IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

LINE-BASED REALISTIC IMAGE GENERATION

Chen Zihan[a], Xie Xiaolan[b], Xu Hong[c], Chen Xuejin[d]
College of Information Science and Engineering
Guilin University of Technology
Jiangan Road 12, Guilin,China

Abstract In recent years, marked by the advent of generative adversarial networks, people have gained very remarkable progress in tackling various image generation tasks. However, great challenges are imposed in generating high-quality natural images from simple lines. This paper improves the pix2pix model utilizing self-attention mechanism and multiple level discriminator, introduces self-attention module and multiple level discriminator to pix2pix to build an improved network architecture transformed from line to image. pix2pix and the improved network were tested using image dataset CelebA and face edge image extracted therefrom. By comparing the realistic face images generated before and after the improvement, it can be intuitively found that the improved network can synthesize more natural facial features, light and shadow. The three objective evaluation scores calculated based on the generated image also prove that the improved network has superior generation effect than pix2pix in overall. Finally, free-hand sketches of human faces were input to the well-trained network. It was further verified that the improved network also had obvious advantages in generating realistic face images using fee-hand sketches.

Keywords: Image generation; generative adversarial network; pix2pix; self-attention; multiple level discriminator.

1. INTRODUCTION

1.1. Topic selection purpose, background and significance

In the fields of computer vision, computer graphics and virtual reality, generating high-quality realistic images with degraded images has long attracted widespread attention, which already has applications in image denoising, image repair, image super-resolution, image coloring and image segmentation. In such realistic image generation characterized by image-to-image transformation, the input can be images such as a photo or images formed by lines. Edge map and freehand sketches formed by lines make up relatively simple and intuitive inputs, and these lines have structured image organization forms, which can express geometric structure information with rich objects or scenes. Using this information, it is possible to lead the transformation model to generate realistic images with specified contents, and further image editing is also possible. In recent years, the emergence of deep neural networks (DNNs) has brought very attractive development prospects for realistic image generation, of which, Generative Adversarial Network (GANs) [3] proposed in NIPS2014 conference has demonstrated great application potential. GAN generates images by zero-sum games between generators and discriminators. Trained generator attempts to generate realistic images based on noise input, while discriminator judges whether the output image of the generator is real or fake. Generating a more realistic image is possible by confrontation and competition between the two. Image generation based on input edge map or free-hand sketch can be expressed as an input line-based image translation problem. Literature [1] and [2] introduce deep learning methods for solving such problems using GAN as the basic framework. In this paper, based on the pix2pix method proposed in

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

literature [2], self-attention (SA) mechanism and multiple level discriminator (MLD) module are introduced to study how to generate realistic face images from sparse face edge maps and fee-hand sketches.

2. LINE-TO-FACE IMAGE GENERATION EXPERIMENT BASED ON PIX2PIX

2.1. pix2pix model

Many image processing problems can be reduced to the problem of "translating" an input image into a corresponding output image. Just as one concept can be expressed in different languages, a scene can be described using RGB images, gradient fields, edge skeleton maps or classification labels, etc. Traditionally, these tasks are usually handled by separate, specifically targeted algorithm mechanisms, but the internal settings of these mechanisms are essentially the same: pixel prediction from pixel, that is, mapping pixels to pixels (pix2pix).

The problem of image-to-image transformation is essentially the mapping of input grid to output grid. Although input and output have different specific surface forms, both are rendered on the same basic structure. Hence, input and output have roughly the same structure, which is considered in the design of pix2pix generator. Previously, this type of generator G will use a codec network, whose input is passed to the bottleneck layer through several down-sampled convolutional layers (encoders). Afterwards, the transfer process is reversed. The up-sampled deconvolution layers (decoders) in the same number as the encoder map back the original resolution. To extract and model the high-frequency parts (i.e., contour information) in the image, pix2pix adopts a new discriminator architecture D, and names it PatchGAN. PatchGAN only punishes the structure on the local image patch scale, which can effectively model the image as a Markov random field (that is, it is assumed that pixels with an interval greater than a patch diameter are independent). The discriminator determines the authenticity of each N×N small block in the image by classification. The N×N small block is generally much smaller than the image in size. By convolution of the entire input image using the discriminator, the final output of D can be obtained by averaging all responses.

