Adaptive Convolution Kernel for Painterly Image Harmonization

Xiao Zhang[1,a*],Yun Jiang[1,b], Shanshan Wang[1,c]

[1]School of Artificial Intelligence, Chongqing Technology and Business University, Chongqing, China

[a*]Corresponding author e-mail:zhangxiao@ctbu.edu.cn
[b]e-mail:jiangyun@email.ctbu.edu.cn
[c]e-mail:wangshanexpression@163.com

Abstract. Image synthesis is a technique for cutting out foreground images and pasting them onto another image. To make the synthesized image harmonious, it is necessary to adjust the color and light of the foreground image so that it blends smoothly with the background image and has a consistent color. When the background image is an artistic style image, this process is called painterly image harmonization. This method can adjust the style of the foreground image so that it is compatible with the background image, producing a visually harmonious composite image while preserving the content of the foreground image. Existing painterly image harmonization methods have relied heavily on AdaIN methods, which often focus on color harmony but ignore the local structure information in the style image. We propose a new method based on dynamic convolution kernel for painterly image harmonization, which can dynamically generate convolution kernel to accommodate different style images during inference. Our method can effectively perceive the spatial structural elements of style images and generate more aesthetically pleasing composite images than AdaIN-based painterly image harmonization methods.

Keywords: image synthesis, image harmonization, painterly image harmonization

1 Introduction

Image synthesis refers to the technique of cutting out foreground from one image and pasting it onto another background image, producing a synthesized image that maintains the harmony and consistency of the image and semantics. The synthesized image generated by simply cutting and pasting the foreground onto the background often has obvious editing traces, and the color, lighting, and semantics of the foreground and background images may not be consistent, making the overall composite image unharmonious and unrealistic. There are some traditional or deep learning-based image synthesis methods that address the image synthesis problem from the perspectives of scene harmony, shadow synthesis, and object placement [1].

R. Appleby et al. (eds.), *Proceedings of the 2nd International Conference on Intelligent Design and Innovative Technology (ICIDIT 2023)*, Atlantis Highlights in Intelligent Systems 10,
https://doi.org/10.2991/978-94-6463-266-8_52

When the background image is a painterly image, image harmonization is also called painterly image harmonization. Integrating painterly images from the real-world image is a difficult task. Due to the sensitivity and subjective nature of human visual system, this synthesized image is easily recognized by humans. These inharmonious elements mainly come from the inconsistency between the style of the foreground and the background and from the sudden change of color gradient at the boundary between the foreground and the background. Traditional image harmonization methods adjust the brightness, saturation, and contrast of the foreground image attributes to make the colors in the painting more harmonious and natural, and the boundary between the foreground and background becomes more natural. However, it is very difficult to adjust the style content of the foreground content, and the differences between different artistic styles are difficult to quantify. The original content is easily destroyed during the style adjustment.

In recent years, with the development of deep learning and style transfer, methods based on Gram matrix [2] and feature map statistics [3] have been proposed, which are widely used in the field of painterly image harmonization. Painterly image harmonization methods can be mainly divided into optimization-based methods [2] and feedforward neural network methods [3,4]. Optimization-based methods set appropriate loss functions, and iteratively update the output image from noise images. Optimization-based methods can usually generate very good artistic effects, but they are slow and difficult to apply. Neural network-based methods require a large amount of artistic image training, Cao et al. [4] perform harmonization operations in the frequency domain, and Yan et al. [5] improve image quality from multiple scales. These methods are slow to train, but once trained, they generate synthesized images very quickly, making them the fastest growing class of methods.

Currently, existing methods focus on transferring the style of the image, and Huang [3] proposes the AdaIN method, which is widely used in the field of real-time style transfer. This method achieves style transfer by aligning the feature maps of the style image and the content image in terms of feature map statistics. However, these transformation methods focus on color style alignment between the content image and the style image, often ignoring the geometric element information in the style image. As shown in Fig. 1, during the image synthesis process, the difference between the lines of the foreground image and the background image can be easily distinguished by the style image with a large difference. To solve these problems, we propose a new method to improve the quality of the generated style image by AdaIN.

