{n} -Y_s \log D_{c_2}(G_2(E_1(x_s))) \tag{5}$$

where D_{c_1} and D_{c_2} are the probabilities of class n output by classifier C_1 and C_2, respectively. Y_s and Y_t are the labels of $\mathbf{x_s}$ and $\mathbf{x_t}$, respectively. The definitions above derive in a multitask training that preserves label information of the synthetic data. In addition, we can generate any amount of training data for training supervised models.

To further constrain classifiers C_1 and C_2, we define a matching loss, formulated as:

$$L_{match} = \sum_{i} |D_{c_1}(x_s) - D_{c_2}(G_2(E_1(x_s)))| + |D_{c_2}(x_t) - D_{c_1}(G_1(E_2(x_t)))| \tag{6}$$

Where i is the class index. This loss improves the classifier performance. Likewise, we define another constraint in the generator to improve the quality of the generated data by training the discriminator to minimize combined loss:

$$L_D = L_{adv} + \gamma_1 L_{ip} + \gamma_2 L_{match} \tag{7}$$

where hyperparameters γ_1 and γ_2 weigh the corresponding objective terms.

We optimize MtC-GAN by alternatively optimizing multitask discriminator and coupled generator for each training iteration until the whole network converge.

4 Experiments and Results

We evaluated the performance of the proposed MtC-GAN mainly on the CTW dataset [10]. Although the most commonly used metric for determining the quality of generative models is the inception score [31], it does not suit our objective of using the generated data to improve the classifier performance. Instead we use two complementary evaluation metrics. First, similar to [28], we deploy the 'Visual Turing Test' to evaluate the visual quality of the generated images. Second, we use generated data to train a classifier, and compare the performance among classifiers with different generation methods.

4.1 GAN Training

We used a recently released Chinese text detection and recognition dataset, the CTW dataset [10]. It is split into training, validation and testing dataset, where the validation dataset was used for evaluating all the experiments. Similar

to [10], we only consider recognition of the top 1000 most frequently observed character categories. In addition, we evaluated a simple classifier to determine the enhancement provided by the generated images. Specifically, the classifier that we used is the ResNet18 [15], whereas the architecture of generator and discriminator was the same as that of the UNIT network [25]. The encoders consisted of 3 convolutional layers as the front-end and 4 basic residual blocks [15] as the back-end. The generators consisted of 4 basic residual blocks as the front-end and 3 transposed convolutional layers as the back-end. The discriminators consist of 6 convolutional layers. Then, an Adam solver [32] was adopted for the MtC-GAN with learning rate of 0.0002, $\lambda_1 = 1$, $\gamma_1 = 1$,$\gamma_2 = 5$.

4.2 Generated Image Quality

In this section, we deployed the 'Visual Turing Test' [28] to quantitatively evaluate the visual quality of the generated images and designed a simple user study where subjects were asked to classify images as being either real or synthetic. Each subject observed a random selection of 40 real and 40 synthetic character images that were randomly presented, and was asked to label the character images as either real or synthetic. We used the classification accuracy for quantitative evaluation, whose outcomes are shown in Table 1. The classification accuracy among subjects was 57%, which is very close to a random selection, i.e., 50%. Consequently, we considered that the subjects were unable to distinguish between real and synthetic images.

Table 1. Results of the 'Visual Turing test' where subjects classified real and synthetic images. The average classification accuracy among subjects was **57%**, close to the **50%** of random selection.

	Selected as real	Selected as synthetic
Ground truth real	225	175
Ground truth synthetic	169	231

Figure 2 shows examples of characters generated using the proposed method that served to quantitatively evaluate its outcomes.

4.3 Classifier Performance

The goal of this study was to use generated data for improving the classifier performance, and thus the classification accuracy was our main concern. Table 2 lists the classification accuracy using different generation methods. We can see that, naively learning from synthetic data can undermine classification accuracy due to the difference between synthetic and real image distributions, whereas the proposed MtC-GAN generation method achieves the best performance among

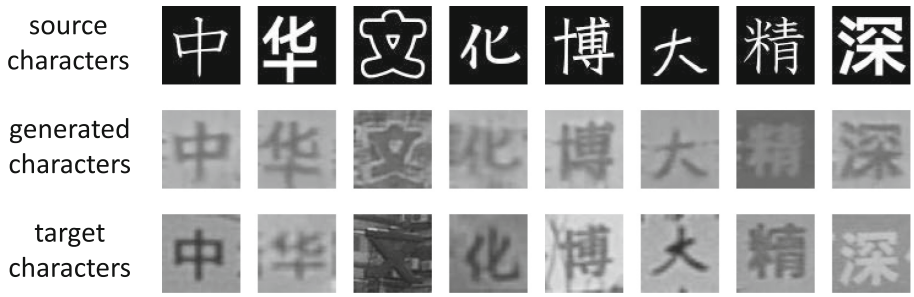

Fig. 2. The generated images using multitask coupled GAN. From top to bottom: source characters, generated characters, target characters.

