High-Quality Synthetic Character Image Extraction via Distortion Recognition

Tomoya Sawada(✉), Marie Katsurai, and Masashi Okubo

Doshisha University, 1-3 Tatara Miyakodani, Kyotanabe 610-0394, Kyoto, Japan
{sawada,katsurai}@mm.doshisha.ac.jp, mokubo@mail.doshisha.ac.jp

Abstract. Digital avatars have become indispensable in the digital age. In Japan, virtual characters using illustration-style avatars have gained popularity and is generating large economic impact. However, the creation of such avatars is a time-consuming and costly process that requires a great deal of expertise. To support avatar creation, research of automatic generation of character design and textures for 3D models have emerged. However, deep learning-based generative models sometimes synthesize corrupted outputs. Methods to detect collapsed outputs from a generative model has not been explored, and users of the generator needs to manually exclude such outputs. In this paper, we propose a method to extract high-quality images from a set of synthetic illustrations, generated by a deep learning model, based on the degree of distortion of the images. As it is difficult to prepare real-world distorted images to train a distortion recognition model, we propose a simple procedure to create pseudo-distorted images. Experimental results showed superior results of the proposed method in distinguishing between human-drawn images and generated images, compared to baseline methods. Furthermore, we sorted the generated images using the confidence level of the trained distortion detection model, and qualitatively confirmed that the proposed method produces results closer to human perception.

Keywords: High-quality character image extraction · distortion recognition · deep learning

1 Introduction

Digital avatars play an essential role in games, films, and in recent years, Metaverse. In Japan, virtual YouTubers (VTubers) are one of the contents that is generated a large economic effect through the use of avatars. In addition to their activities on the video-sharing website YouTube, VTubers have recently appeared on television and been used in corporate advertisements. The appeal of using avatars for such activities is that, avatars can be not only abstractions of real-world figures, but can also be different from oneself or even illustrated characters that do not exist in the real world, which can contribute to the acquisition of new identities in the electronic world. In addition, avatars are being incorporated into many interfaces used for computer-mediated communication, and are

M. Kurosu and A. Hashizume (Eds.): HCII 2023, LNCS 14013, pp. 522–535, 2023.
https://doi.org/10.1007/978-3-031-35602-5_37

gaining ground as a communication tool in virtual space. However, the creation of avatars is not an easy task, as each step of the process requires specialized knowledge such as in sculpturing the shape of the character, drawing textures, and capturing the motion to move the avatar. Therefore, a method to easily synthesize digital avatars by using a deep learning-based generative models has been studied in recent years [13,36].

Generative models may synthesize data that are corrupted which makes the contents unrecognizable. Methods to detect these collapsed generated data are not well researched, and users need to manually exclude the collapsed data from the produced results. There is a research field called image quality assessment (IQA), which evaluates the quality of images, but most of these studies are conducted on natural images, and the behavior of IQA methods on synthetic data has not been examined in detail. In addition, it is pointed out that the results of IQA methods for illustration images differ from those of human perception, and IQA for illustration images is still a developing research field [39]. Therefore, it is difficult to use IQA to detect collapsed synthetic images, especially on unnatural images such as illustrations. A method to extract only high-quality images from a synthetic image set generated by a deep learning model would be very beneficial.

Corruption of synthesized data include creation of unnecessary objects, loss of necessary objects, unnatural textures, and distortions. Among them, we focused on distortion because it is a content-independent collapse. There are studies that focus on local distortion of synthetic images in the field of fake face detection. For example, Guo et al. discovered that human face images synthesized by deep generative models have irregular pupil shapes and proposed a method to automatically detect synthesized images by segmenting the pupil [10]. Detecting the distortions caused by the image generation model will contribute to evaluate the quality of the images according to the degree of the distortions.

We propose a novel approach for extracting high-quality character images from a synthetic image set to support deep learning-based creation of digital avatars. Specifically, we train a deep learning model to detect distorted images. Since it is difficult to collect distorted images in the real world, pseudo-distorted images are synthesized by applying transformations to undistorted images and used as training data. To extract high-quality images from synthetic images provided by and external generative model, we rank images based on probability value, which is the confidence when the distortion recognition model predicted inputs as distorted images. The proposed method was evaluated quantitatively and qualitatively. In the quantitative evaluation, we placed synthetic images in a set of human-drawn images and evaluated whether the synthetic image could be extracted from this set of images. In the qualitative evaluation, we sorted the generated images according to the confidence level of the model and observed the results. The proposed method showed superior results compared to baseline methods.

