Style Image Harmonization via Global-Local Style Mutual Guided

Xiao Yan, Yang Lu, Juncheng Shuai, and Sanyuan Zhang

College of Computer Science and Technology, Zhejiang University, China
csyanxiao@zju.edu.cn syzhang@cs.zju.edu.cn

Abstract. The process of style image harmonization is attaching an area of the source image to the target style image to form a harmonious new image. Existing methods generally have problems such as distorted foreground, missing content, and semantic inconsistencies caused by the excessive transfer of local style. In this paper, we present a framework for style image harmonization via global and local styles mutual guided to ameliorate these problems. Specifically, we learn to extract global and local information from the Vision Transformer and Convolutional Neural Networks, and adaptively fuse the two kinds of information under a multi-scale fusion structure to ameliorate disharmony between foreground and background styles. Then we train the blending network GradGAN to smooth the image gradient. Finally, we take both style and gradient into consideration to solve the sudden change in the blended boundary gradient. In addition, supervision is unnecessary in our training process. Our experimental results show that our algorithm can balance global and local styles in the foreground stylization, retaining the original information of the object while keeping the boundary gradient smooth, which is more advanced than other methods.

1 Introduction

Style image harmonization is a kind of image synthesis technique. It allows artists to create new artworks with existing materials. When pasting keying footage with different styles onto the background image, style image harmonization helps pasted materials to mix the style of the background image and make the overall image harmonious. Artistic image editing is a time-consuming process and is difficult to edit under style images. Due to the sensitivity of the human visual system [1], this disharmony of the synthetic image can cause visual discomfort. These discords mainly stem from (1) the inconsistency between the foreground and the background styles, (2) and a sudden change of gradient at the boundary of foreground and background. The problems of inconsistent styles between the foreground and the background, the incoordination of factors such as color and texture between the two, and the sudden change in gradient of the boundary are remaining to be solved.

Recently, some deep learning methods can be applied to the style image harmonization, but there are still some other problems. Wu et al. [2] proposed

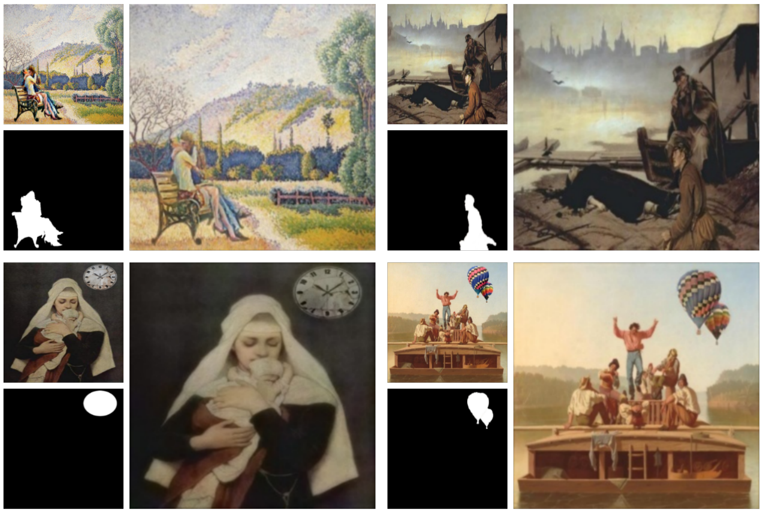

Fig. 1. With the mask of the source image and the composite image, our algorithm can transfer the style of the background to the foreground pasted object and smooth the boundary to make the overall image harmonious.

to combine Generative Adversarial Networks (GANs) [3] and traditional Poisson blending to synthesize real-looking images, but its training required a well-blended Ground Truth as supervision. And there was still a background color seeping into the foreground, causing the foreground to lose its semantics. In addition, artifacts were produced in some areas. Zhang et al. [4] proposed to jointly optimize the Poisson blending, content and style calculated from the deep network to iteratively update the image blend area, but it had obvious foreground distortion caused by excessive style transfer. In conclusion, the main reason for the existing problem caused by the current methods is the use of Convolutional Neural Networks (CNNs) [5–8], which cause the foreground to be affected by the corresponding regional style of the background, resulting in distortion and artifacts. Lately, the transformer-based style transfer proposed by Deng et al. [9] solved the problem that CNNs have difficulty obtaining global information and content leak of input images on style transfer, but it still cannot be applied to the local style transfer, and having a global style on the foreground makes it incompatible with the surroundings.

