Image generation based on multi-channel encoder and dual attention module

Genyuan Zhang*[a], Xiaohua He[a], Jianhao Ding#[b]

[a]College of Media Engineering, Communication University of Zhejiang, Hangzhou, China; [b]School of Humanities and Digital Media, Hangzhou University of Electronic Science and Technology, Hangzhou, China

ABSTRACT

Aiming at the requirements of appearance content generation of industrial products, this paper proposes an intelligent appearance product generation network architecture based on multi-channel encoder and dual attention module. Through multiple encoders, our network can learn the semantic information at different levels of the image, and then intelligently generate the appearance product image through the learned semantic information. Our model takes the line diagram as the input and can also input the sample image of the image to be generated. The network can generate appearance products that are consistent with the contour of the line graph and the color of the sample graph. The experimental results of the algorithm are qualitatively and quantitatively evaluated to verify that the algorithm can effectively generate appearance product images with high quality.

Keywords: Image generation, multi-channel encoder, attention, industrial products

1. INTRODUCTION

With the development of artificial intelligence, especially deep learning, image generation technology has been greatly improved[1-5]. At present, there are three mainstream image generation technologies: automatic regression model[1], variational self-encoder[6], and generating countermeasure network[7]. Generating countermeasure network has become one of the most popular research directions in the field of image generation, even in the whole field of artificial intelligence. Many AI research companies are investing a lot of energy to develop and promote the generation of confrontation networks, including OpenAI, Facebook, Twitter, Apple, and other companies. Yann LECUN, a chief scientist of Facebook Artificial Intelligence Institute, praised the generation of confrontation network and called it "the most exciting idea in the field of machine learning in the past decade." The image generation algorithm based on the generation countermeasure network is also widely used. At present, it can generate digital, face, and other objects, generate various realistic indoor and outdoor scenes, generate specific images according to the object contour, and generate high-resolution images according to low-resolution images. Generative adversarial networks[7] (GANs) is a generative model proposed by Ian Goodfellow et al. in 2014, which is expressed by GANs later. The optimization process of GANs is a minimax game problem. The goal of the generator is to make the performance of the generated data on the discriminator consistent with the performance of the real data on the discriminator. The goal of the discriminator is to accurately distinguish whether the input data is real data or data generated by the generator. Through alternating training and iterative optimization of generator and discriminator, the purpose of confrontation learning is achieved. As a generation model, GANs effectively solve the problem of data generation that can establish natural interpretation and has extremely important enlightening significance for the development of the field of image generation. However, GANs also bring some new problems, especially for flexible manufacturing. Flexible manufacturing needs to generate diverse design results according to conditions, and the generated image resolution meets the requirements of manufacturing. As far as the current method is concerned, there are mainly the following problems:

(1) Lack of diversity

Taking the training of face images with GANs as an example, when generating images with a small resolution, the diversity of images generated by GANs can be guaranteed. However, when the author attempts to generate a higher

*zgy8711@sohu.com; #djh@hdu.edu.cn

International Conference on Computer Application and Information Security (ICCAIS 2021),
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122600V · © The Authors.
· doi: 10.1117/12.2637361

resolution image (e.g., 512) × 512 pixels), the face images generated by GANs are very similar, that is, the problem of insufficient diversity.

(2) Poor quality of high-resolution image generation

There are usually two problems when using the image generation algorithm based on a generation countermeasure network to generate high-resolution images.

First, the generated image is distorted. In practice, it is found that the reason for this problem is that in the training process, when the network is trained to a certain extent, the discriminator will not be updated, resulting in the generator finding a method that can deceive the discriminator, and this method often cannot generate effective images.

Secondly, the generated high-resolution image is blurred. When generating high-resolution images, some high-frequency details are difficult to be learned by the network model, resulting in the loss of high-frequency details in the generated high-resolution images, which is manifested in low image definition.

In order to stably generate high-quality and high-resolution images, a large number of scholars have proposed a series of image generation and promotion algorithms based on generation countermeasure networks, including LAPGAN[8] and PG-GAN[9]. These algorithms have their own characteristics and alleviate the above problems to a certain extent from different angles, but they still do not completely solve the problems of insufficient diversity and poor quality of high-resolution image generation. Aiming at the requirements of appearance content generation of industrial products, this paper proposes an intelligent appearance product generation network architecture based on a multi-channel encoder and dual attention module. Through multiple encoders, our network can learn the semantic information at different levels of the image and then intelligently generate the appearance product image through the learned semantic information. Our framework mainly studies the high-quality image generation of industrial products constrained by line graph to ensure that the generated appearance image of industrial products is consistent with the line graph in outline. Taking the appearance of shoes as a typical case, experiments show the superiority and effectiveness of our method.

