Full-Body High-Resolution Anime Generation with Progressive Structure-Conditional Generative Adversarial Networks

Koichi Hamada(✉), Kentaro Tachibana, Tianqi Li, Hiroto Honda, and Yusuke Uchida

DeNA Co., Ltd., Tokyo, Japan
{koichi.hamada,kentaro.tachibana,tianqi.li,
hiroto.honda,yusuke.a.uchida}@dena.com

Abstract. We propose Progressive Structure-conditional Generative Adversarial Networks (PSGAN), a new framework that can generate full-body and high-resolution character images based on structural information. Recent progress in generative adversarial networks with progressive training has made it possible to generate high-resolution images. However, existing approaches have limitations in achieving both high image quality and structural consistency at the same time. Our method tackles the limitations by progressively increasing the resolution of both generated images and structural conditions during training. In this paper, we empirically demonstrate the effectiveness of this method by showing the comparison with existing approaches and video generation results of diverse anime characters at 1024×1024 based on target pose sequences. We also create a novel dataset containing full-body 1024×1024 high-resolution images and exact 2D pose keypoints using Unity 3D Avatar models.

Keywords: Generative adversarial networks · Anime generation · Image generation · Video generation

1 Introduction

Recently automatic image and video generation using deep generative models has been studied [5,10,21]. These are useful for media creation tools such as photo editing, animation production and movie making. Focusing on anime creation, automatic character generation can inspire experts to create new characters, and also can contribute to reducing costs for drawing animation. Jin et al. [9]

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-11015-4_8) contains supplementary material, which is available to authorized users.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11131, pp. 67–74, 2019.
https://doi.org/10.1007/978-3-030-11015-4_8

focuses on image generation for anime character faces with GAN architecture. However full-body character generation has not been studied enough. Generation of images for anime characters which only focused on face images was proposed, however, its quality was not satisfactory for animation production requirements [9]. To generate full-body characters automatically and add actions to them with high quality is a great help for making new characters and drawing animations. Therefore, we work on generating full-body character images and adding actions to them (i.e., video generation) with high quality.

There remain two problems to applying full-body character generation to animation production: (i) generation with high-resolution, (ii) generation with specified sequences of poses.

Generative Adversarial Networks (GANs) [5] are one of the most promising candidates as a framework applied to a diverse range of image generation tasks [8,9,14,16,17,25]. Recent progress of GANs with hierarchical and progressive structures has been realizing high-resolution and high-quality image generation [10], text-to-image synthesis [23,24], and image synthesis from label map [22]. However, It is still a challenge for GANs to generate structured objects consistent with global structures [4], such as full-body character generation. On the other hand, GANs with structural conditions, such as pose keypoints and facial landmarks, have been also proposed [1,7,13–15,18,19]. However, their image resolution and quality are insufficient.

We propose Progressive Structure-conditional GANs (PSGAN) to tackle these problems by imposing the structural conditions at each scale generation with progressive training. We show that PSGAN is able to generate full body anime characters and animations with target pose sequences at 1024×1024 resolution. As PSGAN generates images with latent variables and structural conditions, PSGAN is able to generate controllable animations for various characters with target pose sequences. Figure 1 shows some example of animation generation results.

2 Proposed Methods

2.1 Progressive Structure-Conditional GANs

Our key idea is to learn image representation with structural conditions progressively. Figure 2 shows generator G and discriminator D architecture of PSGAN. PSGAN increases the resolution of generated images with structural conditions at each scale and generates high-resolution images. We adopt the same architecture of the image generator and discriminator as Progressive GAN [10], except that we impose structural conditions on both the generator and discriminator at each scale by adding pose maps with corresponding resolutions, which significantly stabilizes training. GANs with structural conditions have also been proposed [1,7,13–15,18,19]. They exploit a single-scale condition while we use multi-scale conditions. More specifically, we downsample the full-resolution structural condition map at each scale to form multi-scale condition maps. For each scale,

Fig. 1. Generated images of full-body anime characters at 1024×1024 by PSGAN with a test pose sequence. A generated anime at 1024×1024 by PSGAN is at https://youtu.be/bIi5gSITK0E.