Style image harmonization consists of two parts: stylization and harmonization. There are some excellent methods of style transfer and image harmonization, but simple combinations cannot bring desirable results. Cong et al. [10, 11] found that converting the foreground domain to the background domain helps to guide the harmonization of the foreground with background information. Sofiiuk et al. [12] used an encoder-decoder framework, which combines pre-trained foreground-aware deep high-resolution network to obtain composite images with semantic visuals. However, combining the state-of-the-art methods mentioned above with style transfer models, will still have problems such as distortion of the foreground leading to bad visuals.

In this paper, we present a framework for style images harmonization via global and local styles mutual guided, which solves the problems of foreground distortion, content loss, and semantic inconsistencies caused by the excessive transfer of local style in the existing methods, and realize a better integration of foreground pasted object into the background style image. For the harmonization of foreground and background styles, we come up with a method for style transfer that blends global and local styles. Learn from the Vision Transformer(ViT) to extract global information and CNNs to extract local information, and adaptively fuse them using a multi-scale fusion structure. For the discord caused by the sudden change of gradient at the boundary of foreground and background, we train the blending network GradGAN to recover the gradient smoothness of the blended images, and then fuse style and gradient result in a harmonious deep image. In addition, we improved the blending loss so that the training process does not require any supervision.

Our main contributions are as follows:

- We propose a novel Blending Decoder that learns to extract global and local information from the ViT and CNNs, and blends this information to make the pasted foreground have a more reasonable style.
- Motivated by Liu et al [13], we propose a multi-scale adaptive fusion structure that bridges ViT and CNNs.
- Different from Wu et al. [2], we improve the blending loss so that the training process only needs the foreground object, background image, and mask without supervision.
- Our experimental results show that our algorithm can balance global and local styles in the foreground stylization, retaining the original information of the object while keeping the boundary gradient smooth, which is more advanced than other methods. Some parts of the results as shown in Fig. 1.

2 Relative Work

2.1 Style Transfer

The task of style transfer is to transfer the style of one drawing into another [5, 6, 14, 15]. Early style transfer was achieved by histogram matching [16] or global image statistics transfer [17]. Gatys et al. [5] designed the first style transfer algorithm with neural networks, using the convolutional features of VGG19 [18] and its Gram matrix to represent content and style. Huang et al. [6] proposed model-based arbitrary style transfer by making the mean and standard deviation of each channel of the content images the same as those of the style images, which is commonly used in various generation tasks [19–22]. Li et al. [7] utilized the idea of de-stylization and stylization, making the multi-layer stylization modules enable the styles of rendered images to be transferred on multiple feature scales. Recently, Vision Transformer (ViT) [22] has been widely used in the field of vision [23–26]. In order to solve the locality and spatial invariance of CNNs, Deng et al. [9] proposed transformer-based style transfer and Content-aware Positional Encoding (CAPE) to adapt to different input sizes. However, there are few methods specifically addressing the problem of local style transfer in the harmonization of style images.

2.2 Image Blending

The task of image blending is to paste an area of the cropped source image onto the target image and make the image looks harmonious as a whole. Traditional blending methods use low-level appearance statistics [27–30] to adjust the foreground image. Alpha Blend [31] uses alpha channel values of the foreground and background to blend images. Recent blending techniques primarily use gradient smoothing [32–34], which targets the smooth transition of gradients at blend boundaries due to human sensitivity to regions of the sudden change of gradient. The earliest work [35] reconstructs pixels in blended regions through gradient-domain consistency. A large number of methods are used for realistic image blending [36, 37, 12], only a few methods are for style images. Luan et al. [38] proposed the use of iterative stylistic transfers and refining them with adjacent pixels, which was the first method of blending paintings. Recent methods [2, 4] combined neural networks and Poisson blending to generate realistic images. Jiang et al. [39] used cropping perturbed images to handle the stylistic images blending, which uses 3D color lookup tables (LUTs) to find information such as hue, brightness, and contrast. All of these approaches distort the foreground and lose its semantics. Our algorithm blends global and local styles, then smooths the blended boundary, and improves the deficiencies in the existing methods.