2.2. Experimental dataset

The face image dataset used herein is CelebA dataset [6], which is a large-scale face dataset containing more than 200K celebrity images. This paper will adopt versions cropped and aligned by CelebA dataset, in which each image has a size of 218×178. In order to flesh out the original settings of pix2pix, this paper has cropped the center of the image and adjusted the image size to 256×256. In order to let the network learn more diverse face information, 30,000 face images in the CelebA dataset were randomly selected, covering men and women, young and old, multiple ethnicities and hair styling, plus different expressions.

2.3. Experimental evaluation criteria

To evaluate the overall generation quality of abundant images and evaluate the real performance of the algorithm, this paper will adopt three evaluation parameters: Inception Score (IS) [7], kernel maximum mean discrepancy (KMMD) [8], and Fréchet Inception Distance (FID) [9]. Inception Score

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

is a metric that automatically evaluates the quality of image generation models. The industry has proven that this standard has a good correlation with human scores on realism in generated image. Kernel MMD shows amazing working performance when running in the feature space of pre-trained ResNet, and the sample complexity and computational complexity are relatively low. Fréchet Inception distance is used to measure the difference between two Gaussian distributions, which can be used to evaluate GAN measurement method. In the comparative evaluation of the image quality using the above three evaluation indicators, in general, the higher the IS value, the better, and the lower the KMMD and FID, the better.

2.4. Experimental results

In pix2pix network training, this paper adopts mini-batch SGD and adaptive moment estimation (Adam) optimizer. There are 30,000 samples in the training set used in the experiment. Eight samples were randomly selected each time to calculate the loss function and update the parameters. Fig. 1 shows the convergence of L1 loss function used to train the pix2pix network with the number of iteration steps. It can be seen that L1 loss function converges at around 700,000th step, which indicates that pix2pix network has completed training. The reason why the convergence of L1 loss function is chosen to judge whether the network is well trained is that L1 loss mainly describes the difference between the generated image and the input image, which can reflect the most direct goal of the network.

Subsequently, the well-trained network was used for image generation, and the face image results generated from 1000 randomly selected face edge maps were scored. Several better ones were selected for subjective display and comparison. It can be seen that compared with real image, pix2pix-generated image has acquired a lot of details, such as hair texture, eye highlights, etc. Using the foregoing three evaluation criteria, statistical tests were performed on the 1000 generated images, and the score results are shown in Table 1.

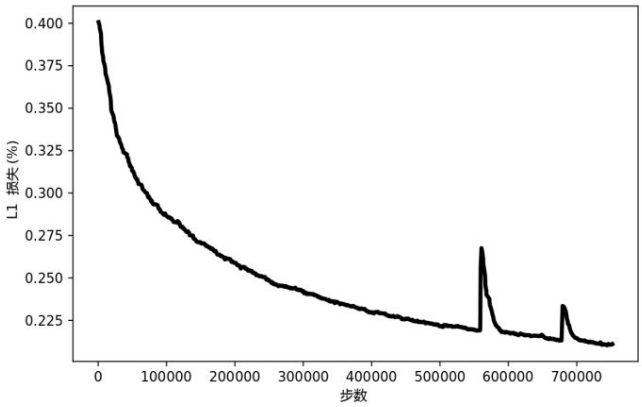


Figure 1. Convergence graph of pix2pix's L1 loss function.

Table 1. Three evaluation scores on pix2pix generation results.

Network	IS	KMMD	FID
pix2pix	2.5284	1.4484	273.2681

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

3. GENERATION MODELS FROM LINES TO HIGH-QUALITY FACE IMAGES

3.1. Self-attention module and pix2pix+SA model

In order to adapt to the condition setting of image-to-image transformation, and encourage the model to directly use the information of the conditional image, so that complex geometric constraints can be imposed on the global image structure more accurately, self-attention module, also called intra-attention, is introduced here. This module is designed as a general-purpose module under the conditional framework, which enables higher layers of the network to perceive conditional images and can be added after any existing modules.