The main contributions of this paper are as follows:

- We propose a simple but effective method for extracting high-quality images from a set of synthetic images using a deep learning based distortion detection model.
- We apply simple transformation to real images to produce pseudo-distorted images to train a distortion recognition model. The proposed method is considered to be robust to the content of the synthetic data.

2 Related Works

2.1 Generative Models

Generative adversarial networks (GAN) was first proposed by Goodfellow et al. [8] as a generative model trained with adversarial supervision. GANs are trained using two deep learning models, a generator that synthesizes images and a discriminator that discriminates whether the input data is real or fake images. The generator receives a latent vector sampled from a probability distribution as the input and attempts to generate realistic images to fool the discriminator. By having the generators and discriminators compete with each other, the generator will eventually produce data of such high quality that they can be mistaken as real. Various studies have been conducted on GAN since its appearance, such as alternative or addition loss functions [1,9,22,23], methods to stabilize training [14,38,40], and also model architectures [2,17]. Karras et al. proposed a generator architecture based on neural style transfer literature, StyleGAN [15], which can automatically learn to separate high-level attributes and stochastic variations from an intermediate latent space. By analyzing generated images of StyleGAN, Karras et al. [16] proposed StyleGAN2, which improved the quality of the generated images by updating the generator and discriminator architectural terms. In most cases, systems designed to support the creation of illustrated characters use methods that have been proven effective on natural images with illustration datasets. In this study, StyleGAN2 is trained on whole body images of illustrated characters and used as an image generator in the experiments.

Diffusion models are rapidly developing in various research fields of deep learning, and generative diffusion models are the most actively studied [12,28]. In our experiments, we have evaluated the proposed method used synthesized images from a GAN model, but we believe our method can be used on images generated by other generative models such as diffusion models.

2.2 Assessing Quality of Synthetic Images

Metrics such as Inception score [30] and Fréchet Inception distance [11] are most commonly used to evaluate image generative models. Both metrics are based on a model called Inception v3 [32], which is trained on a large dataset of natural images called ImageNet [4]. These metrics evaluate the diversity of images produced by the generator and the degree of similarity between the synthetic images and the real images. However, these evaluation metrics cannot evaluate

the visual quality of individual images. Blind IQA (BIQA) is a research field that aims to evaluate the quality of images on a piece-by-piece basis. BIQA is a fundamental metric in image processing operations such as image compression and transfer. Some methods are effective for natural images [37] or unnatural images such as computer graphics [25]. However, for illustrations, it has been reported that recent BIQA methods output results that differ from human perception [3]. In addition, since the validation of IQA methods when applied to synthetic data is still insufficient, the application of conventional IQA methods to our task will result in unintended behavior, such as judging low quality images as high quality and vice versa.

3 Distorted Image Recognition

This section describes how we detected high-quality illustrations from synthetic images, which were automatically produced by a GAN model. Our method is based on the assumption that the confidence of a deep learning model trained to detect distortions on images reflects the quality of the input images. First, we trained a distortion recognition model, which receives images and predict whether these input images were distorted or not. Though, it was difficult to collect real-world distorted illustrations as positive examples to train the distortion recognition model. Therefore, we applied a random transformation to human-drawn illustrations and created pseudo-distorted images to simulate corruptions on synthetic images produced by a generative model. Then, we input synthesized images to the trained model and sorted these images by the confidence of the predictions. We will explain the details in the following sections.

3.1 Pseudo Distorted Image Synthesis

This subsection describes the procedure to create pseudo-distorted images. The precedure to preprocess the input images is presented in Fig. 1. First, we sampled an image from the training set and randomly selected whether to apply transformation to this image. When the transformation was applied, the image was given a label representing the distorted class. The image that was not transformed was labeled as the non-distorted class. After the labeling, a transformation was uniformally selected from four spatial transformations. Sample images of the four transformations used in our experiments are presented on Fig. 2. The details of the transformations are described bellow:

Grid distortion divides an image into multiple grids and stretches or shrinks each grid vertically or horizontally. There are two parameters: the number of vertical and horizontal divisions, and the "limit" parameter that controls the maximum degree of expansion and contraction.

