Image Synthesis with Aesthetics-Aware Generative Adversarial Network

Rongjie Zhang, Xueliang Liu, Yanrong Guo, and Shijie Hao(✉)

Hefei University of Technology, Hefei 230009, Anhui, China
hfut.hsj@gmail.com

Abstract. With the advance of Generative Adversarial Networks (GANs), image generation has achieved rapid development. Nevertheless, the synthetic images produced by the existing GANs are still not visually plausible in terms of semantics and aesthetics. To address this issue, we propose a novel GAN model that is both aware of visual aesthetics and content semantics. Specifically, we add two types of loss functions. The first one is the aesthetics loss function, which tries to maximize the visual aesthetics of an image. The second one is the visual content loss function, which minimizes the similarity between the generated images and real images in terms of high-level visual contents. In experiments, we validate our method on two standard benchmark datasets. Qualitative and quantitative results demonstrate the effectiveness of the two loss functions.

Keywords: Image synthesis · Generative Adversarial Network
Image aesthetics

1 Introduction

The rapid development of Generative Adversarial Network family sheds light on the task of natural image generation. As the basic idea of GAN [7], the generator tries to produce images as real as possible to confuse the discriminator. Various GAN-based models [1,2,8,8,24,28] have been proposed to optimize the instability problems in generating images from different aspects. They have made solid progress in synthesizing natural images by using the standard datasets with legible backgrounds/foregrounds [33], e.g. MNIST [20], CIFAR-10 [18], CUB-200 [36] and so on. In many real-world applications, generating images with good visual aesthetics is highly desirable. Most of the existing GAN models are limited to achieve this goal, as they do not consider the image aesthetics in the learning process.

To address the above issue, in this paper, we propose a novel adversarial network namely AestheticGAN, and synthesize images with better visual aesthetics and plausible visual contents. Our consideration is two-fold. First, people always prefer to images with pleasant appearances, such as vivid color and appropriate

R. Hong et al. (Eds.): PCM 2018, LNCS 11165, pp. 169–179, 2018.
https://doi.org/10.1007/978-3-030-00767-6_16

composition. Therefore, the image generator is expected to be trained with aesthetics awareness. Second, apart from visually appealing, the generated images should also have reasonable visual contents. For example, based on our method, the image scene is quickly recognizable, and the content details are real. So the image generator is also expected to be aware of image semantics. To this end, we design and add two types of loss functions for the DCGAN architecture [29]. The first one is the aesthetics loss, which uses a quantitative score to evaluate the visual aesthetics of an image. The second one is the semantic loss, which measures the high-level semantic similarity between generated and real images [13,21].

The main contributions of this work are listed as follows:

- We attempt to create images with visually appealing images based on adversarial learning. Two types of loss functions are designed and added into the state-of-the-art GAN architecture.
- Extensive experiments are conducted on the AVA and cifar10 datasets. Comparisons in terms of visual appearance, quantitative scores, and user studies all demonstrate the effectiveness of our method.

The remain parts of this paper are organized as follows. We briefly review the related work in Sect. 2, and describe the proposed method in Sect. 3. In Sect. 4, we evaluate our method with qualitative and quantitative experiments. Section 5 finally concludes the paper.

2 Related Works

Since our research is closely related with the fields of GANs and image aesthetics, we briefly introduce their related research in this section.

GAN is a generation model inspired by two-person zero-sum game in Game Theory. Based on the seminal research by Goodfellow et al. [7], many GAN-based variants [1,2,8,24,28] have been proposed, which focus on the model structure extension, in-depth theoretical analysis, and efficient optimization techniques, as well as their extensive applications. For example, in order to solve the problem of disappearance of training gradient, Arjovsky et al. [1] proposes Wasserstein-GAN (W-GAN) and then improves it by adding the gradient penalty [8]. In order to limit the modeling ability of the model, Qi [28] proposes Loss-sensitiveGAN (LS-GAN), which limits the loss function obtained by minimizing the objective function to satisfy the Lipschitz continuity function class, and the authors also give the results of quantitative analysis of gradient disappearance. Further, ConditionalGAN (CGAN) [24] adds additional information(y) to the G and D, where y can be labels or other auxiliary information. InfoGAN [2] is another important extension of GAN, which can obtain the mutual information between hidden layer variables of the input and the specific semantics. Odena et al. [27] proposes that Auxiliary Classifier GAN (AC-GAN) can achieve multiple classification problems, and its discriminator outputs the corresponding tag probability.

