CBA-GAN: Cartoonization style transformation based on the convolutional attention module

Feng Zhang [1], Huihuang Zhao[1, 2], Yu-hua Li, Xiao-man Liang [1, 2]

1 Hunan Provincial Key Laboratory of Intelligent Information Processing and Application, Hengyang, 421002, China;

2 College of Computer Science and Technology, Hengyang Normal University, Hengyang, 421002, China;

3 School of Computer Science and Informatics, Cardiff University, Cardiff, UK;

Abstract: An Art form widely used, cartoonization has been integrated into every part of our life. Although cartoonization has made great progress, there is still a challenge to generate high-quality graphics. In this paper, a new model named CBA-GAN(Convolutional Block Attention Generative adversarial networks) to transform real photos into cartoonish images is proposed. The proposed method can multiply the feature graphs of the input image to achieve adaptive feature optimization, thus making images more cartoonish. At the same time, the proposed method does not produce redundant edges and can handle shadows in the image better. Experimental results on different types of images demonstrate that our method outperforms three representative methods from recent publications and has good robustness.

Keywords: Cartoonization, Convolutional Block Attention, Edge detections, Attention, Generative adversarial networks

1 Introduction

Cartoons are usually created by cartoonists through sketching [1-2], pencil drawing [3], or other artistic painting methods. Animation and comics are widely used as art forms in daily life. In recent years, in addition to the familiar Japanese manga, domestic animation and manga also developed rapidly. While good works in animation continue to emerge, they are relatively rare. The reason for this is that animation productivity is relatively low and production costs are relatively high. In order to save costs and speed up animation production, many comic companies have begun to develop intelligent algorithms to produce high-quality comics. This saves production costs and reduces the workload of cartoonists, allowing them to concentrate on designing. Existing software that are classic tools for processing graphic comics include Adobe After Effects (AE) [4-5] and Adobe Premiere Pro(PR) [6-7]. The popularity of these tools illustrates the importance of cartoonization of real photos and the necessity of developing cartoonization algorithms.

Although non-realistic rendering has been extensively studied in the field of style transfer, its algorithm has limited versatility. In recent years, deep learning-based style transfer methods have been sought after by many researchers, and image style transfer is one of their most representative research directions. CycleGAN created a training model for unpaired content images and style images and obtained high-quality migration results [8-9]. Subsequently, the Tsinghua University team proposed the CartoonGAN

model in 2018, which transformed real scenes into cartoon images based on a generative confrontation network. Using unpaired content images and style images as data sets, the cartoonization trend is set off [10-11]. In 2020, the white box Cartoonize model proposed by the University of Tokyo team in collaboration with ByteDance pushed image cartoonization research to its peak. On the basis of CartoonGAN, this model adds adjustable surface, texture, and structure loss to fine-tune the image cartoon effect. Combined with the Douyin APP, it allows users to automate the cartoonization of pictures [12]. Although the white-box cartoon has achieved great success, the training of this model lacks stability, and it is difficult to achieve the effect shown by the original author.

To solve the problems mentioned above, this article proposes a simple, effective, and novel method for cartoonization of real photos. Our method uses unsupervised learning, real-world photos, and cartoon images as the training set. The training set does not need to be paired. First, a generator is designed, which is based on the block attention module. A lightweight Convolutional Block Attention Module (CBAM) is added to the generator, which contains meaningful features in two dimensions of cross-channel and spatial axis [13]. To enhance the expressiveness of the module, attention is given to more important features in the image, and unimportant features are suppressed. In this manner, the style of the characteristic attributes of the image can be captured more accurately, thereby improving the effect of generating cartoon images. Second, three discriminators are designed according to the different attributes of images. They are used to distinguish between the different properties of generated art and cartoon art, forcing the generator to produce better art in the game. Finally, in the process of model training, we need to reconstruct the loss of the generator and the discriminator. The training loss includes fuzzy resistance

loss, texture resistance loss, edge resistance loss, and surface resistance loss of image generation and cartoon images.

To evaluate our method, data sets of Hayao style, Shinkai style, and Hosoda style were used as the style training set of landscape cartoon images, and data sets of Kyoto_face style and P. A. face style were used as the style training set of face cartoon images. Compared with the state-of-the-art cartoonization methods, in this paper, the method proposed has obvious advantages in the processing of color, texture, edge, and shadow in cartoonization of real photos. Finally, the ablation study of the key components of the model further shows that our CBA-GAN method is preferred in feature extraction and edge preservation of real images, compared to several advanced cartoonization methods.

To summarize, the key contributions of this paper are:

A novel CBA-GAN model is proposed in which unpaired real photos and cartoon images are used as training sets. The method uses an attention mechanism to increase expressive power and pay attention to important features to suppress unnecessary features. It can capture the style on which image feature attributes depend accurately, and generate high-quality cartoon images.