3 Algorithm

3.1 Overview

Our model achieves style image harmonization followed by StyTr^2 [9], WCT [7], and GP-GAN [2]. Our training data pair is $(x, y, m), x, y, m \in R^{W \times H \times 3}$. x is a composite image containing the foreground, y is the corresponding background image, and m is mask of the foreground. The goal is to blend the foreground of x into the entire image, maintain its texture and semantics while transferring style, and smoothly transitioning the paste boundary gradient with the surrounding gradient. We learn from a generator with an encoder-decoder structure to turn the source image into the target image. The objective function is shown in Eq. 1.

$$target = style(BD(TE(p), CE(p))) + poisson(GradGAN(p)), p = (x, y, m). \tag{1}$$

TE is Transformer Encoder, CE is CNN Encoder, BD is Blending Decoder, $GradGAN$ is a network for image preliminary blending, $style$ and $poisson$ are style constraint and poisson constraint for images, and *target* is the generation target.

Our framework contains five main components: A global style transfer based on the transformer, a local style transfer based on CNN, a global and local blending module, a gradient smoothing module, and a style-gradient fusion module. Fig. 2 provides an overview of our framework. First, we will introduce global and local mutual guided style transfer (Sec. 3.2), so that the source cropped area has the target style while more in line with its semantics. Second, gradient-guided image fusion (Sec. 3.3) to make the stylized areas smoother in the gradient at the paste boundary. Finally, we detail the objective function (Sec. 3.4).

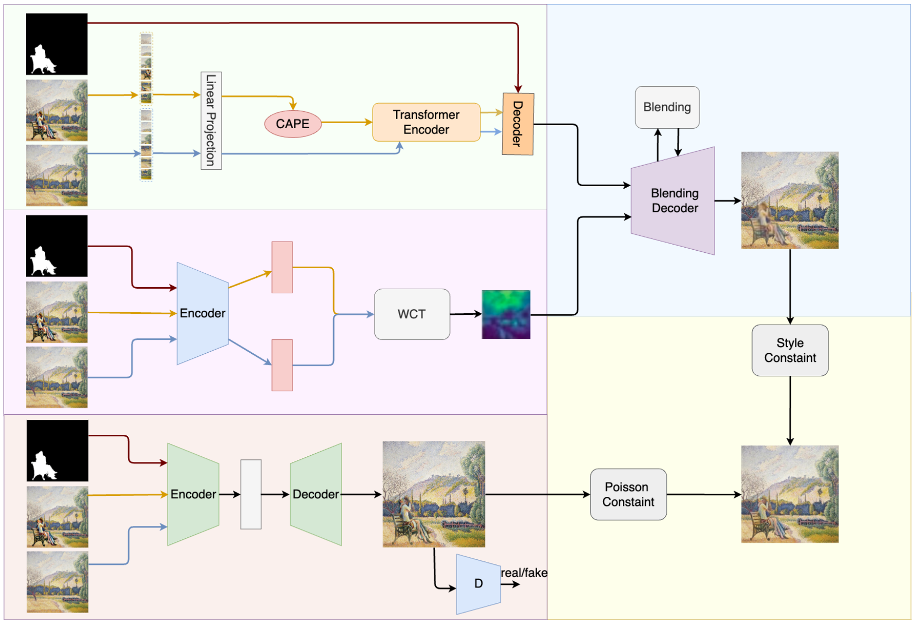

Fig. 2. Our framework contains five main components. Global style transfer based on the transformer (green part) and local style transfer based on CNN (pink part) encode style and content image separately. The global and local blending module Blending Decoder (blue part) decodes the latent code to get an image of the stylized foreground. GradGAN (orange part) generates a gradient smooth image of the boundary of the pasted area, blending styles and gradients (yellow part), and so on getting the final stylistic image harmonization output.

3.2 Global and Local Mutual Guided Style Transfer

To solve the discord caused by excessively style transfer, we propose a global and local mutual guided style transfer, using a multi-scale fusion structure to bridge transformer and CNN. Blend the extracted global and local style to make the foreground has a more reasonable style. The framework consists of two encoders and one decoder.

Encoder We utilize WCT decoders, which extract features from high-level (relu_5) to low-level (relu_1) via VGG19 [18] for Whitening and Coloring Transform (WCT).

We use transformer encoder [9] as another encoder, which contains a style encoder and a content encoder. As shown in Fig. 2, the composition and background images are used to obtain the patch sequence of images through linear projection respectively, and Content-aware Positional Encoding (CAPE) is used only for the content sequence. The input sequence is encoded as Q (query) K (key) V (value), giving the sequence outputs of style and content respectively.