2. METHOD

Based on the infrastructure of generation countermeasure network, this section designs the algorithm and model structure in combination with the characteristics of appearance products, and proposes an appearance product generation model based on multi-channel encoder and dual attention module. The goal is to learn the statistical laws of appearance products and generate high-quality appearance products according to the laws. In this model, the style attention module in the dual attention module is used to enhance the style representation and learning of micro images. At the same time, the style loss is introduced into the loss function to restrict the style of the generated results. The multi-channel encoder in the model has multiple branches with different field sizes. The branch network with large field of view pays attention to the macro structure, while the branch network with small field of view pays attention to the fine texture, so that the algorithm can not only generate the appearance products with reasonable macro content, but also generate the surface texture of micro products relatively carefully. Generally speaking, this model strives to generate images with good performance in content integrity and style continuity.

2.1 Network architecture

At present, the most advanced algorithms of image generation use the generation countermeasure network as the infrastructure. The training strategy of mutual game between generator and discriminator in the generation countermeasure network urges the generator to simulate the distribution of real data, making the generation result closer to the real image. The model can be roughly divided into two networks: generator and discriminator. In the training stage, generator and discriminator participate in the training together; In the test phase, only the generator is used to generate the image without the participation of the discriminator.

As shown in Figure 1, the generator network can be further divided into three parts: a multi-channel encoder in the front section of the network, a space style attention module in the middle of the network and a decoder in the back section of the network. The input to the generator is the input image (line graph). The final output image of the algorithm is the appearance product image generated according to the line graph. The input of the discriminator is the repaired image and the corresponding ground truth. Score them respectively and try to distinguish them. The multi-channel encoder in this algorithm can realize multi-scale feature extraction and multi-scale feature fusion, as shown in Figure 1. It has three parallel convolutional neural network branches. In this algorithm, the receptive field of some branches is rapidly increased by using expansion convolution[9], so that the final receptive field sizes of the three branches are different. Compared with the ordinary convolution of the same size, the expanded convolution does not change the number of pixels actually participating in the convolution point multiplication, but expands the size of the input window by sampling the pixels at a certain number of intervals. The multi-channel encoder has three parallel convolutional neural network branches. The form of input data is that the three branch networks accept the same input, and the three branch networks have exactly the same structure and configuration except for different expansion coefficients. The spatial attention module constructs a global view for each position in the image, effectively captures the long-distance dependence between different positions in space, makes each position obtain more information from more semantically similar positions, provides more basis for repairing missing positions, and enhances the reliability and accuracy of the repair results; The style attention module can enhance the model's representation of image style, cooperate with the style loss in the loss function, and restrict the repair results to maintain the style of the original image.

In addition to maintaining the content structure of the input line graph, the image generation task should also maintain the style of the guide image. This algorithm uses the style attention module to enhance the representation of the image style, and uses the style loss function to constrain the repair results. As mentioned above when introducing the style attention module, the feature Gram matrix is used to represent the style of the image. Specifically, the resulting image and the corresponding real image are respectively input into the vgg19 network pre trained on the Imagenet dataset (shown in Figure 2). The network structure is shown in Figure 2. Then extract the output features of the five convolution layers "conv1_1", "conv2_1", "conv3_1", "conv4_1" and "conv5_1" of the two images in the vgg19 network respectively, and calculate the Gram matrix for the features of each layer to obtain the style representation of the layer. The final style loss is calculated as follows:

$$L_{style} = \sum_{l} \left[\frac{1}{4C_l^2 N_l^2} \sum_{i,j} \left(Y_{ij}^l - \hat{Y}_{ij}^l \right)^2 \right]$$

where $\hat{Y}_{ij}^l$ and Y_{ij}^l respectively represent the $\hat{y}$ and y elements in the Gram matrix of l layer and, while C_l and N_l respectively represent the number of channels and width height product of l layer features.