Elastic distortion [31] is a transform that moves image pixels within a local area using a displacement field. The distance that a pixel can move and the smoothness of the displacement field can be controlled by the parameters "alpha" and "sigma", respectively.

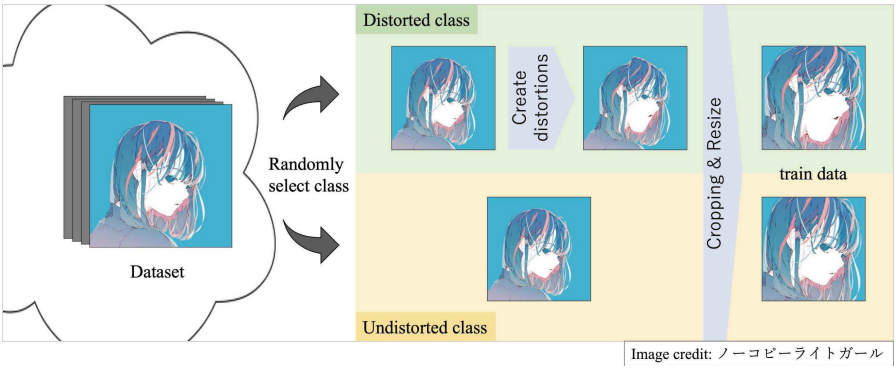

Fig. 1. Procedure of preprocessing images from the dataset as inputs to distortion detection model. Images labeled as distorted class is transformed using a spatial transformation method.

Table 1. Parameters used when applying the transform. When the "value" column is a range, a single value is uniformally sampled between the range.

transform	parameter	value
Grid distortion	number of grid divisions	15
	limit	0.6
Elastic distortion	alpha	200
	sigma	10
Piece-wise affine	number of rows and columns	$[3, 6]$
	scale	$[0.03, 0.05]$
Thin-plate-spline	initial point position	$[0.3, 0.7]$
	distance from initial point to the destination	$[-0.25, 0.25]$

Piece-wise affine transformation can create local distortions by randomly moving the neighbourhood of regularly placed grid of points via affine transformations. It is possible to control the number of rows and columns and also how much each point can be moved via "scale" parameter.

Thin-plate spline transformation is a method to warp an image by moving points on an image to another point and linearly interpolate the pixels between the points. Although the points and destinations to be moved are originally set manually, in this study they are uniformally sampled from a certain range.

The parameters used in the selected transformation were randomly sampled from a certain range for each training image. Table 1 are the parameters of each transform and the values we used in our experiments. Then, the images were randomly flipped horizontally. Finally, all images were resized and cropped to 224×224 pixels and the pixel values were normalized to $[-1, 1]$.

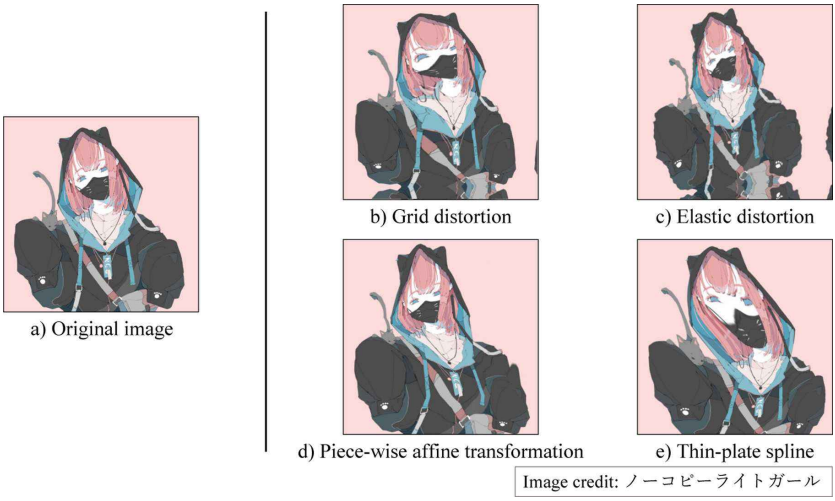

Fig. 2. Sample images of the four spatial transforms used to train our distortion recognition model.