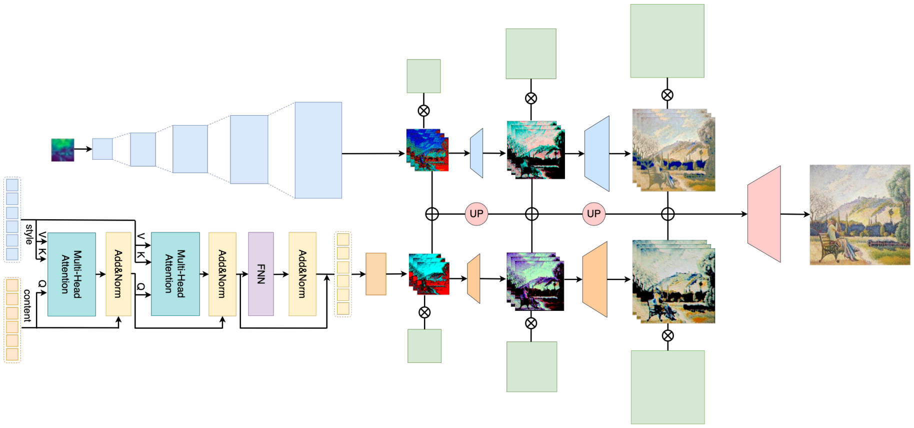

Fig. 3. We propose an adaptive multi-scale Blending Decoder. Using a multi-scale fusion structure to connect the equivalent feature maps of transformers and CNNs, bridge transformer decoders and CNN decoders, which blends global and local styles.

Adaptive Multi-scale Blending Decoder Inspired by ASFF [13], we propose an adaptive multi-scale fusion structure that bridges the WCT decoder and the transformer decoder.

As shown in Fig. 3, in the CNN decoder path, we change the VGG19 decoder's structure in the WCT to extract three feature maps of different scales. In the transformer decoder path, the style sequence is represented as K and V, and the content sequence is represented as Q. The transformer decoder layer contains multi-head attention and a Factorization Machine supported Neural Network (FNN). The output sequence of the transformer is in the shape of $\frac{W \times H \times C}{64}$. Then use the decoder to obtain three feature maps that are equivalent to the CNN decoder output size.

Multi-scale fusion structure connects the equivalent feature maps, adaptively learning fusion spatial weights, specifically shown in Eq. 2.

$$y_{ij} = \alpha \cdot up(\alpha_{ij}^{1} \cdot x_{ij}^{C1} + \alpha_{ij}^{2} \cdot x_{ij}^{T1}) + \beta \cdot up(\beta_{ij}^{1} \cdot x_{ij}^{C2} + \beta_{ij}^{2} \cdot x_{ij}^{T2}) + \gamma \cdot up(\gamma_{ij}^{1} \cdot x_{ij}^{C3} + \gamma_{ij}^{2} \cdot x_{ij}^{T3}), \tag{2}$$

Where y_{ij} represents the pixel value at (i, j) of the output image, $x_{ij}^{C1}, x_{ij}^{C2}, x_{ij}^{C3}$ represent pixel values at (i, j) of the feature maps of three different scales of CNN path, $x_{ij}^{T1}, x_{ij}^{T2}, x_{ij}^{T3}$ represent the pixel values at (i, j) of the feature map of three different scales of the transformer path, $\alpha_{ij} \in \alpha, \beta_{ij} \in \beta, \gamma_{ij} \in \gamma$ represent the spatial weight values at different scales, *up* represents upsampling, and α, β, γ represent the fusion weights of different scales. At the meantime, make $\alpha_{ij}^{1} + \alpha_{ij}^{2} = 1, \beta_{ij}^{1} + \beta_{ij}^{2} = 1, \gamma_{ij}^{1} + \gamma_{ij}^{2} = 1$, where α, β, γ is defined by Eq. 3.

$$\alpha_{ij}^{l} = \frac{e^{\lambda_{\alpha_{ij}^{l}}}}{e^{\lambda_{\alpha_{ij}^{1}}} + e^{\lambda_{\alpha_{ij}^{2}}}}, \beta_{ij}^{l} = \frac{e^{\lambda_{\beta_{ij}^{l}}}}{e^{\lambda_{\beta_{ij}^{1}}} + e^{\lambda_{\beta_{ij}^{2}}}}, \gamma_{ij}^{l} = \frac{e^{\lambda_{\gamma_{ij}^{l}}}}{e^{\lambda_{\gamma_{ij}^{1}}} + e^{\lambda_{\alpha_{ij}^{2}}}}. \tag{3}$$

Spatial fusion weights at each scale are obtained through back propagation. The middle output y is decoded as a fusion map after the style transfer.