The pix2pix+SA model structure adopted herein has two SA modules placed symmetrically on the second layer of the generator input network and the penultimate layer of the output network. The reason for considering the introduction of only two symmetrical SA modules is: multiple experiments indicate that if each pair of input and output networks of the generator is symmetrically added with a pair of SA modules, except that computational burden is significantly increased, quality of the generated image cannot be significantly improved.

3.2. Multiple Level Discriminator and pix2pix+SA+MLD Model

Based on the hypothesis that two pixels with an interval greater than the diameter of an image block are independent of each other, the discriminator used by pix2pix is partitioned. Such hypothesis is correct in some cases. For instance, it has been proven in texture generation and style transfer. Nevertheless, such hypothesis is invalid for global structure constraint, so it is impossible for partitioned discriminator to grasp the global structure information and then lead the generator to perceive the global face structure.

Pix2pix's discriminator adopts block-by-block convolution to distinguish real/synthesized images in a local receptive field much smaller than the input image. The average value of all responses is provided to the discriminator D as the final output, which is a hypothesis based on independent pixels (pixel spacing is greater than one block diameter). However, structural constraint belongs to global information spanning the entire image and block discriminator lacks the ability to capture such global information, so another global discriminator Dg is added here, which has a receiving domain as large as the entire image and can capture the global structure information. The block discriminator Dp and the global discriminator Dg share weights in the first few layers, which is because these discriminators share the same lower characteristics.

3.3. Experimental results

3.3.1 Experiment on pix2pix+SA model

As shown in Fig. 2, it can be seen from L1 loss function that pix2pix+SA has entered convergence at around 400,000th step, which is faster and smoother than pix2pix convergence. Although in the pix2pix+ SA simulation, the speed of calculating a batch is reduced by nearly one fold compared to

IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

pix2pix, in general, introduction of SA strengthens the generator capacity and further improves overall learning ability of the network.

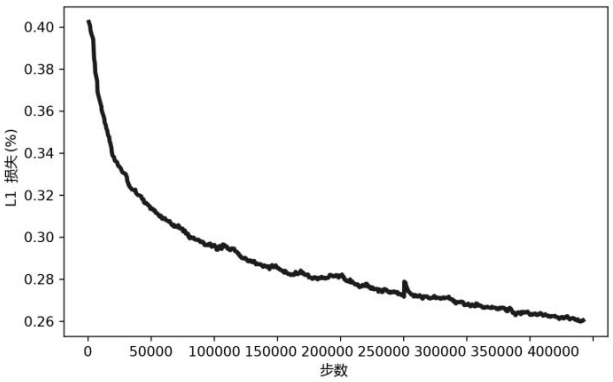


Figure 2. Convergence graph of L1 loss function of pix2pix+SA.

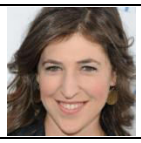
Face edge map	Ground Truth	pix2pix	pix2pix+SA
			

Figure 3. pix2pix+SA model image generation results.

Seen from the generated face image, as shown in Fig. 3, the introduced SA has a good binding effect on geometric construction of the five sense organs, but clearer images can be generated in pix2pix where it is blurred. As shown in Table 3.1.1, compared to pix2pix, pix2pix+SA has a significant drop in both MMD and FID scores, which indicates that the generated image is more realistic. Nonetheless, IS score is decreased here. Further research has been conducted on this. First, the 200 generated images used for the scoring test were increased to 1,000, while IS score only increased by no more than 0.1. Then, several dozens of images with poor generation quality were excluded, and IS score was even dropped by more than 0.3. This suggests that image with poor generation quality has a greater impact on IS score. In IS calculation, images with poor generation quality are deemed as expression of diversity.