An improved boxed model based on the block attention module is proposed. In the generator, in order to distinguish the importance of image features, a U-shaped network based on a lightweight convolutional attention module is designed, which can pay attention to important features and ignore unnecessary features in a two-dimensional space across channels and spaces.

At the same time, to solve the problem of image pixel overflow during the model training process, a Tanh

activation function is added to the generator to normalize the pixel value to [-1,1]. To keep the sharp edges in the discriminator network, an edge discriminator is designed to distinguish the edges of the image. It promotes the adversarial loss of the edges, and keeps the sharpness of the edges. Finally, ablation studies on key components of the model further demonstrate that our method is more effective than existing methods and is more effective for cartoonization of real photos.

2 Related Work

Image cartoon processing. Drawing real-world cartoons is a unique craft that is in high demand. In recent years, the generation and processing of cartoon images have been widely studied. The creation of cartoon images includes many aspects such as contour, structure, and texture. It can be seen that halftone texture has become a unique problem of cartoon image texture optimization. Halftone refers to the grayscale, which refers to the tone value of the picture expressed by the size or density of the dots. Using the spatial integration of human vision, black and white pixels approximate the intensity of local small areas. According to the steps of drawing contour lines in the process of cartoon drawing, a path-based image cartooning method is proposed, which reduces the artifacts in the course of drawing by adjusting the scanning path of the image. Based on the path-based method [14-16], Knuth et al. proposed to add edge detection in the pre-training process to maintain the edge of the image[17]. Later, Buchanan et al. proposed a method of optimizing structural similarity to retain the detailed structure of images [18-19]. A semantic segmentation method for cartoon images based on visual perception is proposed by Noh et al. [20]. This method maps the brightness of pixels or regions to the overall screen mode to achieve a perceptually distinguishable and prominent screen tone structure. The learning-based method trains the

network model through an end-to-end approach to predict the texture of the sketch. The variational autoencoder maps the input cartoon to a potential space and describes the probability of the cartoon in the potential space in a probabilistic manner [21]. Although existing methods have achieved success in dealing with the texture aspects of cartoon drawings, they focus on one aspect of the cartoon creation process. There is still no satisfactory result in the quality of transforming ordinary images into cartoons.

Style transfer. The traditional image style transfer is mainly based on the synthesis of the rendering texture of the physical model [22]. The image quilting method forms a new texture by stitching existing small images into new images and is based on the image style transfer of the physical model [23-24]. Subsequently, researchers proposed a sample image-based method. The fast style transfer method, which improves the Ashikhmin algorithm, establishes a domain consistency measurement technique and improves the efficiency of image pixel matching [25-26]. Although these methods have obtained relatively good results, they can only extract the underlying features of the image, and cannot extract the deep features. When the texture and color of the image are more complex, the resulting image synthesis effect is relatively rough.

With the development of deep learning, its application in style transfer has also developed rapidly. For example, Gatys et al. proposed a creative method of image style transfer based on a neural network. This method transfers the style of the style image to the content image and generates a new image. They use the VGG network for pre-training, express the extracted feature maps as content images, and continuously optimize new images so they not only have the content of the content image but also match the texture information of the style image [27]. However, this method belongs to supervised learning and

requires one-to-one correspondence between content images and style images. When the image contains complex objects, the style may be transferred to areas with different semantics, resulting in misalignment. For cartoon style transfer, edges and shadows are indispensable elements, and the method proposed by Gatys et al. cannot maintain edges and shadows well. Subsequently, Li et al. proposed a method combining the Markov random field model and deep convolutional neural network with discriminator training. This method divides the image feature map into several regions and matches them to achieve image style transfer, but its partial matching is prone to error, resulting in new images with semantic errors [28].

Generative adversarial networks and style transfer. Since it was proposed in 2014, the generative adversarial network has been sought after by a large number of researchers. It has shown strong advantages in image generation and is widely used in image-to-image translation [29], image carbonization [30], face generation [31], etc. Li et al. combined the Markov random field with the generative adversarial network to conduct adversarial training on the generated model and obtain a generated image with better quality. However, this method requires paired data sets, which are difficult to find for cartoon images [32]. Since then, researchers have proposed a series of unsupervised generative adversarial networks, such as CycleGAN [33], DCGAN [34], DualGAN [35], etc. Yet, these methods are based on the iterative optimization of image divergence distribution to carry out antagonistic training, and the process of image style transfer cannot be controlled.

3 Method

According to the original Generative Adversarial Nets definition [27], a mapping function is defined to

describe the process of converting a real image to a cartoon image, which is the working process of a generator . Through this mapping function, the random Gaussian noise variable corresponding to the real image is mapped to the Gaussian distribution corresponding to the cartoon image. The training data set used by the mapping function is as follows. The real image represents the number of real images in the training set and the cartoon image represents the number of cartoon images in the training set. The discriminator D is used to distinguish whether the peak value of the image distribution is close to the peak value of the Gaussian distribution of the generated image or of the cartoon image. At the same time, the antagonism loss of the two images is output to promote the continuous proximity of the peak value of the Gaussian distribution of the generated image and of the cartoon image, in order to achieve the purpose of optimizing the generator G.