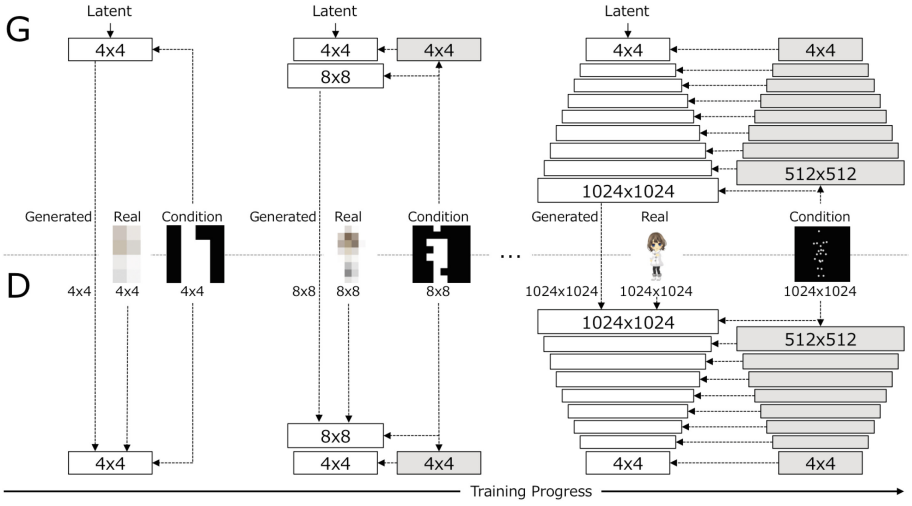

Fig. 2. Generator (G) and Discriminator (D) architecture of PSGAN.

the generator generates an image from a latent variable with a structural condition and the discriminator discriminates the generated images and real images based on the structural conditions. $N \times N$ white boxes stand for learnable convolution layers operating on $N \times N$ spatial resolution. $N \times N$ gray boxes stand for non-learnable downsampling layers for structural conditions, which reduce spatial resolution of the structural condition map to $N \times N$. We use M channels for representation of M-dimensional structural conditions (e.g. M keypoints).

2.2 Automatic Dataset Construction with Exact Pose Keypoints from Unity 3D Models

We create a novel dataset containing full-body high-resolution anime character images and exact 2D pose keypoints using the Unity[1] 3D models for various poses, in a similar manner as is done in [3,20] for photo-realistic images. We use various motions and costumes of full-body character models to create this dataset. The four key features of our methodology are the following: (1) Pose Diversity: To generate smooth and natural animation we prepare a very wide variety of pose conditions. We generate high-resolution images and pose keypoint coordinates of various poses for reproducing smooth and natural continuous motion by capturing images and exactly calculating the coordinates while each Unity 3D model is moving with each Unity motion. (2) Exact pose keypoints: Direct calculation of pose keypoint coordinates from the Unity model makes it possible to calculate the coordinates with no estimation error. (3) Infinite number of training images: An infinite number of synthetic images with keypoint maps are obtained by generating 3D modeled avatars using Unity automatically. Various images with keypoints can be created by replacing detachable items for each Unity 3D model. (4) Background elimination: We can set the background color to white and erase unnecessary information to avoid negative effects on image generation.

3 Experiments

We evaluate the effectiveness of the proposed method in terms of quality and structural consistency of generated images on the Avatar Anime-Character dataset and DeepFashion dataset. We show comparisons between our method and existing works.

3.1 Datasets

In this section, we describe our dataset preparation methodology. For PSGAN we require pairs of image and keypoint coordinates. We prepare the original Avatar Anime-Character dataset synthesized by Unity, and DeepFashion dataset [12] with keypoints detected by Openpose [2].

[1] Unity: https://unity3d.com.

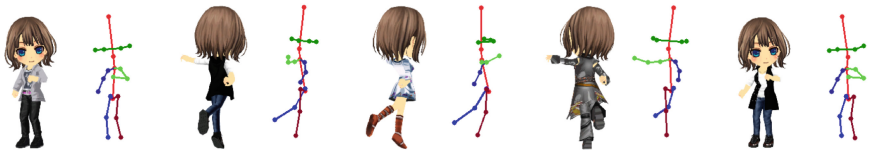

Fig. 3. Samples of Avatar Anime-Character dataset.