Despite of the rapid development of GANs, there are few works that specifically designed for the task of aesthetic image generation.

The computational aesthetics has attracted attentions in recent years [6,14]. The purpose of the research on computational aesthetics is to endow machine with the ability to perceive the attractiveness of an image qualitatively or quantitatively. The extraction of aesthetics-aware features plays a key role in this direction before the deep learning era. Previous research efforts [4,15,23,26,35] have shown some success in extracting aesthetic features. For 3D objects, [10] proposes to employ multi-scale topic models to fit the relationship of features from the multiple views of objects. However, most of them are handcrafted and task-specific. With the continuous development of deep learning, extracting the deep features of aesthetics images becomes the best way to solve the above problems. A lot of CNN-based models such as [22,25] have been proposed to improve the results. The applications are mainly targeted on the task of image aesthetic evaluation [17,22,34]. What's more, Hong et al. [12] propose a multi-view regularized topic model to discover Flickr users's aesthetic tendency and then construct a graph to group users into different aesthetic circles. Based on it, a probabilistic model is used to enhance the aesthetic attractiveness of photos from corresponding circles [11]. Although existing GAN models have achieved great success, they are still limited in producing "beautiful and real" images. Based on adversarial learning, Deng et al. [37] enhance image aesthetics in terms of scene composition and color distribution. This work is different from the theme of our research, as the enhancement model of [37] tries to optimize the parameters of cropping and re-coloring for an existing natural image. For our method, we directly synthesize an image without any prior information on the input side, e.g. a meaningless noise image.

3 Proposed Method

We formulate the problem of automatic aesthetic image generation as an adversarial learning model. We first introduce the overall architecture of our proposed framework shown in Fig. 1. Then we present the details of the newly-added loss functions.

3.1 Overall Framework

Basically, GAN is a pair of neural networks (G;D): the generator G and the discriminator D. G maps a vector z from a noise space N^z with a known distribution p_z into an image space N^x. The goal of G is to generate p_g (the distribution of the samples $G(z)$) to deceive the network D. And goal of D is to try to distinguish p_g (the distribution of a generated image) from p_{data} (the distribution of a real image). These two networks are iteratively optimized against each other in a minimax game (hence namely "adversarial") until the convergence. In this context, the GAN model is typically formulated as a minimax optimization of

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

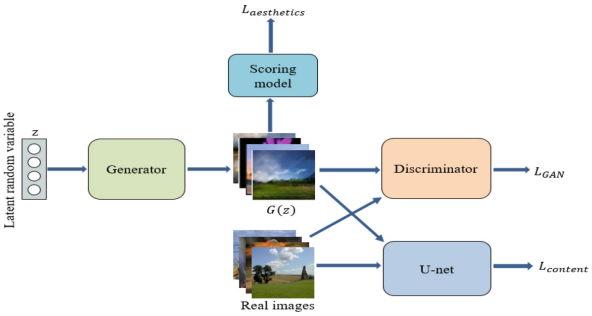


Fig. 1. The overall architecture of the proposed system.

Specifically, as for the structure of G and D, we choose fully convolutional networks as in DCGAN [29]. As shown in Fig. 2, there are a series of fractionally-stride convolutions in G and a series of convolution layers in D.