Table 2. Classification accuracy of different generation methods

Generation method	Classification accuracy
Real data only	76.3%
Real data + source data(x_s)	75.5%
Real data + synthtext2014 [12]	78.5%
Rael data + synthtext2016 [13]	78.2%
Real data + SimGAN [28]	77.2%
Real data + CycleGAN [23]	77.8%
Real data + UNIT [25]	78.5%
Real data + proposed MtC-GAN	**80.7%**

Table 3. Classification accuracy of different classifiers with and without the generated images

Classifier	Real data	Real data+MtC-GAN
ResNet18 [15]	76.3%	**80.7%**
ResNet34 [15]	78.5%	**82.2%**
VGG16 [16]	81.3%	**83.5%**

the compared methods, suggesting that multitask training can improve the classifier performance.

To further verify the effectiveness of the proposed method, we use different classifiers, whose accuracies are listed in Table 3. Every classifiers using data generated from the proposed MtC-GAN exhibits the best performance. Furthermore, the ResNet18 with multitask training can have better performance than the ResNet34 [15] without multitask training. It shows that if we can generate images which are realistic enough, we can train a shallow network enjoying the comparable performance with a deep one.

5 Conclusions

We propose a multitask coupled GAN (MtC-GAN) for realistic annotation-preserving image synthesis. The generated scene Chinese character images improve the performance of character classifiers. Both qualitative and quantitative evaluations demonstrate the effectiveness of the proposed MtC-GAN method and its superior performance. The experimental results also suggest that if we can generate images which are realistic enough, we can train a shallow network enjoying the comparable performance with a deep one.

Acknowledgement. This research was supported in part by GD-NSF (No. 2017A030312006), the National Key Research and Development Program of China (No. 2016YFB1001405), the National Natural Science Foundation of China (No.: 61673182, 61771199, 61502176), GDSTP (No.: 2014A010103012, 2017A010101027), GZSTP (No. 201607010227) and Fundamental Research Funds for the Central Universities (No. 2017BQ058).

References

1. Bojarski, M., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
2. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3D object detection network for autonomous driving. In: IEEE CVPR, vol. 1, p. 3 (2017)
3. Björklund, T., Fiandrotti, A., Annarumma, M., Francini, G., Magli, E.: Automatic license plate recognition with convolutional neural networks trained on synthetic data. In: 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6. IEEE (2017)
4. Masood, S.Z., Shu, G., Dehghan, A., Ortiz, E.G.: License plate detection and recognition using deeply learned convolutional neural networks. arXiv preprint arXiv:1703.07330 (2017)
5. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Meghini, C., Vairo, C.: Deep learning for decentralized parking lot occupancy detection. Expert Syst. Appl. **72**, 327–334 (2017)
6. Song, X., Kanasugi, H., Shibasaki, R.: DeepTransport: prediction and simulation of human mobility and transportation mode at a citywide level. In: IJCAI, pp. 2618–2624 (2016)

7. Shi, B., Wang, X., Lyu, P., Yao, C., Bai, X.: Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4168–4176 (2016)
8. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2017)
9. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: towards accurate text recognition in natural images. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5086–5094. IEEE (2017)
10. Yuan, T.-L., Zhu, Z., Xu, K., Li, C.-J., Hu, S.M.: Chinese text in the wild. arXiv preprint arXiv:1803.00085 (2018)
11. De Campos, T.E., Babu, B.R., Varma, M.: Character recognition in natural images (2009)
12. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014)
13. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2315–2324 (2016)
14. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082 (2014)
19. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
20. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint (2016)
21. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 318–333. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_20
22. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arXiv preprint (2017)
23. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2017)
24. Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGan: unsupervised dual learning for image-to-image translation. arXiv preprint (2017)
25. Liu, M.-Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, pp. 700–708 (2017)
26. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2030–2096 (2016)

27. Azadi, S., Fisher, M., Kim, V., Wang, Z., Shechtman, E., Darrell, T.: Multi-content GAN for few-shot font style transfer. arXiv preprint arXiv:1712.00516 (2017)
28. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 3, p. 6 (2017)
29. Zhao, J., et al.: Dual-agent GANs for photorealistic and identity preserving profile face synthesis. In: Advances in Neural Information Processing Systems, pp. 65–75 (2017)
30. Liu, M.-Y., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems, pp. 469–477 (2016)
31. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
32. Kinga, D., Adam, J.B.: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)