Table 2. Comparison of the three evaluation scores on pix2pix and pix2pix+SA generation results.

Network	IS	KMMD	FID
Pix2pix	2.5284	1.4484	273.2681
pix2pix+SA	2.4136	0.9376	210.7394

3.3.2. Experiment on pix2pix+SA+MLD model

As can be seen from Fig. 4, pix2pix+SA+MLD converges faster, but the speed of calculating a batch in actual training is reduced by a fold compared to pix2pix+SA. In general, however, pix2pix+SA+

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

MLD enhances discriminator capacity, further improves overall learning ability of the network and converges faster.

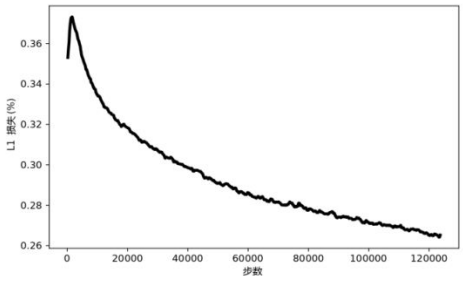


Figure 4. Convergence graph of L1 loss function of pix2pix+SA+MLD model.

Face edge map	Ground Truth	pix2pix	pix2pix+SA	pix2pix+SA+MLD
				

Figure 5. pix2pix+SA+MLD model image generation result.

Seen from the generated image, as shown in Fig. 5, the image generated by pix2pix+SA+MLD has more details compared to pix2pix+SA, and the overall light and shadow is more vivid than that of pix2pix+SA, indicating a deeper understanding of light and shadow by the network. The three evaluation scores of the generation results are shown in Table 3.

Table 3. Comparison of the three evaluation scores on pix2pix, pix2pix+SA and pix2pix+SA+MLD generation results.

Network	IS	KMMD	FID
Pix2pix	2.5284	1.4484	273.2681
pix2pix+SA	2.4136	0.9376	210.7394
pix2pix+SA+MLD	2.4145	1.2588	255.5071

Table 3 shows that although pix2pix+SA+MLD has higher scores than pix2pix, but is less scored than pix2pix+SA. Careful observation of the generated face image reveals that images generated bypix2pix+SA+MLD will show a large area of black spots near the nose. Although pix2pix+SA also shows such phenomenon in individual images, from the point of view of score, pix2pix+SA has better generation quality as a whole, which is because frequency of black spots is much lower in pix2pix+SA after addition of MLD. Although the latter two networks have respective advantages and disadvantages in the face image generated on the extracted edge map, pix2pix+SA+MLD demonstrates stronger generation ability in the face edge sketches drawn by hand. This will be discussed in the next section.

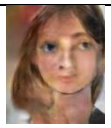
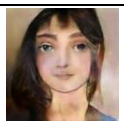
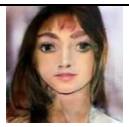
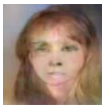
IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

3.4. Realistic image generation by free-hand sketches

To test the generation effect of free-hand sketches, two free-hand sketches of female face were finally selected after continuous attempts. The image generation results are shown in Fig. 6. Compared with pix2pix and pix2pix+SA, pix2pix+SA+MLD shows an obvious advantage in image generation by free-hand sketches.

Figure 6. Three model results of realistic image generation by free-hand sketches.

Free-hand sketch	pix2pix	pix2pix +SA	pix2pix+SA+ MLD
			
			

After adjusting the input free-hand sketches, it is found that appropriately reducing image contrast and adding local lines to the nose can effectively suppress black spots, as shown in Fig. 7. A pair of illustrations was used to extract face edge map, and the details were manually supplemented. Three generation results were obtained via pix2pix+SA+MLD network and the trained fixed parameters. Large areas of dark spots were visible above the nose in the first generated image. After proper addition of lines to the nose, the dark spots were significantly reduced. This suggests that after introducing more network structures into pix2pix, the network has higher requirements on the details of the input sketch. When the input sketch lacks some high-frequency information to express the details, pix2pix+SA+MLD network may show black spot effect, which is not easy to improve by strengthening network training.