We applied this distorted image creation pipeline while a batch of images was sampled as the input to the model. This made our method select the transformation and the parameters every time an image was sampled, which allowed us to prepare a pseudo-infinite number of distorted images. Creating distorted images on-the-fly, prevents the distortion recognition model from over-fitting to the training samples, and it also enables our method to be trained with limited amount of training data.

3.2 Training a Distortion Recognition Model

Model Architecture. The architectures of deep learning models have been remarkably developed since the advent of deep learning, and their performance has also been improved. Especially, the convolutional neural networks (ConvNets) which are composed of multiple convolutional layers has been useful in the field of image processing because of its superior compatibility with images. In this study, we adopt a type of ConvNet architecture called ConvNeXt presented by Liu et al. [19]. ConvNeXt is inspired by a deep learning model that employs an architecture which differs from that of ConvNet, called vision transformers (ViT) [5], while retaining ConvNet's compatibility with image processing. ViT has attracted attention in recent years and, like ConvNets, is used in various fields of image processing. ViT employs a mechanism called self-attention, which is capable of extracting global features of an image, in contrast to ConvNet, which uses convolution layers that are good at extracting local features. However, self-attention is computationally expensive and requires a longer time for training and testing. ConvNeXt, on the other hand, is a model in which self-attention is replaced by a depth-wise convolution layer with a large kernel

size, and the order of activation functions and convolution layers is made similar to the basic building block of ViT. Since it is an fully convolutional neural network, it maintains compatibility with image processing of ConvNets, while incorporating the advantages of ViT by making the basic structure similar. In recent years, ViT has performed better in the field of image processing compared to ConvNets, but in our preliminary experiments, we did not find any improvement in the loss values in the early stages of training with Swin Transformer v2 [18], which is a type of ViT.

Hyper-parameters. Four types of ConvNeXt have been proposed, depending on the number of parameters. In this study, we used ConvNeXt-Tiny configuration, which has the smallest number of parameters. Cross entropy was used for the loss function, and AdamW [21] was used as the optimizer with a learning rate of 0.001. We used the batch size of 128 and trained the ConvNeXt for 100 epochs on a RTX Titan GPU. The learning rate was adjusted using a cosine annealing scheduler [20] so that the learning rate was gradually reduced to zero by the end of training. For the four spatial transforms, we selected one of the transforms to be applied with equal probability. The parameters for each transform were sampled uniformly from the ranges shown in Table 1 for each image.

Dataset. Our dataset consists of 260,000 images randomly selected from the Danbooru dataset[1], and divided into 250,000, 5,000, and 5,000 images for training, validation, and testing, respectively. We have limited the random selection of images from Danbooru dataset to those that have a source URL attached in their metadata. From the validation and testing set, 2,500 images each were randomly extracted before training the model and one of the four transforms was applied to each image. These transformed images were saved so that distorted and non-distorted classes were kept unchanged during training. For the training set, images were selected with a probability of 50% whether spatial transform was applied or not, so that the ratio of distorted images to undistorted images was $1:1$.

Implementation Details. Among the four image transforms, grid distortion, elastic distortion and piece-wise affine transformation were performed using the implementation in albumentations[2], a library for image augmentation. For the thin-plate spline transformation, we modified an open-source implementation[3] to enable random selection of parameters. We used PyTorch [26], an deep learning library, for training the distortion recognition model. We used the implementation of ConvNeXt model in PyTorch's framework for modeling and training deep learning models, called transformers of huggingface [35]. To speed up the training, we used automatic mixed-precision [24], which is a method to stably train deep learning models using half-precision floating point numbers.

[1] https://www.gwern.net/Danbooru2021.
[2] https://albumentations.ai/.
[3] https://github.com/cheind/py-thin-plate-spline.

4 Evaluation

We evaluated the proposed method both quantitatively and qualitatively. For the quantitative evaluation, we put illustrations synthesized by the image generator into a set of real images drawn by illustrators and compared the performance of extracting the generated images. For the qualitative evaluation, we present and discuss the results of reordering of the synthesized images by the predicted visual quality using actual generated images. More detailed explanations are given in the following subsections.