Avatar Anime-Character Dataset. We create a dataset of full-body 1024×1024 high-resolution images with exact 2D pose keypoints from Unity 3D Avatar models based on the above proposed method. We divide several continuous actions of one avatar into 600 poses, and calculate keypoints in each pose. We conduct such process for 69 kinds of costumes, and obtain 47,400 images in total. We also obtain 20 keypoints based on the location of the bones of the 3D model. Figure 3 shows samples of created data. Anime characters (left of pair) and pose images (right of pair) are shown.

DeepFashion Dataset. The DeepFashion dataset (In-shop Clothes Retrieval Benchmark) [12] consists of 52,712 in-shop clothes images, and 200,000 cross-pose/scale pairs. All images are in 256×256 resolution and are richly annotated by bounding box, clothing type and pose type. However, none of them has keypoint annotations. Following [14], we use Openpose [2] to extract keypoint coordinates from images. The number of keypoints is 18 and examples with less than 10 detected keypoints are omitted.

3.2 Experimental Setups

We use the same stage design and the same loss function as [10]. We train networks with 600k images and structural conditions for each stage and use WGAN-GP loss [6] with $n_{critic} = 1$. We use a minibatch size 16 at the stage for 4×4 – 128×128 image generation and gradually decrease it to 12 for 256×256, 5 for 512×512, and 2 for 1024×1024 respectively due to GPU memory constraints. We use M channels of structural conditions as pose keypoints. M is 20 for the Avatar Anime-Character dataset and 18 for DeepFashion dataset. At each scale, the single pixel value at the corresponding keypoint coordinate is set to 1 and -1 elsewhere. For downsampling the condition map, we use max-pooling with kernel size 2 and stride 2 at each scale. We train the networks using Adam [11] with $\beta_1 = 0$, $\beta_2 = 0.99$. We use $\alpha = 0.001$ at the stage for 4×4 – 64×64 image generation and gradually decrease it to $\alpha = 0.0008$ for 128×128, $\alpha = 0.0006$ for 256×256, $\alpha = 0.0002$ for 512×512, and $\alpha = 0.0001$ for 1024×1024 respectively.

3.3 Avatar Anime-Character Generation at 1024×1024

We show examples of a variety of anime characters and animations generated at 1024×1024 by PSGAN. Figure 1 shows generated results of full-body anime

characters at 1024×1024 with a test pose sequence. We can generate new full-body anime characters by interpolating latent variables corresponding to anime characters with different costumes (character 1 and 2) for various poses. By fixing the latent variables and giving continuous pose sequences to the network, we can generate an animation of the specified anime characters[2].

3.4 Comparison of PSGAN, Progressive GAN, and PG2

First, we evaluate structural consistency of PSGAN compared to Progressive GAN [10]. Figure 4 shows generated images on the DeepFashion dataset (256×256) by Progressive GAN and PSGAN. We observe that Progressive GAN is not capable of generating natural images consistent with their global structures (for example, left four images). On the other hand, PSGAN can generate plausible images consistent with their global structures by imposing the structural conditions at each scale.

Fig. 4. Comparison of structural consistency with [10] on DeepFashion dataset.

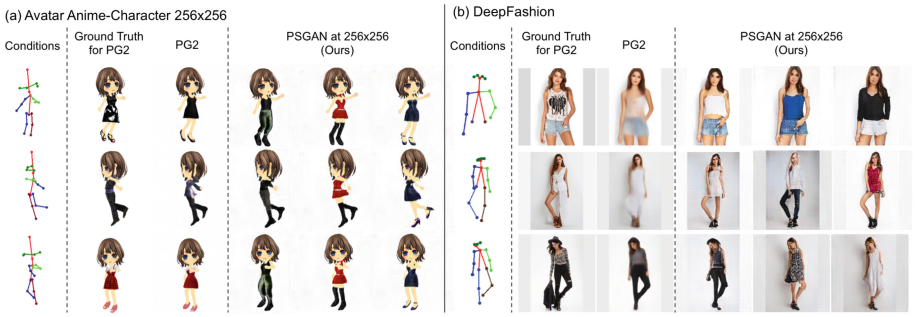

Fig. 5. Comparison of generated image quality based on pose conditions with [14] at 256×256 on (a) Avatar Anime-Character dataset and (b) DeepFashion dataset.