As shown in figure 1, the framework is outlined in this article, where the input is a real photo $X \in R^{w \times h \times c}$ and the output is a caricature image $Y \in R^{w \times h \times c}$, w, h, c indicating the width, height, and the number of channels of the image, respectively. A model called CBA-GAN is designed, which is based on CartoonGAN and CBAM. The model consists of a generator and three discriminators, D_s , D_t and D_e, D_s which are used to distinguish generated images x' and cartoon images y after denoising and smoothing, and promoting the adversarial loss of image smoothing to suppress image noise. D_t, it is used to distinguish the generated image and cartoon image after graying, facilitates adversarial training of image textures to preserve sharp textures, and D_e is used to distinguish the edges of generated image and cartoon image to keep the clear edge.

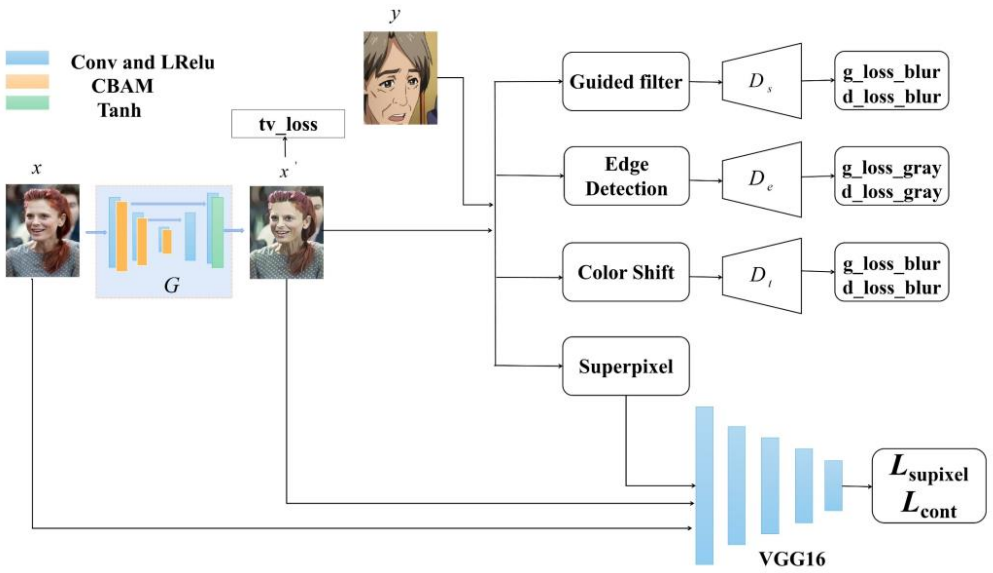


Fig.1 Schematic diagram of the model architecture

4 Models

4.1 CBA-GAN

Our model is mainly composed of a generator G and three discriminators D_s, D_t and D_e. The generator G of the model comprises a network similar to the UNet structure and the light weight attention convolution module, CBAM, whose main function is to map the random noise z corresponding to real photos x to the distribution of cartoon images y. Through continuous iterative optimization, the distribution of random noise z will continuously learn to fit the distribution of cartoon image y, to ensure the generated image has a cartoonish style. In terms of the structure of the generator, a U-shaped full convolutional network is used for image generation, which is composed of the encoder, residual network, and decoder. Encoder, each layer contains a convolution layer, an activation layer, a CBAM layer, and a down-sampling layer, among which the last layer has no down-sampling layer. The encoding and compression of image features are realized through the above operations. Residual network, which uses four residual network blocks with the same structure. It can learn the expected mapping more easily and ensure the image gradient. Each layer of the decoder consists

of a convolution layer, an activation layer, and an upsampling layer. Similarly, the last layer of the decoder also has no upsampling layer. Through the above operations, image features are reconstructed and images with cartoon-style are obtained. The structure of the CBA-GAN is shown in figure 2 below.