For the image generator, we trained a well-known GAN model, StyleGAN2 [16] using illustrations of characters. The generator was capable of synthesize images of illustrated full-body characters with a resolution of 512×256 pixels.

4.1 Quantitative Evaluation

To quantitatively evaluate the proposed method, we conducted an experiment to predict the authenticity of illustration images. Specifically, we mixed synthetic images into a set of real illustrations collected from Shutterstock[4] with permission to use in publications. Then neural networks were used to predict whether the images were automatically generated or drawn by illustrators.

As comparison methods, we used Illustration2Vec (I2V) [29] and the discriminator used to train the image generator. I2V is a model for vectorizing illustrations proposed by Saito and Matsui. Since I2V uses image tagging as a learning task to train the vectorization network, it can not only vectorize illustrations but also assign appropriate tags according to their contents. In our experiments, among the 1,539 tags that I2V can assign, we used the tag "no human" which means that there is no human drawn in the image. If the model is convinced that this tag should be assigned, we could judge that the character in the image is collapsed. We used the pretrained model, distributed by the authors[5]. The discriminator was trained to discriminate whether the input images are from the dataset or generated images. Since the discriminator has knowledge about the generator, one criterion is to exceed the performance of the discriminator.

The output of each model for an input image was a probability value. Predictions for each model were defined as the result of thresholding this probability values. We used precision, recall, F1-score, and accuracy metric to evaluate the performance of the models. Precision measures the proportion of positive predictions that are actually correct. Recall, measures the proportion of actual positive examples that a classifier was able to identify correctly. The F1-score is an evaluation index that takes both precision and recall into account. The accuracy measures the percentage of correct predictions out of all predictions made by a classifier. Since the results change as the generated images change, the test was conducted on 5 different sets of randomly generated illustrations.

[4] https://www.shutterstock.com/.
[5] https://github.com/rezoo/illustration2vec.

Table 2. Results of the authenticity inference performance of the proposed and comparison methods. The number after @ indicates the threshold value and **bold** values indicate the best performing values.

model@threshold	Precision	Recall	F1-score	Accuracy
I2V@0.0001	0.671 ± 0.015	0.658 ± 0.017	0.652 ± 0.019	0.658 ± 0.017
Discriminator@0.5	0.692 ± 0.011	0.595 ± 0.012	0.536 ± 0.019	0.595 ± 0.012
Discriminator@0.3	0.705 ± 0.011	0.660 ± 0.014	0.640 ± 0.017	0.660 ± 0.014
ConvNeXt@0.5 (Ours)	0.765 ± 0.009	0.646 ± 0.017	0.601 ± 0.024	0.646 ± 0.017
ConvNeXt@0.3 (Ours)	$\mathbf{0.782 \pm 0.008}$	$\mathbf{0.688 \pm 0.015}$	$\mathbf{0.659 \pm 0.020}$	$\mathbf{0.688 \pm 0.015}$

Results. Table 2 presents the calculated metrics when thresholding with different values for each model. The values reported are the average and standard deviation of the results of five runs. It is shown that the proposed method with a threshold value of 0.3 has the best performance compared to the comparison method in all evaluation metrics.

Among the methods compared in this experiment, the discriminator had the lowest performance. This is due to the objective when training GANs, where the loss becomes low if the discriminator misidentifies synthesized images as real, even if these images are corrupted. The generator has the knowledge to fool the discriminator, which is independent of visual quality, and thus the results of the discriminator were less accurate than other methods. We further analyse this behavior in Sect. 4.2.

Using the thresholds of 0.3 and 0.5 for "no human" tag of I2V resulted to infer that all images contained human characters. Therefore, for I2V only, we selected an appropriate threshold value by inputting several images to the model and observing the output values. What should be noted in the results of I2V compared to the proposed method is that the evaluation was based on the contents in the images. We believe that better results can be obtained by considering both the quality of the image and its contents. This is an issue to be addressed in our future works.

4.2 Qualitative Evaluation

For qualitative evaluation, we generated several images and sorted them in the order of the probability that the proposed method classified them as the undistorted class. In the remainder of this subsection, we describe the images used in the evaluation, the results of the sorting, and the discussion on the results.