3.3 Gradient-guided Image Fusion

After obtaining a stylized foreground, sudden change of the boundary gradient can still cause visual discomfort. Therefore, we introduce gradient-guided image fusion based on style transfer. Utilize the style information of the stylized image and the gradient information of the fusion image to make the final output more harmonious. To achieve this goal, the GAN is firstly trained to generate gradient fusion images, and then the gradient and style are fused using style and gradient constraint.

GradGAN We use the VGG encoder-decoder [18] structure as a generator to fuse the composite image to obtain an image after gradient smoothing. And use the patch discriminator to train against the generator to get a more realistic one.

Fusion We smooth the gradient of the foreground boundary of the composite image using GradGAN, but it is still not harmonious. Different from GP-GAN [2] which needs the gradient of the original image, we fuse the gradient and the style of the generated image to obtain the final deep blending image. The low-level style and the high-level gradient are used for fusion, and the underlying features are extracted as the style constraint and the high-level features are extracted as gradient constraint.

$$\begin{aligned} S(x, x^{style}) &= \sum_{ij} ||F(x_{ij}) - x_{ij}^{style}||_2, \\ G(x, x^{grad}) &= \sum_{ij} ||P(x_{ij}^{grad}) - L(x_{ij})||_2. \end{aligned} \tag{4}$$

The style constraint and gradient constraint for the target are shown in Eq. 4. x^{style} is the image generated by style transfer, x^{grad} is GradGAN blending image, P is gradient operator, F is the filter and L is the Laplace operator that extracts the low-level style and the high-level gradient of the image respectively.

$$T(x) = \omega \cdot S(x, x^{style}) + \varphi \cdot G(x, x^{grad}). \tag{5}$$

The final optimization goal is shown in Eq. 5. ω is style reserved parameters and φ is gradient reserved parameters.

3.4 Optimization

Style Transfer Loss The result of the style transfer should have the same content as the composite image and the same style as the background image [7]. Therefore, the style transfer loss consists of three parts: the content perception loss, the style perception loss, and the reconstruction loss. This loss is defined by Eq. 6.

$$\begin{aligned} L_c &= \frac{1}{N_c}\sum_{i=1}^{N_c} ||\phi_i[x] - \phi_i[x^{compose}]||_2, \\ L_s &= \frac{1}{N_s}\sum_{i=1}^{N_s} ||Gram_i[x] - Gram_i[x^{background}]||_2, Gram_i = \phi_i[\cdot]\phi_i[\cdot]^T, \end{aligned} \tag{6}$$

where x is output image, $x^{compose}$ is the composite image, $x^{background}$ is the background image, and ϕ_i is the feature map of middle layer of VGG19.

Self-supervision can help network training [40]. Therefore, we use the reconstruction loss $L_{rec} = L_c(x, I) + L_s(x, I)$ to learn more accurate content and style representations. Input two identical images I, both as content images and as style images.

$$L_{transfer} = \lambda_c L_c + \lambda_s L_s + \lambda_{rec} L_{rec}. \tag{7}$$

The total style transfer loss is shown in Eq. 7. $\lambda_c, \lambda_s, \lambda_{rec}$ is set separately for the weights of each loss.

Blending Loss Different from GP-GAN which requires Ground Truth to recover low-resolution coarse images, our model only uses composite images and background images to recover preliminary fusion image. Specifically, blending loss consists of three parts: the generating adversarial loss, the perceptual loss, and the gradient loss.

$$\begin{aligned} &L_{adv}^G = E_{x' \sim P_{data}(x')}[\log D(x')] + E_{I \sim P_{data}(I)}(\log(1 - D(decoder(encoder(I))))), \\ &L_{adv}^D = -E_{x' \sim P_{data}(x')}[\log D(x')] - E_{x \sim P_G}[\log(1 - D(x)]. \end{aligned} \tag{8}$$

Generative adversarial loss is defined by Eq. 8, where I is the source composite image, x' is the real background image, x is the generated image, and D is the discriminator.