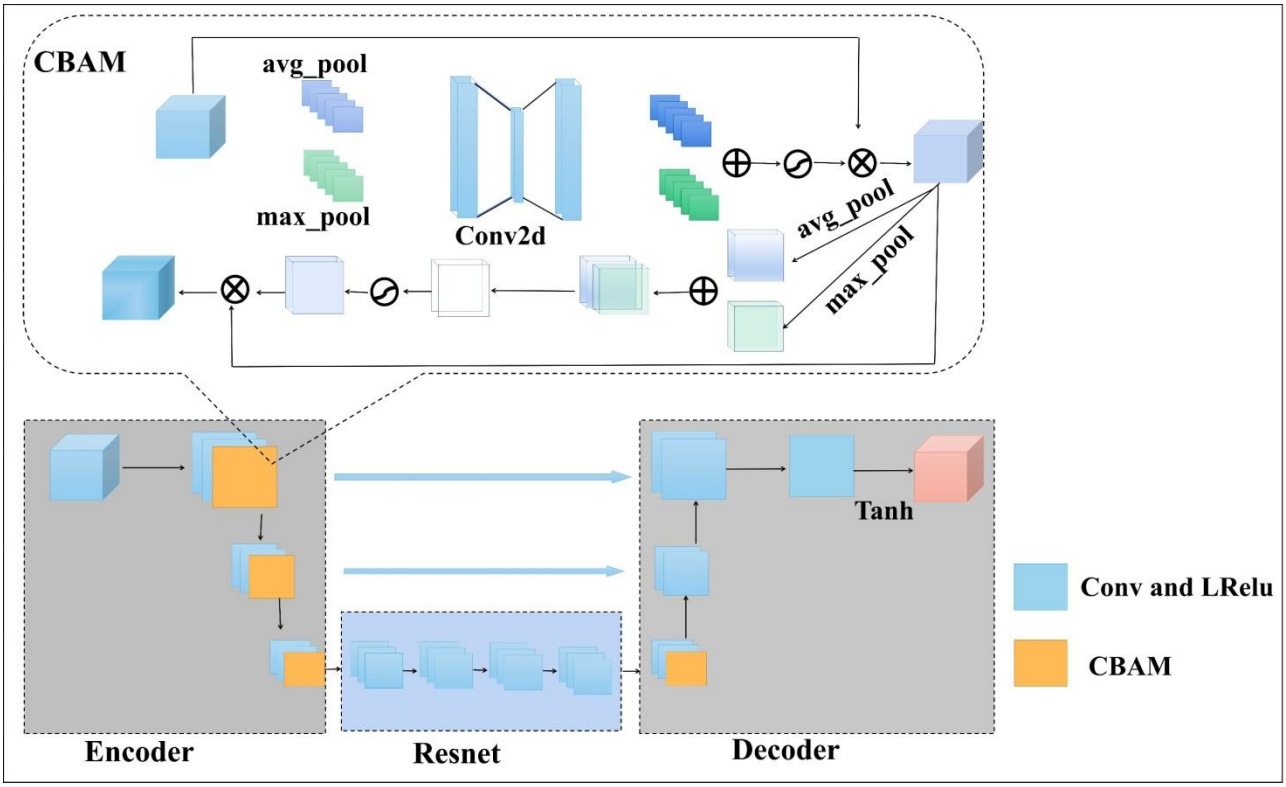


Fig. 2 The specific generator structure of CBA-GAN.

Pixel overflow is prone to occur in the image training process, which leads to spots in the training image. Therefore, the hyperbolic tangent activation function is added to the image generated by the output of our previous generator. This normalizes the pixel value of the function image (1,1) and effectively prevents pixel overflow. Further, it can also improve both the nonlinear neural network model and the training effect of the image. The specific formula of the Tanh activation function is shown in equation 1.

$$\tanh(x) = g(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{1}$$

This is an odd function that goes through the origin and traverses quadrants I and III. The value range is (-

1, 1). Using this as the activation function of the last layer of generator G can effectively solve the problem of image pixel overflow.

As an important part of the generative adversarial network, discriminator D is mainly used to distinguish whether the image comes from the distribution of generated images or the distribution of real images. Compared with discriminators that only perform classification, the discriminator used in this paper requires an auxiliary generator to generate cartoon-like images. The generation of cartoon-like images depends on the local features of the images, so it is very important to design a suitable discriminator. Therefore, PatchGAN, which has fewer parameters and is fast in operation, is used as the discriminator. The image is divided into several patches with a size of 32×32, and each patch is judged to be true or false. Finally, the image is returned to the generator for adversarial training.

4.2 Loss Function

Figure 1 above shows the network framework proposed by us, and its loss function consists of four parts: (1) Fuzzy antagonistic loss $L_{blur}(G, D_s)$, which is derived from the use of discriminator D_s to make the image generated by generator G suffer less noise interference and maintain the gradient of edge; (2) Texture antagonism loss $L_{gray}(G, D_t)$, which mainly avoids color and other unnecessary factors affecting the migration of image texture and ensures that the model only considers image texture; (3) Edge antagonism loss $L_{edge}(G, D_e)$, which preserves clear edges for images in the process of cartoonization; and (4) content loss $L_{cont}(G_i, G_o)$, which ensures the retention of image content in the process of cartoonization.