Non-duplicate Confirmation. In the research area of generative models, it has been pointed out that generators may synthesize copies of data that exist in the training set [6,33], which may lead to copyright issues. We confirmed that the generated images shown in the figures of this paper are not copies of the training data by using method for detecting data copies proposed by Pizzi

Fig. 3. Results of sorting synthetic images via the confidence of deep learning model's predictions. Images presented on this figure were confirmed that they are not replications of images in the training set. Zoom in for best view.

et al. [27]. Specifically, a deep learning model trained to detect copied images was used to embed arbitrary images to feature vectors, and these vectors can be used to calculate the similarity between other images. The feature vectors extracted from the synthesized images and the training data were used to calculate the cosine similarity. We observed the top 10 similar images sorted by the calculated cosine similarity and confirmed that the synthesized images were not copies of the training data.

Results. The sorted synthetic images for the three models used in our evaluation are presented in Fig. 3. The proposed method was able to reorder the images in an order close to that of human senses. The image in the 5th position is collapsed, but not distorted, so it may have been difficult to determine with the proposed method, which did not learn degradations of images other than distortion.

As we pointed out in Sect. 4.1, it can be seen that the generator was able to produce images that can fool the discriminator even though the images were collapsed. The character image sorted to be the highest quality of the discriminator has a silhouette that looks like a character, but the face and arms are not generated. Furthermore, the ranking of the 7th image is low despite the fact that it is a relatively clean synthesized image, since it has no missing body parts and the details of the face are clear.

The results show that I2V's "no human" tag output looks at whether the body parts are aligned. This can be seen from the fact that the 5th image was ranked higher than the other methods. The images ranked 6th, 7th, and 8th have simplified body parts or no arms, whereas the image ranked 5th has well-defined legs and arms, which is considered to be the reason for its high ranking. However, the 5th-ranked image is not favorable as a result, because although the head is present, it does not generate any facial details.

4.3 Limitation

It is nearly impossible to predict what kind of distortion an image generator will produce, since it depends on a variety of factors, including the data used for training, the training method of the generator, and the architecture of the model. There are limitations to reproducing this complex distribution of distortions using only parametric methods. The study of degraded image restoration has a similar limitation, since images are degraded using simple transformations and used to train the restoration model. In real-world super resolution, which is the study of super-resolving degraded low-resolution images, methods that train neural networks capable of reproducing the distribution of real-world image degradation from external datasets have emerged [7,34]. This research direction will improve the performance of our task. In addition, as mentioned in Sect. 4.1, the proposed method evaluates high quality images by considering only spatial distortion, therefore the aggregation with the content-aware method is our future issue.

5 Conclusion and Future Works

In this study, we presented a distortion recognition-based method for extracting high-quality images from a synthetic image set. We transformed images from the training data to automatically synthesize distorted image samples to train the classification model. We demonstrated that simple spatial transformation methods can adequately simulate distortions on synthetic data produced by deep learning-based generative models. Moreover, our method can be trained without any knowledge of the synthetic data, therefore has some robustness in the contents of the synthesized images. We believe that this study will help to create high-quality avatars, which have become an important communication tool in virtual spaces.

We used distortion as the corruption on synthesized images, but other corruptions can be compared to find more effective transformations to improve our method. Also, our method uses the visual quality as the basis for the ranking, but we believe that methods focusing on the contents can also be explored. In our future work, we will study methods that take both content of the image and visual quality into account.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 214–223 (2017)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (ICLR) (2019). https://openreview.net/forum?id=B1xsqj09Fm
3. Chen, Y., Zhao, Y., Li, S., Zuo, W., Jia, W., Liu, X.: Blind quality assessment for cartoon images. IEEE Trans. Circuits Syst. Video Technol. **30**, 3282–3288 (2020). https://doi.org/10.1109/TCSVT.2019.2931589
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
5. Dosovitskiy, A., et al.: An Image is Worth 16×16 Words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2020). https://openreview.net/forum?id=YicbFdNTTy
6. Feng, Q., Guo, C., Benitez-Quiroz, F., Martinez, A.M.: When Do GANs replicate? on the choice of dataset size. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6701–6710 (2021)
7. Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3599–3608 (2019). https://doi.org/10.1109/ICCVW.2019.00445
8. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 27, pp. 2672–2680 (2014)
9. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved Training of Wasserstein GANs. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
10. Guo, H., Hu, S., Wang, X., Chang, M.C., Lyu, S.: Eyes tell all: irregular pupil shapes reveal GAN-generated faces. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2904–2908 (2022). https://doi.org/10.1109/ICASSP43922.2022.9746597
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30, pp. 6626–6637 (2017)
12. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840–6851 (2020)
13. Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., Liu, Z.: AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. ACM Trans. Graph. **41** (2022). https://doi.org/10.1145/3528223.3530094
14. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Advances in Neural Information Processing Systems, pp. 12104–12114 (2020)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410, June 2019
16. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8110–8119, June 2020