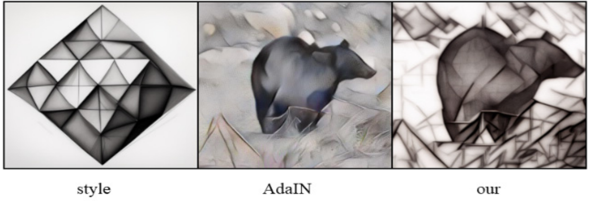

Fig. 1. Applying AdaIN and our method to the entire image, our method generates images that include more local structure information from the style image.

Our model is based on the encoder-decoder architecture. To better capture the spatial local information in the style image, we designed a simple kernel prediction net that can generate a series of convolution kernels from the feature map of the style image. These convolution kernels form the decoder, which decodes the synthesized image from the feature map of the content image. Experimental results show that our method can better capture the structural information of the feature map than existing methods.

2 Related Work

2.1 Image Harmonization and Painterly Image Harmonization

Traditional image harmonization methods often rely on various statistical information or textures in the image, including color gradient matching or multi-scale statistics, Song et al. [6] used the information of image grayscale mean scale to estimate scale, and adjusted the lighting information, ultimately achieving lighting coordination. Some deep learning-based image harmonization methods use paired data for learning, this type of method uses datasets generated by real image editing, and the dataset may be different from the real pasted image. When the background image is a painterly image, the foreground image needs to be transformed into the same style as the background image, while maintaining coordination. This type of work is relatively rare. Luan et al. [7] transferred the related statistical data of the neural network to the inserted object and introduced multiple losses to ensure spatial and scale consistency. Zhang et al. [8] used a new Poisson gradient calculation method and introduced style loss into the Poisson gradient loss. Peng et al. used AdaIN and GAN methods, changed the feature map statistics of the content image to be the same as the feature map of the background image, and used local discriminators for adversarial learning. These methods only consider the transfer of style in the color feature map, but do not consider the transfer of geometric information in the style image.

2.2 Kernel prediction

Kernel prediction refers to the technology in which the model dynamically generates convolution kernels during inference. Unlike traditional methods that are trained directly, dynamic generation of convolution kernels has better scalability and can avoid overfitting to a certain extent, but it is more difficult to train. Chen [9] et al. proposed a method for style transfer using dynamic convolution kernels, but this method can only transfer styles within a limited range. Niklaus et al. [10] extended the kernel prediction method to the video field. Xue et al. [11] used CNN to generate dynamic kernels and applied them to video frame synthesis.

3 Method

Our method consists of an encoder-decoder structure and is trained using a generative adversarial method. The model structure is shown in Fig. 2.

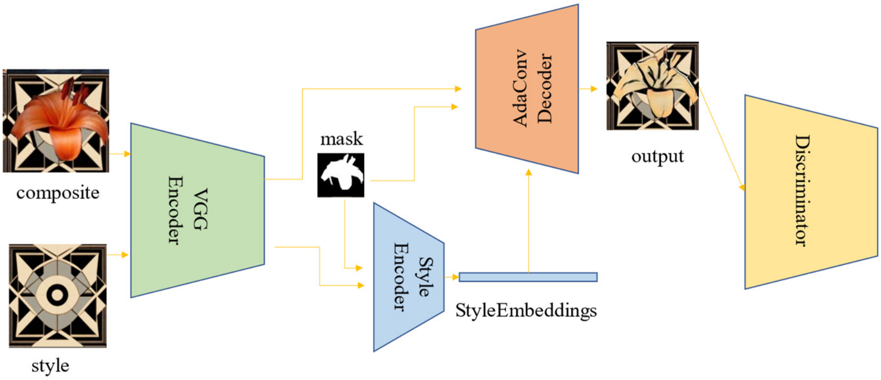

Fig. 2. Model overview figure

The style image and content image are respectively input into a fixed VGG network to obtain their feature map embeddings. We train the decoder and style encoder, and combine the feature maps of the two to generate the synthesized image. The decoder and discriminator are adversarially trained to make the output of the decoder more harmonious.