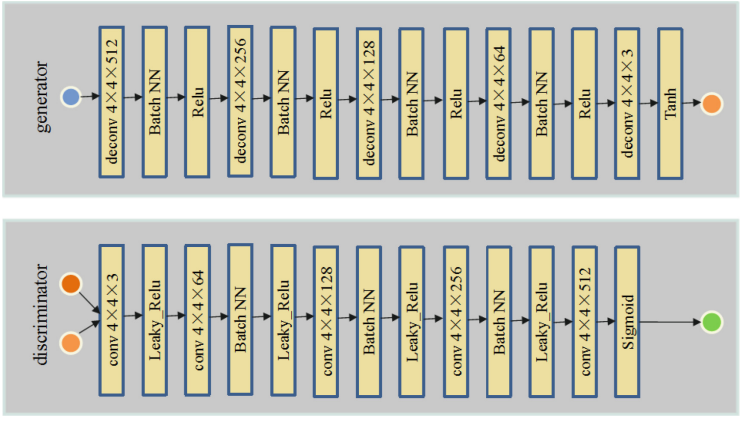


Fig. 2. The network of the G and D.

We can see that the above target function only seeks for the consistency between p_{data} and p_g in a broadly statistical sense. It has no explicit control over the visual appealingness and the content realness. So we extend the total loss function with two additional losses:

$$L_{total} = \alpha_1 L_{GAN} + \alpha_2 L_{aesthetics} + \alpha_3 L_{content} \tag{2}$$

In the formulation, L_{GAN} is the original GAN loss, $L_{aesthetics}$ is the aesthetics loss, and $L_{content}$ is the content loss. α_1, α_2, and α_3 denote their weights. In the following, we introduce the details of the two added losses.

3.2 Loss Function

Aesthetics-Aware Loss. In order to generate a visually appealing image, we propose to apply the aesthetics scoring model [17] to boost the image aesthetics, i.e. maximizing the obtained score, or minimizing $(1 - score)$. The key point is to learn a deep convolutional neural network that is able to accurately rank and rate visual aesthetics. In the network, the scoring ability is subtly encoded in its network architecture in the following aspects. First, the Alexnet [19] is fine-tuned based on a regression loss that predicts continuous numerical value as aesthetic ratings. Second, a Siamese network [3] is used by taking image pairs as inputs, which ensures images with different aesthetic levels have different ranks. The whole network is trained with a joint Euclidean and ranking loss. Moreover, they add attribute and content category classification layers and make the model be aware of fine-grained visual attributes. As demonstrated in [17], the overall aesthetic evaluation model is able to provide aesthetic scores which are well consistent with human rating. Therefore, we use the obtained scores as the aesthetic-aware loss:

$$L_{aesthetics} = \| (1 - S(\tilde{x})) \| \tag{3}$$

where $S(\tilde{x})$ is the aesthetic score of the generated image $\tilde{x}$.

Content-Aware Loss. Our synthesized images are also expected to have meaningful visual semantics. So we design the content-aware loss. In many image processing tasks [13,21], the content loss is considered. It is usually based on the activation maps produced by the ReLU layers of the pre-trained VGG network. Different from measuring pixel-wise distance between images, this loss emphasizes similar feature representation in terms of high-level content and perceptual quality. Since we aim to generate images with both good aesthetics and reasonable details, we need a network that is more suitable to our task. So we replace VGG with a more advanced U-net network [30], as its structure is able to preserve more image details by combining the concept features ("what it is") and the locality features ("where it is"). We denote $\psi_i()$ as the feature map extracted after the $i-th$ convolutional layer of the U-net. Then our content loss is defined as:

$$L_{content} = \frac{1}{C_i H_i W_i} \| \psi_i(\tilde{x}) - \psi_i(x) \| \tag{4}$$

where C_i, H_i, and W_i are the number, height and width of the feature maps, xs are real images and $\tilde{x}$s are generated ones.