[2] An illustration video for adding action to full-body anime characters with PSGAN is at https://youtu.be/0LQlfkvQ3Ok.

Next, we evaluate image quality on pose conditional image generation of PSGAN compared to Pose Guided Person Image Generation (PG2) [14]. PG2 requires a source image and a corresponding target pose to convert the source image to an image with the structure of the target pose. Meanwhile, PSGAN generates an image with the structure of the target pose from latent variables and the target pose and does not need paired training images. Figure 5 shows generated images of PSGAN and PG2 on the 256×256 resolution version of the Avatar dataset and DeepFashion dataset. We pick the weight parameter for L1 loss of PG2 (which affects image quality) to 1.0. The input image of PG2 is omitted. We can observe the generated images of PSGAN are less blurry and more detailed than PG2 due to structural conditions imposed at each scale.

4 Conclusion

In this paper, we have demonstrated smooth and high-resolution animation generation with PSGAN. We have shown that the method can generate full-body anime characters and the animations based on target pose sequences at 1024×1024 resolution. PSGAN progressively increases the resolution of generated images with structural conditions at each scale during training and generates detailed images for structured objects, such as full-body characters. As PSGAN generates images with latent variables and structural conditions, it is able to generate controllable animations with target pose sequences. Our experimental results demonstrate that PSGAN can generate a variety of high-quality anime characters from random latent variables, and smooth animations by imposing continuous pose sequences as structural conditions. Since the experimental setting still remains limited, such as one avatar and several actions, we plan to conduct experiments and evaluation in various conditions. We plan to make the Avatar Anime-Character dataset available in the near future.

References

1. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: Proceedings of CVPR (2018)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of CVPR (2016)
3. Chen, W., et al.: Synthesizing training images for boosting human 3D pose estimation. In: Proceedings of 3D Vision (2016)
4. Goodfellow, I.: NIPS 2016 tutorial: generative adversarial networks. arXiv:1701.00160 (2017)
5. Goodfellow, I.J., et al.: Generative adversarial nets. In: Proceedings of NIPS (2014)
6. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of Wasserstein GANs. In: Proceedings of NIPS (2017)
7. Hu, Y., Wu, X., Yu, B., He, R., Sun, Z.: Pose-guided photorealistic face rotation. In: Proceedings of CVPR (2018)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of CVPR (2017)

9. Jin, Y., Zhang, J., Li, M., Tian, Y., Zhu, H.: Towards the high-quality anime characters generation with generative adversarial networks. In: Proceedings of NIPS Workshop on Machine Learning for Creativity and Design (2017)
10. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, and stability, and variation. In: Proceedings of ICLR (2018)
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimizations. In: Proceedings of ICLR (2015)
12. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of CVPR (2016)
13. Ma, L., Sun, Q., Georgoulis, S., Gool, L.V., Schiele, B., Fritz, M.: Disentangled person image generation. In: Proceedings of CVPR (2018)
14. Ma, L., Sun, Q., Jia, X., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. In: Proceedings of NIPS (2017)
15. Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., Wang, H.: Geometry-contrastive generative adversarial network for facial expression synthesis. arXiv:1802.01822 (2018)
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proceedings of ICLR (2016)
17. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: Proceedings of ICML (2017)
18. Si, C., Wang, W., Wang, L., Tan, T.: Multistage adversarial losses for pose-based human image synthesis. In: Proceedings of CVPR (2018)
19. Siarohin, A., Sangineto, E., Lathuiliere, S., Sebe, N.: Deformable GANs for pose-based human image generation. In: Proceedings of CVPR (2018)
20. Varol, G., et al.: Learning from synthetic humans. In: Proceedings of CVPR (2017)
21. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Proceedings of NIPS (2016)
22. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of CVPR (2018)
23. Zhang, H., et al.: Stackgan++: realistic image synthesis with stacked generative adversarial networks. TPAMI (2018)
24. Zhang, Z., Xie, Y., Yang, L.: Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In: Proceedings of CVPR (2018)
25. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of ICCV (2017)