The encoder part uses a pre-trained VGG-19 on image classification tasks. The encoder outputs the feature maps of the style image and the synthesized image. Unlike other image harmonization or style transfer methods, in order to achieve better stylization effects, we only use the feature maps of relu4_1 layer. Compared with AdaIN, which uses more shallow feature maps, this processing method makes the output image less affected by the content of the original image. In the training process, the parameters of the encoder part are kept fixed, which makes the training faster and the output more stable. After being processed by the encoder, the synthesized image and the style image output a feature vector of (512,32,32) dimensions.

The feature map of the style image and the reduced to (32,32) size mask are concatenated and input into StyleEncoder. StyleEncoder consists of three convolutional layers and pooling layers, and the output of the last convolutional layer is mapped to a 512-dimensional vector by a fully connected layer. In this case, the mask acts as a guide for the model.

AdaConvDecoder module takes the feature map of the synthesized image and the reduced mask as input, and dynamically generates convolution kernels in the StyleEnbeddings module. The decoder consists of three layers, each of which has a size of twice that of the previous layer, and each part is composed of several dynamic convolution layers and up-sampling layers. Each dynamic convolution layer takes StyleEmbedding as input and generates convolution kernel weights and biases through a fully connected layer. Finally, the weights and biases are combined to form a new convolution kernel. The up-sampling layer uses the nearest neighbor sampling to avoid the chessboard effect. The output of the decoder is finally input into the discriminator. The discriminator will try to distinguish which areas are harmonious. The discriminator and decoder adversarially train to improve the discriminator's ability to generate harmonious areas. Style loss and content loss are defined by Formulas 2 and 3; The total loss is defined by Formula 1:

$$L = L_c + \lambda L_s \tag{1}$$

$$L_c = \left\| E(I_f) - E(I_h) \right\|_2 \tag{2}$$

$$L_s = \sum_{i=1}^{L} \|\mu(E_i(I_s)) - \mu(E_i(I_h))\|_2 + \sum_{i=1}^{L} \|\sigma(E_i(I_s)) - \sigma(E_i(I_h))\|_2 \tag{3}$$

where λ is a hyperparameter that controls the ratio of content loss and style loss; L_c is the content loss; L_s is the style loss; I_f is the foreground image; I_h is the harmonized image; I_s is the background image; μ is the mean function; σ is the variance function; E is the VGG encoder; E_i is the output of each layer of the encoder.

4 Experiments

We use the COCO dataset as the foreground image and WikiArt as the background image to conduct experiments. The experimental platform is Pytorch 2.0+RTX 4080. We use PHDNet [4] and AdaIN as our comparison methods. Fig. 3 shows the implementation results of our method and other advanced methods. Compared with other methods, our method is more likely to use the existing elements in the style image when generating images. This strategy makes our method better at integrating the features of the style image into the image content.

Fig. 4 shows more images generated by our method. Compared with other methods, the advantage of our method is that it protects the content of the original image and fully utilizes the features of the style image. By avoiding excessive color changes and unnecessary information loss, our method can generate more natural and stylized images.