Figure 7. Generation effect of removing dark spots by adjusting the input.

	Originally generated image	Complement nose contour information	Reduce contrast
Input sketch			
			

4. CONCLUSION

Based on the pix2pix network, this paper introduces a self-attention mechanism and a multiple level discriminator to build a new deep learning architecture that transforms lines to realistic images. By generating high-quality realistic face images from lines (face edge map and free-hand sketch), it is

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

verified that the proposed improvement measures are effective for face image generation, which not only improves the network's ability to learn the global structure and local structure of the image, but also enables the generated image to contain more rich details, thus more realistic. To generate a clearer image, relatively speaking, the improved method herein requires that the input lines be more accurate on the semantic level. At this point, although free-hand sketches are sparser, its semantic information may be more accurate compared to edge maps with unnecessary lines, so they are more advantageous in generating realistic images. Hence, in the subsequent work of this study, we will first build a training set of free-hand face sketches, and then attempt to make free-hand sketches and real image pairs with pix2pixHD as the baseline algorithm. It is possible to use the training scheme of SketchyGAN [1] to gradually improve the ratio of free-hand sketches in the training set, and finally reach the goal of generating higher quality realistic face images by attempts to introduce MRU module like SketchyGAN to filter the wrong lines in the edge map.

References

[1] Wengling Chen, James Hays., 2018, SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis. *In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018),* 9416-9425.

[2] Phillip Isola,Junyan Zhu,Tinghui Zhou, et al., 2017, Image-to-Image Translation with Conditional Adversarial Networks. *In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 5967-5976.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al., 2014, Generative Adversarial Nets. *In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2014)*, 1-9.

[4] Alexei A. Efros, Thomas K. Leung., 1999, Texture Synthesis by Non-parametric Sampling. *In Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV 1999)*,1033-1038.

[5] William T. Freeman, Thouis R. Jones, Egon C. Pasztor., 2002, Example-Based Super-Resolution. *IEEE Computer Graphics and Applications*, Vol.22 No.2, 56-65.

[6] James Hays, Alexei A. Efros., 2008, Scene Completion Using Millions of Photographs. *Communications of the ACM*, Vol.51No.10, 87-94.

[7] Tao Chen, Mingming Cheng, Ping Tan, et al., 2009, Sketch2photo: Internet Image Montage. *ACM Transactions on Graphics*, Vol.28 No.5, 124:1-124:10.

[8] Mathias Eitz, Ronald Richter, Kristian Hildebrand, et al., 2011, Photosketcher: Interactive Sketch-Based Image Synthesis. *IEEE Computer Graphics and Applications*, Vol.31 No.6, 56-66.

[9] Diederik P. Kingma, Max Welling., 2014, Auto-Encoding Variational Bayes.*arXiv preprint arXiv: 1312.6114,* 1-14.

[10] Aäron van den Oord, NalKalchbrenner, Oriol Vinyals, et al., 2016, Conditional Image Generation with PixelCNN Decoders. *arXiv preprint arXiv: 1606.05328,* 1-13.

[11] Mehdi Mirza, Simon Osindero., 2014, Conditional Generative Adversarial Nets. *arXiv preprint arXiv: 1411.1784*, 1-7.

[12] Han Zhang, Tao Xu, Hongsheng Li, et al., 2017, StackGAN: Text to Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks. *In Proceedings of the 2017 IEEE International Conference on Computer Vision (CVPR 2017),* 5908-5916.

[13] Han Zhang, Tao Xu, Hongsheng Li, et al., 2018, StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. I*EEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1-16.

2019, Volume 3, Issue 2, 65-73, DOI 10.6723/TERP.201912_3(2).0006

[14] Qifeng Chen, VladlenKoltun. Photographic Image Synthesis with Cascaded Refinement Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017), Oct. 2017, 1520-1529.

[15] Tingchun Wang, Mingyu Liu, Junyan Zhu, et al. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), June 2018, 8798-8807.