Therefore, the loss function is defined as:

$$L_{total}(G, D) = \alpha L_{blur}(G, D_s) + \beta L_{gray}(G, D_t) + \lambda L_{edge}(G, D_e) + \omega L_{cont}(G_i, G_o) \quad (2)$$

In this experiment, the parameters of the loss function of each part of the generator are : $\alpha = 0.1$, $\beta = 1$, $\lambda = 0.1$, $\omega = 200$. The parameters of the loss function of each part of the discriminator are : $\alpha = 1$, $\beta = 1$, $\lambda = 0.1$, $\omega = 0$. This setup better balances the ratio of cartoon characters to real photos during the style transition, thus resulting in a better cartoon image.

(1) Fuzzy antagonistic loss $L_{blur}(G, D_s)$ imitates the sketch created by the artist in the early stage of creation, without excess noise, and the edge of the image is relatively clear. In this experiment, the image is processed by guided filtering to preserve the edge and gradient of the image, represented as H_{gf} . The method takes image I_i as the input and itself as the guide graph, and the resulting image is a fuzzy image with noise, texture, and details removed. The discriminator D_s is used to determine whether the noiseless edge-preserving graph is derived from the generated image or cartoon image, and it is fed back to the generator G to learn the useful information after blurring. Where I_o represents cartoon image, $G(I_i)$ represents image generated by G , and represents guided filtering operation.

$$L_{blur}(G, D_s) = \log D_s(H_{gf}(I_o, I_o)) + \log(1 - D_s(H_{gf}(G(I_i), G(I_i)))) \tag{3}$$

(2) Texture antagonistic loss $L_{gray}(G, D_t)$, is used to avoid unnecessary factors such as color and light affecting the high-frequency features of the image extracted from the model, and the single-channel texture is used to represent the image.

$$H_{cs}(I_{rgb}) = (1 - \varphi)(\theta_1 \times I_r + \theta_2 \times I_g + \theta_3 \times I_b) + \mu \times S \tag{4}$$

Where, H_{cs} represents image grayscale operation, I_r , I_g and I_b represents three color channels, and S represents the standard grayscale image of RGB color image conversion.

$$L_{gray}(G, D_t) = \log D_t(H_{cs}(I_o, I_o)) + \log(1 - D_t(H_{cs}(G(I_i)))) \tag{5}$$

(3) The content loss $L_{cont}(G_i, G_o)$ is calculated in the feature space of VGG19 after superpixel segmentation to ensure that the content of the generated image is unchanged in structure. Meanwhile, the local features are combed through the coefficient control of L_1 norm.

$$L_{cont}(G_i, G_o) = \| VGG_n(G(I_i)) - VGG_n(G(I_o)) \| \tag{6}$$

(4) Edge extraction loss $L_{edge}(G, D_e)$ uses the convolution operation imitating Sobel operator to extract the edges of generated images and cartoon images respectively, then input the extracted edges into the discriminator D_e, and finally calculate the $L_{edge}(G, D_e)$ between the two edge images.

$$L_{edge}(G, D_e) = \log D_e(H_{eg}(I_o)) + \log(1 - D_e(H_{eg}(G(I_i)))) \tag{7}$$

Where represents an edge extraction operation, I_o represents a cartoon image, and I_i represents a generated image.

5. Experiments

It is a model implemented in Tensorflow and Python. The proposed CBA-GAN model can provide a comparative experiment for future photo-style cartoonization. All experiments are performed on NVIDIA 1080 GPU.

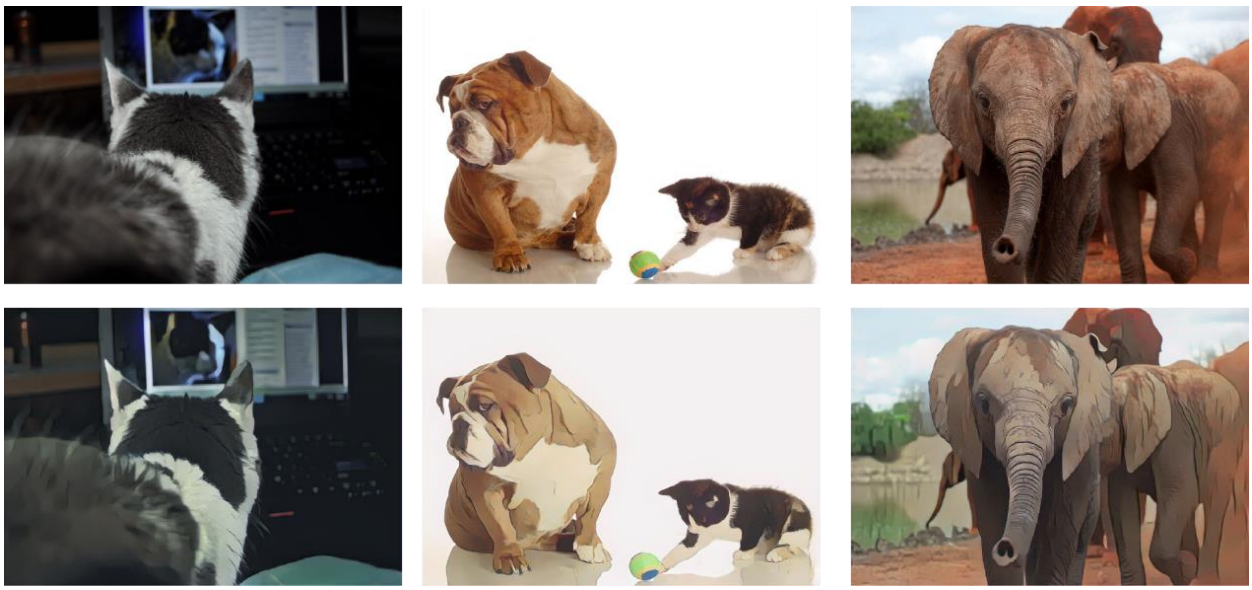
CBA-GAN can use the artist's cartoon collection as stylistic data. And the stylistic data can be generated through training, and turn real photos into cartoons. Since cartoon-style data sets come from various cartoon videos, it is difficult to obtain paired images. Therefore, our model uses unsupervised learning to learn unpaired data. Adam algorithm was used for optimization during model training, with the learning rate at 0.2 and the batch size was set to 9. We carried out pre-training 50,000 times to train the generator G, and then set 60,000 iterations to train the whole model. Generally, the convergence stopped around 40000 times.