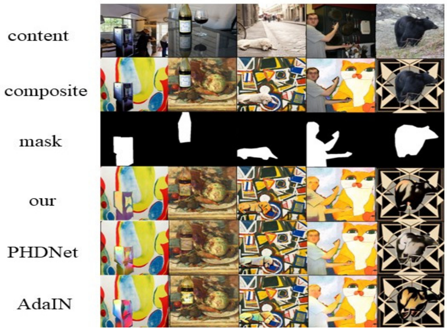

Fig. 3. Compared with other advanced methods

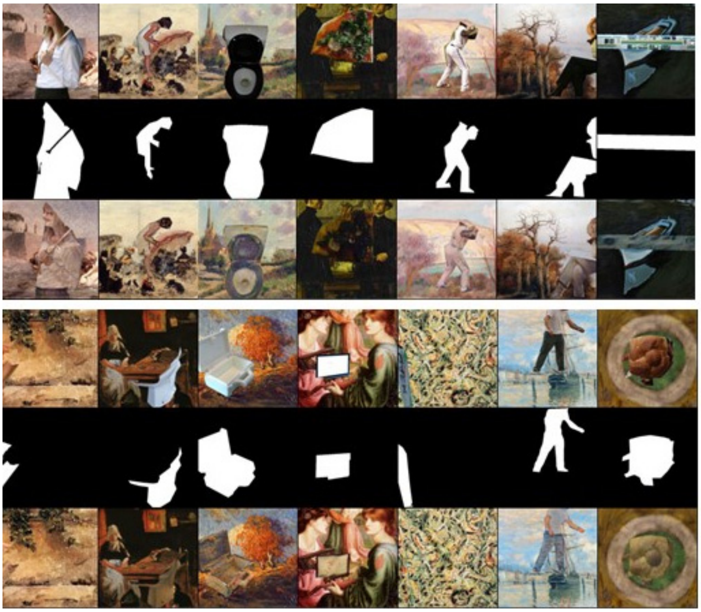

Fig. 4. Other painterly image harmonization results

5 Conclusion

Our model dynamically generates the convolutional kernel of the decoder during the inference process, which enables the model to pay more attention to the elements of the style image during style transformation. This approach can increase the credibility of the synthesized image. In addition, our model uses generative adversarial methods to focus on the modified regions of the image, which makes the generated image more harmonious. In the field of image synthesis, our research provides new ideas and methods for developing more efficient and accurate image synthesis technology. In addition, our research results also provide useful references and applications in the field of art design and other related fields.

Acknowledgment

This work was supported in part by Natural Science Foundation Project of CQ CSTC (cstc2018jcyjAX0057), Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN201800814).

References

1. Niu, L., Cong, W., Liu, L., Hong, Y., Zhang, B., Liang, J., & Zhang, L. (2021) Making images real again: A comprehensive survey on deep image composition. https://doi.org/10.48550/arXiv.2106.14490
2. Gatys, L.A., Ecker, A.S., & Bethge, M. (2016) Image style transfer using convolutional neural networks. In: CVPR. Las Vegas. pp. 2414-2423. https://openaccess.thecvf.com/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html
3. Huang, X., & Belongie, S. (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1501-1510. http://openaccess.thecvf.com/ content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html
4. Cao, J., Hong, Y., & Niu, L. (2023) Painterly image harmonization in dual domains. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver. pp. 268-276. https://doi.org/10.1609/aaai.v37i1.25099
5. Yan, X., Lu, Y., Shuai, J., & Zhang, S. (2022). Style Image Harmonization via Global-Local Style Mutual Guided. In: ACCV. Macao. pp. 2306-2321. https://openaccess.thecvf.com/content/ACCV2022/html/Yan_Style_Image_Harmonization_via_Global-Local_Style_Mutual_Guided_ACCV_2022_paper.html
6. Song, S.; Zhong, F.; Qin, X.; and Tu, C. (2020) Illumination Harmonization with Gray Mean Scale. In: Computer Graphics International Conference. Geneva. 193-205. https://link.springer.com/chapter/10.1007/978-3-030-61864-3_17
7. Luan, F.; Paris, S.; Shechtman, E.; and Bala, K. (2018) Deep painterly harmonization. In: Computer graphics forum. Bintan. https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13478
8. Zhang, L., Wen, T., and Shi, J. (2020) Deep image blending. In: WACV. CO. 231-240. https://openaccess.thecvf.com/content_WACV_2020/html/Zhang_Deep_Image_Blending_WACV_2020_paper.html
9. Chen, D., Yuan, L., Liao, J., Yu, N., & Hua, G. (2020) Explicit fifilterbank learning for neural image style transfer and image processing. IEEE transactions on pattern analysis and machine intelligence. 43:2373--2387. https://doi.org/10.1109/TPAMI.2020.2964205
10. Niklaus S., Mai Li, and Liu F. (2017) Video frame interpolation via adaptive convolution. In: CVPR. Honolulu. 2270– 2279. https://openaccess.thecvf.com/content_iccv_2017/html/Niklaus_Video_Frame_Interpolation_ICCV_2017_paper.html
11. Xue T., Wu J., Bouman K.L. (2016) Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: NIPS. Barcelona. 43. https://proceedings.neurips.cc/paper_files/paper/2016/hash/03afdbd66e7929b125f8597834fa83a4-Abstract.html