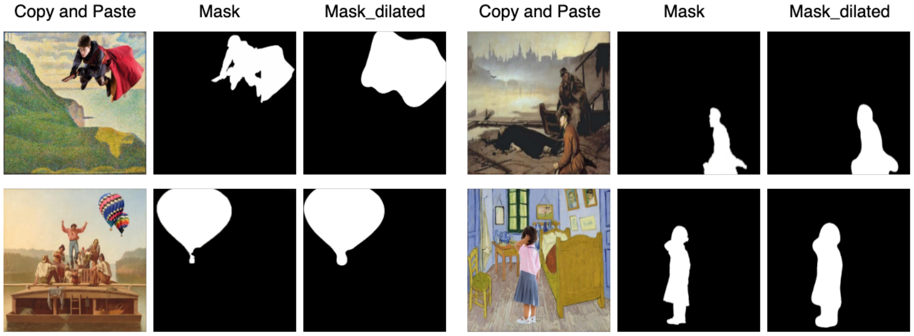

Fig. 4. The source composite image and *mask* and *mask_dilated* corresponding to the foreground. We dilate the *mask* and calculate only the gradient loss inside the *mask_dilated*.

The perceptual loss is L2 loss [41], which accelerates training and produces sharp images compared to the L1 loss [42]. We use L2 loss to make the output content the same as the foreground content of the composite image, and the output style the same as the background style.

$$L_{grad} = \sum_{ij}(|x_{i,j} - x_{i-1,j}| + |x_{i,j-1} - x_{i,j}|). \tag{9}$$

The gradient loss penalizes the output gradient. Since only the gradient of the composite image boundary needs to be smoothed, we dilate the mask and only calculate the gradient loss inside the dilated mask. Mask and dilated mask are shown in Fig. 4. This loss is defined by Eq. 9, where x is the generated image.

$$L_{harmony} = \lambda_{adv} L_{adv} + \lambda_2 L_2 + \lambda_{grad} L_{grad}. \tag{10}$$

The total fusion loss is shown in Eq. 10. $\lambda_{adv}, \lambda_2, \lambda_{grad}$ is set separately for the weights of each loss.

4 Experiments

4.1 Implementation Details

This section describes the implementation details of our method. For style transfer branches, we use StyTr^2 [9] and VGG19 [18] as pre-trained models, adopt Adam [43] optimizer, and employ warm-up training strategies [44]. The initial learning rate is set to 5×10^{-4}, and the decays to 10^{-5}. The $conv1\_1$, $conv2\_1$, $conv3\_1$, $conv4\_1$ of VGG19 are chosen as style representation and $conv4\_1$ as content representation. λ_c is set to 7, λ_s is set to 10, and λ_{rec} is set to 10 in Eq. 7. For GradGAN branches, we adopt Adam optimizer, where α is set to 10^{-4}, β_1 is set to 0.9, and β_2 is set to 0.999. ω is set to 1, φ is set to 1, λ_2 is set to 10 in Eq. 10. All images are reshaped into 256×256, and the datasets we used are from [2, 38].

4.2 Experimental Results

In this section, we compare our method with the existing methods. Qualitative and quantitative comparisons were made, including ablation experiments, comparative experiments, and user studies.

Ablation Experiments To fully illustrate the need for global and local blending, and gradient fusion, example results with different degrees of texture information richness are shown. As shown in the Fig. 5, the full model is superior to other baseline models. The result of w/o global (WCT [7] combine with GradGAN) is that the local style around the foreground transfer to the object, which causes the object to lose its original semantics and produce distortion. The result of w/o global&grad (WCT [7]) seems to be more harmonious with the entire image in the gradient, but the original information in the foreground is seriously missing due to its local style transfer. The result of w/o local (stytr^2 [9] combined with GradGAN) is that this position-independent global style makes the foreground incompatible with the surrounding. The result of w/o local&grad (StyTr^2 [9]) retains the original information, but it is still relatively abrupt in the entire image because it does not transfer local style and may have a big change in the gradient. The result of w/o grad (only style transfer) is more reasonable than the first two in style processing, but there is still a certain degree of texture loss. In contrast, The full model shows the best results, with the foreground area retaining its original texture and semantics while transferring the background style, and aligning with the surrounding style, while its boundary and background gradients are smoother.