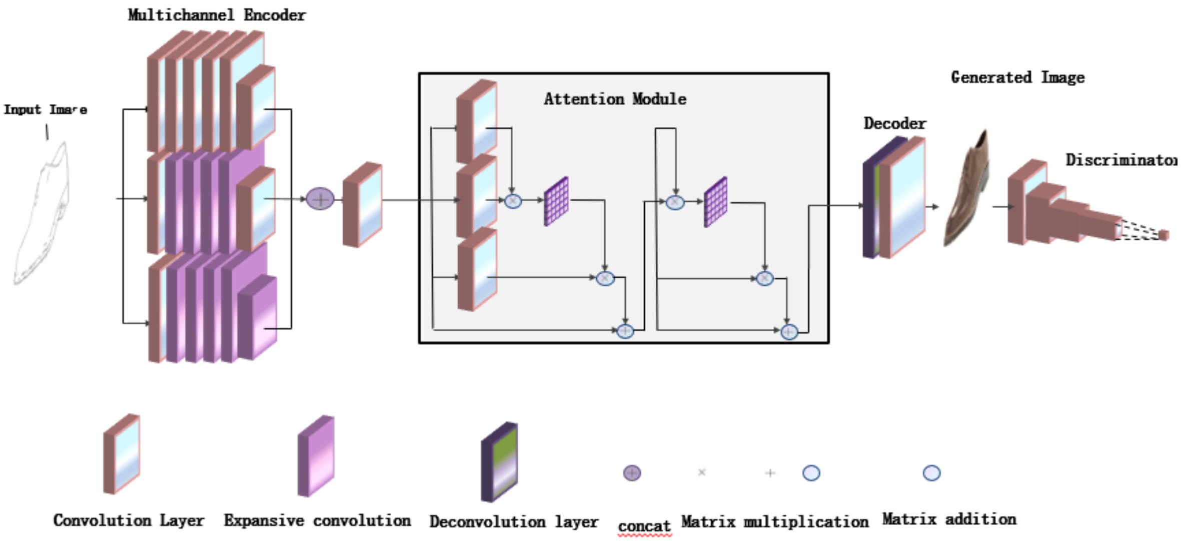

Figure 1. Intelligent generation network structure diagram of appearance products based on multi-channel encoder and dual attention module.

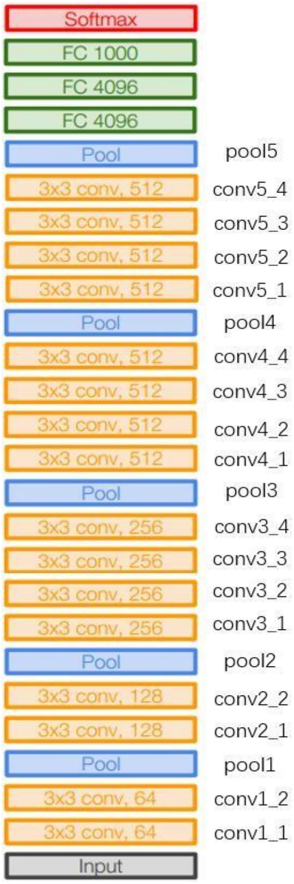

Figure 2. VGG network structure.

After the mean square error between gram matrices of each layer is calculated, it is accumulated layer by layer to obtain the overall style loss. The decoder structure used in this algorithm is shown in Table 1. The main task of the encoder is to use the latent variable Z to generate the converted image, restore the structure information of the original image on the basis of content features, and add domain feature information to realize cross domain image conversion.

The residual block structure used by the decoder is basically the same as that in the encoder. The characteristic image is expanded to the same size as the original image through two layers of up sampling. Leaky relu is used as the activation function in these two layers. The last layer of deconvolution does not change the size of the characteristic image, but changes the number of channels, and the number of compressed channels becomes 3, corresponding to RGB channels respectively; At the same time, the last layer of deconvolution uses tanh activation function to limit the output value to [-1,1], which is convenient to generate the converted image.