3.3 Training Details

In training the proposed GAN model, input images are resized to 96×96 and then randomly cropped to 64×64, which reduces the potential over-fitting problem. The horizontal flipping of cropped images is also applied for random data augmentation. We use the ADAM technique [16] for optimization. As for the

learning rates lr_G and lr_D, we set them as 0.002 for both the generator network and the discriminator network. β_1 and β_2 are set as 0.5 and 0.999. We trained the proposed model in the experiments for 10000 epochs with minibatch size of 256. In implementation, we found out that reducing the learning rate during the training process helps to improve the image quality. Therefore, the learning rates are reduced by a factor of 2 in every 1000 epoch. We empirically set $\alpha_1 = 1, \alpha_2 = 0.15, \alpha_3 = 0.1$ in our experiments.

4 Experiments

In this section, we evaluate the proposed AestheticGAN on public benchmark datasets. Apart from the direct visual comparison, we also use quantitative measures and user study to validate its effectiveness.

4.1 Datasets for Training

The Aesthetic Visual Analysis (AVA) dataset is by far the largest benchmark for image aesthetic assessment. Each of the 255,530 images is labeled with aesthetic scores ranging from 1 to 10. In this study we select a subset of them, i.e., 25,000 images, based on the semantic tags provided in the AVA data for analysis. What's more, to further illustrate the applicability of our method, we also compare our model and its competitors on the cifar10 dataset.

4.2 Visual and Quantitative Comparison

We conduct visual comparison between the results of DCGAN and our model in Figs. 3 and 4. First, from Fig. 3, both DCGAN and our model generate images with good appearances at the first glance. However, the DCGAN results have less appropriate image composition, and less realistic image contents. In contrary, we can easily recognize the scene category and image contents of our results. Second, similar trends can be observed from Fig. 4, although they are not as clear as in Fig. 3. Furthermore, by comparing all the resultant images of Figs. 3 and 4, we can see in general that the model trained on AVA has superior performance than the one trained on cifar10 in terms of visual aesthetics, which indicates the data-driven property of GAN-based models.

We adopt the four different metrics for quantitative assessment. The first two are inception score [31] and Freéchet inception distance (FID) [9] that are commonly used in evaluating the performance of GAN-based image synthesis. Since our goal is to make the model aesthetics-aware, we also use two state-of-the-art evaluation models namely NIMA [32] and ACQUINE [5]. Among them, the NIMA estimates aesthetic qualities in aspects of photographing skills and visual appealingness. ACQUINE achieves more than 80% consistency with the human rating. Of note, larger values of inception score/NIMA/ACQUINE, and smaller FID values denote better quality, respectively. From Tables 1 and 2, we can see that our DCGAN+aesthetic+content achieves much better performances than

the baseline DCGAN model. Additionally, from Table 2, the aesthetic performance of AVA dataset is consistently better than that of cifar10 dataset, which echoes the above visual results.

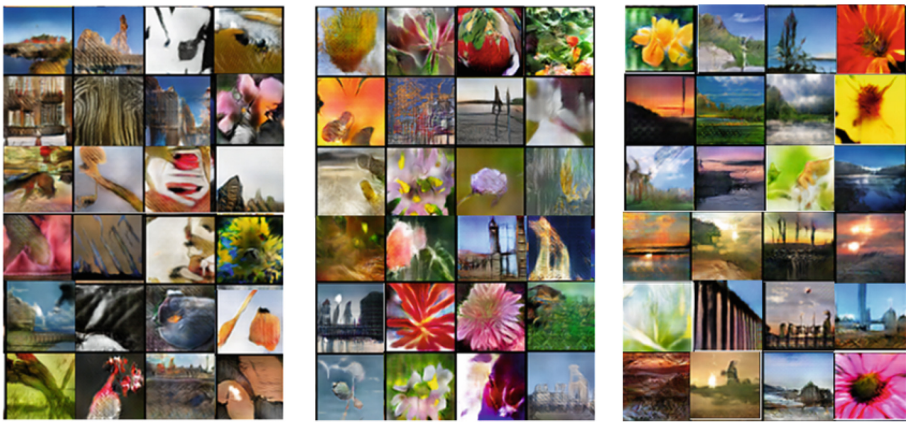
We also perform an ablation study. Apart from the baseline DCGAN and our DCGAN+aesthetic+content, we build an intermediate version DCGAN+aesthetic. Figure 5(b) has better lightness, vivid color and composition than Fig. 5(a). Furthermore, the contents in Fig. 5(c) are more realistic than those in Fig. 5(b). The results in Tables 1 and 2 are also consistent with the above observations. This experiment empirically validate the two losses, respectively.