17. Liu, B., Zhu, Y., Song, K., Elgammal, A.: Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. In: International Conference on Learning Representations (ICLR) (2021). https://openreview.net/forum?id=1Fqg133qRaI
18. Liu, Z., et al.: Swin transformer V2: scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12009–12019 (2022)
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986 (2022)
20. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: International Conference on Learning Representations (ICLR) (2017). https://openreview.net/forum?id=Skq89Scxx
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2018). https://openreview.net/forum?id=Bkg6RiCqY7
22. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2794–2802 (2017)
23. Mescheder, L., Geiger, A., Nowozin, S.: Which Training Methods for GANs do actually Converge? In: Proceedings of the 35th International Conference on Machine Learning (ICML), vol. 80, pp. 3481–3490 (2018)
24. Micikevicius, P., et al.: Mixed Precision Training. In: International Conference on Learning Representations (ICLR) (2018). https://openreview.net/forum?id=r1gs9JgRZ
25. Ni, Z., Zeng, H., Ma, L., Hou, J., Chen, J., Ma, K.K.: A gabor feature-based quality assessment model for the screen content images. IEEE Trans. Image Process. **27**, 4516–4528 (2018). https://doi.org/10.1109/TIP.2018.2839890
26. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS). 32, pp. 8026–8037 (2019). https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
27. Pizzi, E., Roy, S.D., Ravindra, S.N., Goyal, P., Douze, M.: A self-supervised descriptor for image copy detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14532–14542, June 2022
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695 (2022)
29. Saito, M., Matsui, Y.: Illustration2Vec: a semantic vector representation of illustrations. In: SIGGRAPH Asia 2015 Technical Briefs, pp. 1–4 (2015). https://doi.org/10.1145/2820903.2820907
30. Salimans, T., et al.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 29, pp. 2234–2242 (2016)
31. Simard, P.Y., Steinkraus, D., Platt, J.C., et al.: Best practices for convolutional neural networks applied to visual document analysis. In: International Conference on Document Analysis and Recognition (ICDAR), vol. 3, pp. 958–963 (2003)
32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826, June 2016

33. Tinsley, P., Czajka, A., Flynn, P.: This Face Does Not Exist... But It Might Be Yours! Identity Leakage in Generative Models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1320–1328 (2021)
34. Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., Song, H.: Unsupervised real-world image super resolution via domain-distance aware training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13385–13394 (2021)
35. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6
36. Wu, Y., Deng, Y., Yang, J., Wei, F., Chen, Q., Tong, X.: AniFaceGAN: animatable 3D-aware face image generation for video avatars. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022). https://openreview.net/forum?id=LfHwpvDPGpx
37. Yang, X., Li, F., Liu, H.: A survey of DNN methods for blind image quality assessment. IEEE Access **7**, 123788–123806 (2019). https://doi.org/10.1109/ACCESS.2019.2938900
38. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient GAN training. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 7559–7570 (2020)
39. Zhao, Y., Ren, D., Chen, Y., Jia, W., Wang, R., Liu, X.: Cartoon image processing: a survey. Int. J. Comput. Vision **130**, 2733–2769 (2022). https://doi.org/10.1007/s11263-022-01645-1
40. Zhao, Z., Singh, S., Lee, H., Zhang, Z., Odena, A., Zhang, H.: Improved consistency regularization for gans. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 11033–11041 (2021)