5.1 Partial experimental results of our method

This paper has carried out cartoon-style experiments on images of faces, food, animals, and landscapes, as well as other images, and obtained good results. Some of the results generated by CBA-GAN are shown in figure 3.

(a) Person (b) Food

(c) Scenery (d) City



)e) Animals

Fig. 3 Part of the experimental results, the original images on the first row and the results on the second row. Zoom in for details.

It can be seen that our method clearly preserves the edges, textures, and colors of the real image, and the shadow part of the output image is essentially the same as the real image.

5.2 Comparison with different methods

Real photos were randomly selected for testing and the results obtained by comparing the test structures are shown in figure 4. We compare CBA-GAN with the recent White-box cartoon and AnimeGAN methods based on GAN. We randomly selected the input real photos for testing and compared the test results to obtain the qualitative results shown in figure 4.

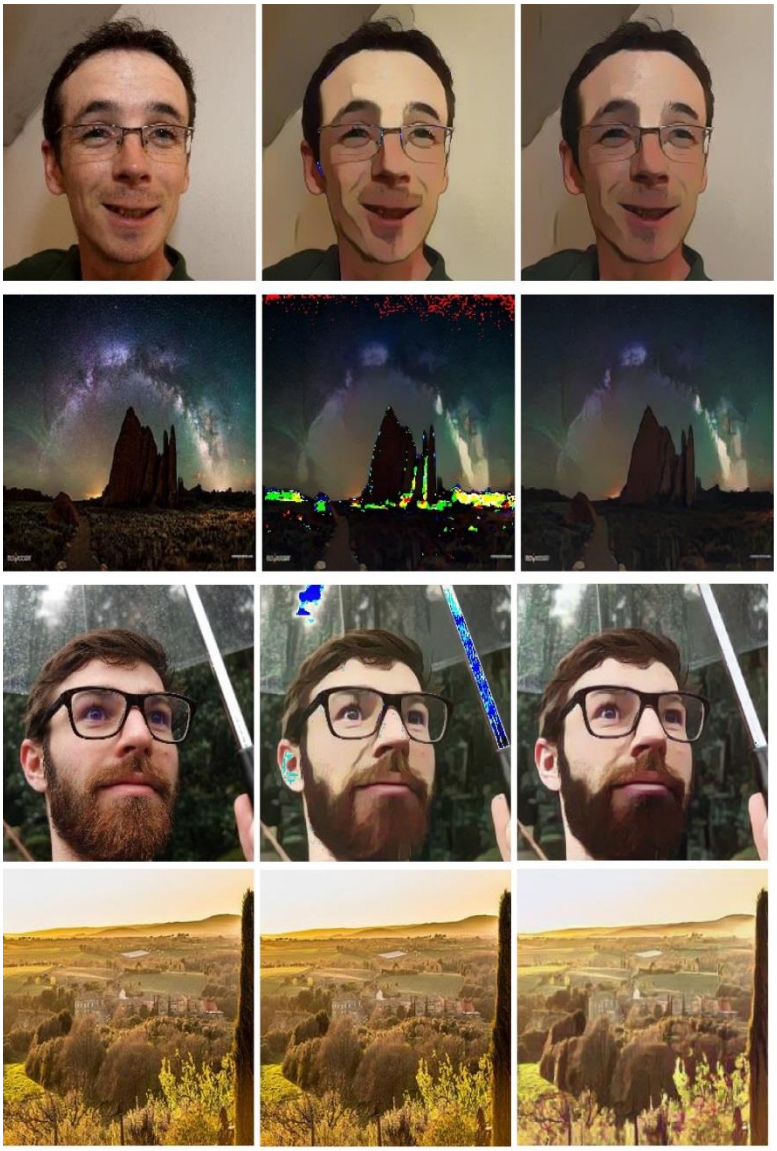
Fig. 4 Experimental results comparison. (a)The original picture, (b)White-box cartoon, (c)AnimeGAN_Hayao style, (d)AnimeGAN_Paprika style, (e)Our method. Zoom in for details.