Comparative Experiment We compare our method with three others (See Fig. 6): GP-GAN [2], SSH [39], and Deep Image Blending [4]. GP-GAN is an image fusion algorithm that combines Poisson fusion and GAN, and is trained in a supervised manner. However, the color is transferred to the foreground from around, making it inconsistent with the original semantics, and it also creates artifacts in some areas. SSH uses dual data enhancement to crop perturbed images, and uses 3D color lookup tables to find information such as hue, brightness, contrast, etc., to process both real and stylistic images. However, the pasted boundary of the processed style image is very obtrusive. Deep Image Blending is an improved method based on GP-GAN, using VGG19 for style transfer, and joint optimization of Poisson loss and content style loss to blend deep images. But it distorts and produces artifacts more heavily in the foreground. Due to using VGG19 [18] as a style transfer network, the foreground has a distinctly localized style, which is unrealistic. Our model shows the best results, balancing global and local styles in the foreground stylization, maintaining the original information of the object, and making the surrounding gradient smoother.

Fig. 7 shows the results of stylizing and harmonizing the image foreground using the mainstream styles transfer models [9, 7], harmonization models [10, 11], and stylization and harmonization of combination models separately. As

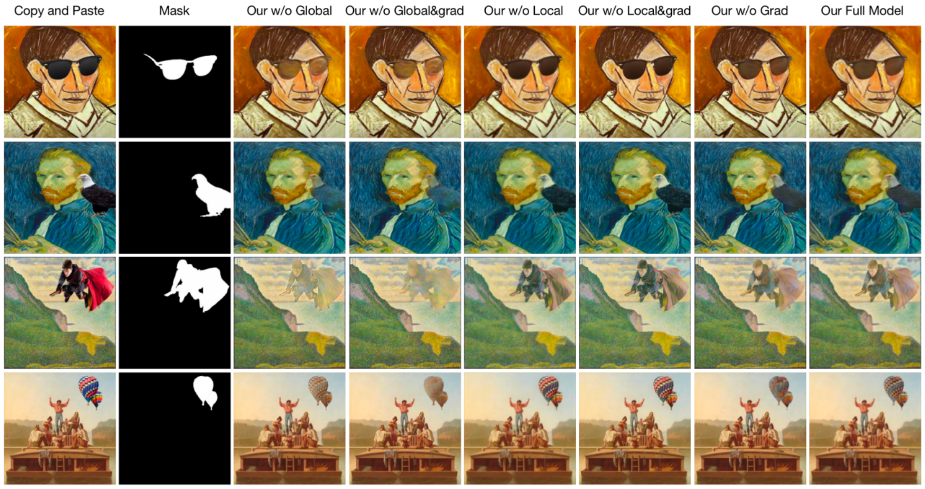

Fig. 5. Experimental study of ablation of global, local, and gradient fusion. The full model shows the best result, however other baseline models either distort the original information or are not in harmony with the surrounding, or there are problems such as a sudden change in the gradient.

shown in the Fig. 7, the style transfer models of the traditional CNN places more emphasis on local style, and produces content distortion, such as the result of the WCT that distorts the eye part of the foreground character. Recent style transfer using transformers places more emphasis on global style, but is less coordinated with the surroundings, and the variation of gradient makes the visual effect more obtrusive. The harmonization models do not handle stylistic images very well. BargainNet [10] and Dovenet [11] converted the foreground domain to a background domain, with background information guiding the harmonization of the foreground. But we can see that the brightness of the foreground is relatively obtrusive relative to the surrounding pixels. There are translation failures in some images: the foreground lacks style information and does not fit well into the background. The result of stylization and harmonization of combination models still has foreground distortion or boundary pixel obtrusion, and the style image harmonization is not well handled. We blend global and local styles, smooth the boundary, and achieve good results.

User Studies We conducted user studies to quantitatively evaluate the experimental results. The first experiment verifies the quality of the image generated by judging whether the provided image has been edited by the user. The second experiment compares the quality of ours and others by selecting the optimal one from the images generated by different methods. At the same time, we also measured the user reaction time to further verify the effectiveness and robustness of our method.