(a) DCGAN (b) our method

Fig. 3. Comparison of the experiments on the AVA between DCGAN (left) and our method (right)

(a) DCGAN (b) our method

Fig. 4. Comparison of the experiments on the CIFAR10 between DCGAN (left) and our method (right)



(a) DCGAN (b) DCGAN+aesthetic (c) our model

Fig. 5. Result images for 3 different loss functions (Color figure online)

Table 1. Inception scores and FIDs for different methods on CIFAR10 and AVA datasets

Method	Inception score		FID	
	CIFAR10	AVA	CIFAR10	AVA
Real data	$11.24 \pm .12$	$14.37 \pm .68$	$7.82 \pm .11$	$8.15 \pm .09$
DCGAN	$6.64 \pm .14$	$7.45 \pm .29$	$37.71 \pm .24$	$69.81 \pm .15$
DCGAN+aesthetic	$6.92 \pm .17$	$7.69 \pm .26$	$36.64 \pm .32$	$64.93 \pm .23$
DCGAN+aesthetic+content	$7.13 \pm .12$	$8.05 \pm .22$	$34.33 \pm .27$	$62.54 \pm .18$

Table 2. The aesthetic scores of NIMA and ACQUINE for different methods on CIFAR10 and AVA datasets

Method	NIMA		ACQUINE	
	CIFAR10	AVA	CIFAR10	AVA
Real data	$6.54 \pm .25$	$7.98 \pm .69$	8.58 ± 1.29	$9.23 \pm .75$
DCGAN	$4.59 \pm .20$	$5.56 \pm .23$	$5.29 \pm .89$	$6.48 \pm .44$
DCGAN+aesthetic	$4.85 \pm .20$	$5.74 \pm .23$	$5.76 \pm .76$	$6.86 \pm .42$
DCGAN+aesthetic+content	$5.02 \pm .19$	$5.96 \pm .24$	$6.19 \pm .87$	$7.15 \pm .53$

4.3 User Study

We also conduct an experiment of user study. We built a ranking system and distributed it to a total of 30 participants. All participants were shown three sets of 330 images, where each image set were generated by three different loss configurations. We asked all participants to rank the images in range of 1–5, where 1 means the lowest aesthetic quality and 5 is the highest one. In order

to avoid the random and systematic errors, the images generated by different loss configurations are listed randomly. Also, we randomly repeatedly provide some images, and ignore the scores when a participant ranked differently on the repeated images. The statistics are shown in Fig. 6, which again demonstrates the effectiveness of the added losses.