Figure 4 shows the qualitative results. It can be seen from the figure 4 that White-box cartoons and AnimeGAN cannot handle the cartoonization of image styles well (figures 4b-4d). Our method can realize adaptive feature optimization and enhance the semantic connection between various parts of the image. At the same time, this method can handle image shadows and edges reasonably without generating redundant edge lines.

5.3 Ablation studies

Some results of the ablation of the Tanh function has been shown in figure 5. Ablation of the Tanh function leads to pixel overflow in the image during training. As shown in figure 5, some overflowing pixels can be seen on the image. This is due to the pixel value of the generated image being too large and therefore out of bounds, after the convolution operation of the encoder and decoder. For comparison, four images are selected randomly after 30000 iterations. Each group contains face images and landscape images. It can clearly see that

after adding the Tanh activation function, the overflow of image pixels is basically gone.



(a) (b) (c)

Fig. 5 Ablation is studied by removing the Tanh function. (a)Input image, (b)Generated image without tanh function, (c)We added generated image with tanh function. Zoom in for details.

Funding Statement: This work was supported by National Natural Science Foundation of China (61772179), Hunan Provincial Natural Science Foundation of China(2020JJ4152), the Science and Technology Plan Project of Hunan Province (2016TP1020).

Conflicts of Interest: We declare that we have no conflicts of interest to report regarding the present

study.

References

[1] Yang M., Lin S., Luo P., et al. Semantics-driven portrait cartoon stylization. 2010 IEEE International Conference on Image Processing, 2010: 1805-1808

[2] Lu H., Zhang M., Xu X., et al. Deep fuzzy hashing network for efficient image retrieval. IEEE transactions on fuzzy systems, 2020, 29(1): 166-176

[3] Wang C., Zhang J., Yang B., et al. Sketch2Cartoon: composing cartoon images by sketching. Proceedings of the 19th ACM international conference on Multimedia, 2011: 789-790

[4] Sofyan A. F., Purwanto A.. Digital Multimedia: Animasi, Sound Editing, & Video Editing: Contoh Kasus dengan Adobe After Effects, Adobe Soundbooth, dan Adobe Premiere Pro. Penerbit Andi, 2020

[5] Lu H., Zhang Y., Li Y., et al. User-oriented virtual mobile network resource management for vehicle communications. IEEE transactions on intelligent transportation systems, 2020, 22(6): 3521-3532

[6] Epifanovsky E., Gilbert A. T. B., Feng X., et al. Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. The Journal of chemical physics, 2021, 155(8): 084801

[7] Wang P., Wang D., Zhang X., et al. Numerical and experimental study on the maneuverability of an active propeller control based wave glider. Applied Ocean Research, 2020, 104: 102369

[8] Zhu J. Y., Park T., Isola P., et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, 2017: 2223-2232

[9] Zheng Q., Zhu J., Tang H., et al. Generalized Label Enhancement with Sample Correlations. IEEE Transactions on Knowledge and Data Engineering, 2022(Early Access)

[10] Chen Y., Lai Y. K., Liu Y. J.. Cartoongan: Generative adversarial networks for photo cartoonization. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 9465-9474

[11] Ma C., Li X., Li Y., et al. Visual information processing for deep-sea visual monitoring system. Cognitive Robotics, 2021, 1: 3-11

[12] Wang X., Yu J.. Learning to cartoonize using white-box cartoon representations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8090-8099

[13] Woo S., Park J., Lee J. Y., et al. Cbam: Convolutional block attention module. Proceedings of the European conference on computer vision (ECCV), 2018: 3-19

[14] Velho L., Gomes J.. Stochastic screening dithering with adaptive clustering. Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, 1995: 273-276

[15] Naiman A. C., Lam D. T. W.. Error Diffusion: Wavefront Traversal & Contrast Considerations. Graphics Interface, 1996, 96: 78-86

[16] Nakayama Y., Lu H., Li Y., et al. WideSegNeXt: semantic image segmentation using wide residual network and NeXt dilated unit. IEEE Sensors Journal, 2020, 21(10): 11427-11434

[17] Knuth D. E.. Digital halftones by dot diffusion. ACM Transactions on Graphics (TOG), 1987,6(4): 245-273