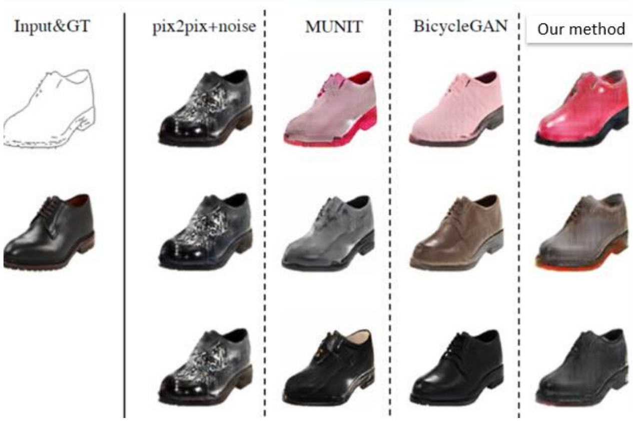

Figure 3. The results generated by our algorithm.

3. EXPERIMENTAL RESULTS AND ANALYSIS

The experiment is divided into two stages: training and testing. In the training phase, the size of the image is 256 × 256, the same size as the image in the training set. The discriminator receives the generated image and the real image in turn and gives scores respectively. The generator and discriminator are trained alternately. The goal of the discriminator is to distinguish them, and the goal of the generator is to produce repair results that can confuse the discriminator. In the experiment, Adam optimizer is used to optimize the parameters of generator and discriminator. The learning rate is 0.0001, and the two momentum parameters are 0.5 and 0.9 respectively. Set the batch size to 1. In the test phase, the line graph is used as the input of the generator, and the generated image is obtained after being processed by the generator. In the test stage, there is no need to participate in the discriminator and calculate the loss function.

Table 1. Decoder structure.

Layer	Size
z	128*64*256
Resblock	28*64*256
Resblock	128*64*256
Resblock	128*64*256
3*3*128 deconv, stride 2	256*128*128
3*3*64 deconv, stride 2	512*256*64
1*1*3 deconv, stride 1	512*256*3
Output	512*256*3

Experiments are done on the data set edge2shoes and compared with the latest methods (shown in Figure 3). In this experiment, the values of two evaluation indexes of four algorithms on test set I are calculated and sorted into Table 1. The larger the values of PSNR and SSIM (shown in Table 2), the higher the generation quality. The bold value of each row in the table represents the optimal value obtained on this index. It can be seen from the data in the table that this algorithm is ahead of other algorithms in two indicators.

Table 2. Metric results.

	Our method	Munit	BicycleGAN	Pix2pix
PSNR	**26.566**	24.826	26.006	23.386
SSIM	**0.810**	0.712	0.791	0.729

4. CONCLUSION

In order to better generate high-quality images, we propose an image generation method based on double attention mechanism and multi-scale. Our method can learn the statistical features of images from different semantic scales. At the same time, we can also learn the dependence of image style and content field range from a global perspective through the attention mechanism of style and content. In the future, we will introduce transformer architecture into the model to further improve the quality of image generation.

ACKNOWLEDGMENT

This paper is partly funded by Research demonstration application of key technologies of product appearance image generation for flexible manufacturing (No: LGG20F020003), Research and system development of the rapid display technology for the super large 3D scene of mobile wearable VR system (No: LGG18F020003), and Research on key technologies and system development of art image restoration based on style constraints (No: LGG20F020002).

REFERENCES

[1] van Oord, A., Kalchbrenner, N. and Kavukcuoglu, K., “Pixel recurrent neural networks,” Inter. Conf. on Machine Learning, 1747-1756 (2016)

[2] Kingma, D. P. and Welling, M., “Auto-encoding variational bayes,” Machine Learning, (2014). arXiv:1312.6114

[3] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., et al., “Generative adversarial networks,” Advances in Neural Information Processing Systems, 2672-2680 (2014).

[4] Gulrajani, I., Ahmed, F., Arjovsky, M., et al., “Improved training of wasserstein gans,” (2017). arXiv preprint arXiv:1704.00028

[5] Karras, T., Aila, T., Laine, S. and Lehtinen, J., “Progressive growing of GANS for improved quality, stability, and variation,” (2017). arXiv preprint arXiv:1710.10196

[6] van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A. and Kavukcuoglu, K., “Conditional image generation with PixelCNN decoders,” CoRR, (2016). abs/1606.05328

[7] Radford, A., Metz, L. and Chintala, S., “Unsupervised representation learning with deep convolutional generative adversarial networks,” (2015). arXiv preprint arXiv:1511.06434

[8] Makhzani, A. and Frey, B. J., “PixelGAN autoencoders,” CoRR, (2017). abs/1706.00531

[9] Yu, F., Koltun, V. and Funkhouser, T., “Dilated residual networks,” Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 472-480 (2017).