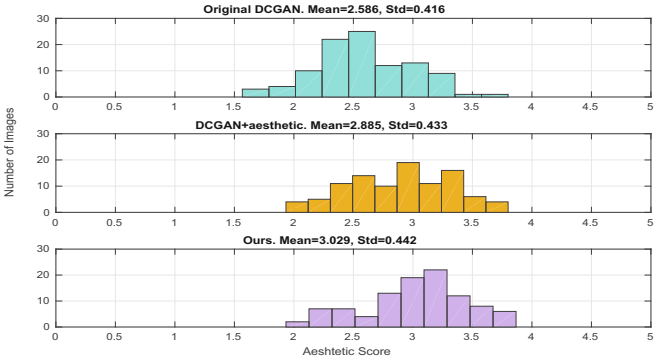


Fig. 6. User study on the AVA dataset

5 Conclusion

In the paper, we proposes a novel AestheticGAN to synthesize more challenging and complex aesthetic images. We enrich the loss function by designing two types of loss functions to train G. The aesthetics-aware loss helps to enhance aesthetic quality of the generated images, while the content-aware loss enforces them to be semantically meaningful. Various experimental results validate the effectiveness of our model. Of note, from Tables 1 and 2, we can see that the overall quality of GAN-generated images is still far from real-world natural images. We plan to narrow this gap by considering fine-grained aesthetic attributes as our future research.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants 61632007, 61502139, 61772171, and 61702156, in part by Natural Science Foundation of Anhui Province under grants 1608085MF128 and 1808085QF188 and in part by Anhui Higher Education Natural Science Research Key Project under grants KJ2018A0545.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv preprint arXiv:1701.07875 (2017)
2. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In: NIPS (2016)

3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR. IEEE (2005)
4. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006). https://doi.org/10.1007/11744078_23
5. Datta, R., Wang, J.Z.: ACQUINE: aesthetic quality inference engine-real-time automatic rating of photo aesthetics. In: ICMR. ACM (2010)
6. Deng, Y., Loy, C.C., Tang, X.: Image aesthetic assessment: an experimental survey. IEEE Signal Process. Mag. **34**(4), 80–106 (2017)
7. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)
8. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein GANs. In: NIPS (2017)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a nash equilibrium. arXiv preprint arXiv:1706.08500 (2017)
10. Hong, R., Hu, Z., Wang, R., Wang, M., Tao, D.: Multi-view object retrieval via multi-scale topic models. IEEE Trans. Image Process. **25**, 5814 (2016)
11. Hong, R., Zhang, L., Tao, D.: Unified photo enhancement by discovering aesthetic communities from flickr. IEEE Trans. Image Process. **25**, 1124 (2016)
12. Hong, R., Zhang, L., Zhang, C., Zimmermann, R.: Flickr circles: aesthetic tendency discovery by multi-view regularized topic modeling. IEEE Trans. Multimed. **18**, 1555 (2016)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
14. Joshi, D., et al.: Aesthetics and emotions in images. IEEE Signal Process. Mag. **28**, 94 (2011)
15. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: CVPR. IEEE (2006)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 662–679. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_40
18. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. University of Toronto (2009)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
20. LeCun, Y.: The MNIST database of handwritten digits (1998). http://yann.lecun.com/exdb/mnist/
21. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint (2016)
22. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rapid: rating pictorial aesthetics using deep learning. In: MM. ACM (2014)
23. Luo, Y., Tang, X.: Photo and video quality evaluation: focusing on the subject. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 386–399. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88690-7_29

24. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
25. Murray, N., Marchesotti, L., Perronnin, F.: AVA: a large-scale database for aesthetic visual analysis. In: CVPR. IEEE (2012)
26. Nishiyama, M., Okabe, T., Sato, I., Sato, Y.: Aesthetic quality classification of photographs based on color harmony. In: CVPR. IEEE (2011)
27. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. arXiv preprint arXiv:1610.09585 (2016)
28. Qi, G.J.: Loss-sensitive generative adversarial networks on lipschitz densities. arXiv preprint arXiv:1701.06264 (2017)
29. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
31. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NIPS (2016)
32. Talebi, H., Milanfar, P.: NIMA: neural image assessment. IEEE Trans. Image Process. **27**, 3998 (2018)
33. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: ArtGAN: artwork synthesis with conditional categorial GANs. arXiv preprint arXiv:1702.03410 (2017)
34. Tian, X., Dong, Z., Yang, K., Mei, T.: Query-dependent aesthetic model with deep learning for photo quality assessment. IEEE Trans. Multimed. **17**, 2035 (2015)
35. Tong, H., Li, M., Zhang, H.-J., He, J., Zhang, C.: Classification of digital photos taken by photographers or home users. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, pp. 198–205. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30541-5_25
36. Welinder, P., et al.: Caltech-UCSD birds 200. California Institute of Technology (2010)
37. Yubin Deng, C.C.L., Tang, X.: Aesthetic-driven image enhancement by adversarial learning. arXiv preprint arXiv: 1707.05251 (2017)