[18] Buchanan J. W., Verevka O.. Edge preservation with space-filling curve half-toning. Graphics Interface. CANADIAN INFORMATION PROCESSING SOCIETY, 1995, 75-75

[19] Kang S., Wu H., Yang X., et al. Discrete-Time Predictive Sliding Mode Control for a Constrained Parallel Micropositioning Piezostage. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 52(5): 3025-3036

[20] 20. Noh H., Hong S., Han B.. Learning deconvolution network for semantic segmentation. Proceedings of the IEEE international conference on computer vision, 2015: 1520-1528

[21] Liu Z. S., Siu W. C., Wang L. W.. Variational AutoEncoder for Reference based Image Super-Resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 516-525

[22] Meyer D. J., Wiertlewski M., Peshkin M. A., et al. Dynamics of ultrasonic and electrostatic friction modulation for rendering texture on haptic surfaces. 2014 IEEE Haptics Symposium (HAPTICS). IEEE, 2014: 63-67

[23] Efros A. A., Freeman W. T.. Image quilting for texture synthesis and transfer. Proceedings of the 28th annual conference on Computer graphics and interactive techniques, 2001: 341-346

[24] Lu H., Tang Y., Sun Y.. DRRS-BC: Decentralized routing registration system based on blockchain. IEEE/CAA Journal of Automatica Sinica, 2021, 8(12): 1868-1876

[25] Qian X. Y., Xiao L., Wu H. Z.. Fast style transfer Computer Engineering, 2006, 32(21): 15-17

[26] Sun J., Li Y.. Multi-feature fusion network for road scene semantic segmentation. Computers & Electrical Engineering, 2021, 92: 107155

[27] Goodfellow I. J., Pouget-Abadie J., Mirza M., et al. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014: 2672-2680

[28] Li C., Wand M., Combining Markov random fields and convolutional neural networks for image synthesis. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 2479-2486

[29] Isola P., Zhu J. Y., Zhou T., et al. Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1125-1134

[30] Armand E., Bajgiran O. S., Rahman S., et al. A multi-objective model for order cartonization and fulfillment center assignment in the e-tail/retail industry. Transportation Research Part E: Logistics and Transportation Review, 2018, 115: 16-34

[31] Makrushin A., Wolf A.. An overview of recent advances in assessing and mitigating the face morphing attack. 2018 26th European Signal Processing Conference (EUSIPCO), 2018: 1017-1021

[32] Li C., Wand M.. Precomputed real-time texture synthesis with markovian generative adversarial networks. European conference on computer vision, 2016: 702-716

[33] Zhu J. Y., Park T., Isola P., et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE international conference on computer vision, 2017: 2223-2232

[34] Fang W., Zhang F., Sheng V. S., et al. A method for improving CNN-based image recognition using DCGAN. Computers, Materials, and Continua, 2018, 57(1): 167-178

[35] Yi Z., Zhang H., Tan P., et al. Dualgan: Unsupervised dual learning for image-to-image translation. Proceedings of the IEEE international conference on computer vision, 2017: 2849-2857

Biographies of Each Co-author

Feng Zhang received his Master degree in 2022 from Hengyang Normal University. Now she is an Assistant Professor in the College of Computer Science and Technology, Hengyang Normal University. Her research interests include image style transfer and video image style transfer.

Hui-Huang Zhao received his Ph.D. degree in 2010 from XiDian University. He was a Sponsored Researcher in the School of Computer Science and Informatics, Cardiff University. Now he is a Professor in the College of Computer Science and Technology, Hengyang Normal University. His main research interests include image style transfer, compressive sensing, machine learning, and image processing.

Yu-Hua Li received the Ph.D. degree in general engineering from the University of Leicester. He is currently a Senior Lecturer with the School of Computer Science and Informatics, Cardiff University. His current research interests include machine learning, pattern recognition, data science and semantic similarity analysis. His research projects have been funded by government, charity and industry.

Research Highlights

1. A novel CBA-GAN model is proposed in which unpaired real photos and cartoon images are used as training sets. The method uses an attention mechanism to increase expressive power and pay attention to important features to suppress unnecessary features. It can capture the style on which image feature attributes depend accurately, and generate high-quality cartoon images.
2. An improved boxed model based on the block attention module is proposed. In the generator, in order to distinguish the importance of image features, a U-shaped network based on a lightweight convolutional attention module is designed, which can pay attention to important features and ignore unnecessary features in a two-dimensional space across channels and spaces.
3. To solve the problem of image pixel overflow during the model training process, a Tanh activation function is added to the generator to normalize the pixel value to [-1,1]. To keep the sharp edges in the discriminator network, an edge discriminator is designed to distinguish the edges of the image. It promotes the adversarial loss of the edges, and keeps the sharpness of the edges. Finally, ablation studies on key components of the model further demonstrate that our method is more effective than existing methods and is more effective for cartoonization of real photos.

Declaration of interests

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.