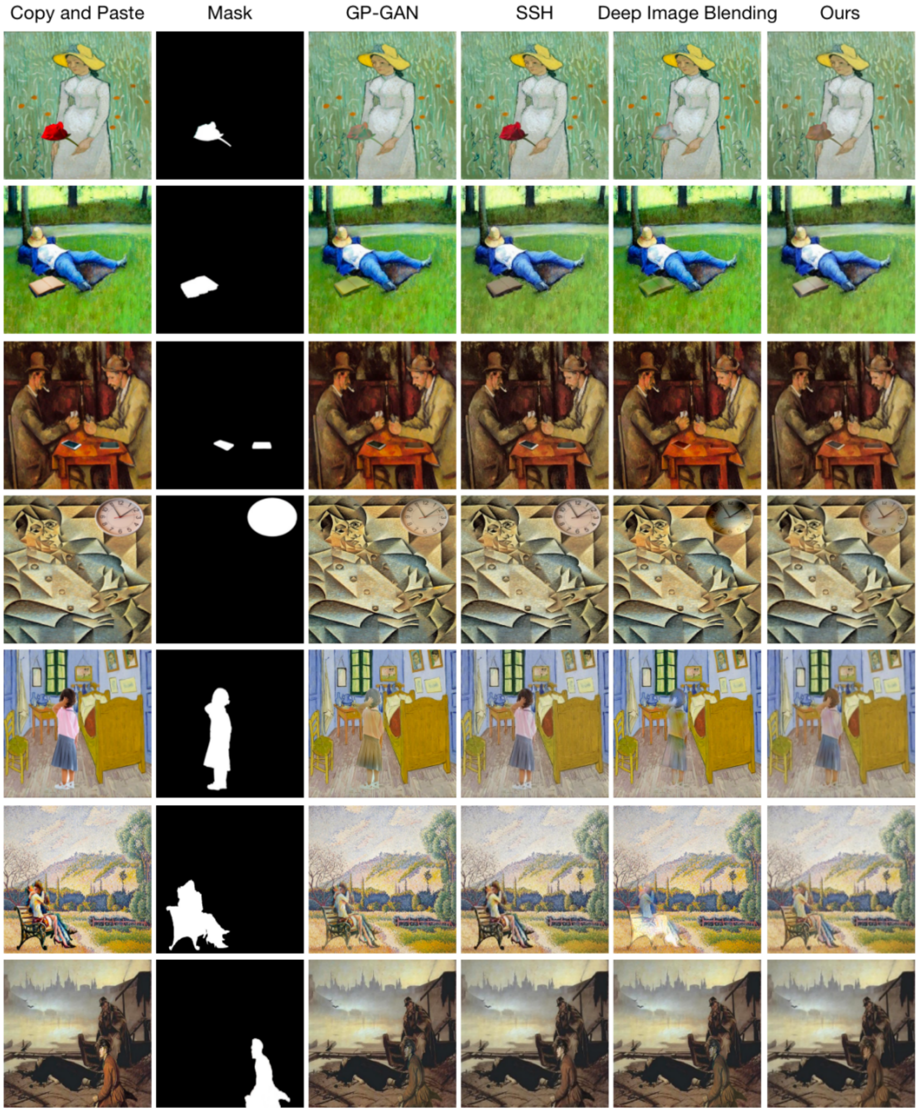

Fig. 6. Comparison of our method with others. The results of GP-GAN and Deep Image Blending show that the foreground is distorted and artifacts are produced. The gradient of pasted boundary processed by SSH is very obtrusive. Our model shows the best results, balancing global and local styles, maintaining the original information of the object, and smoother with the surrounding gradients.

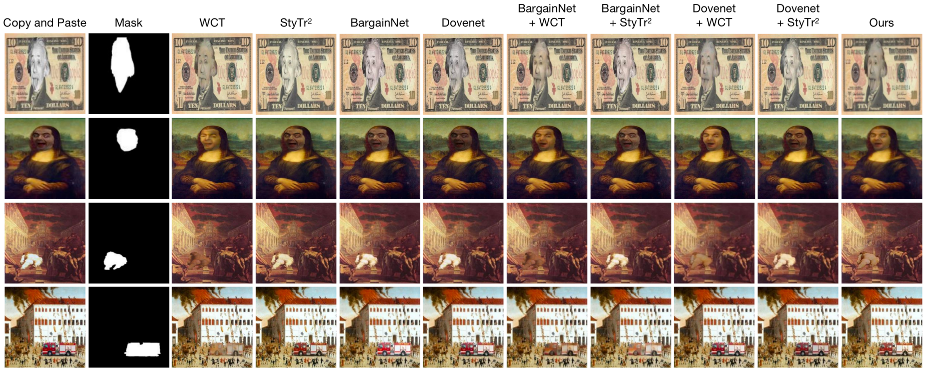

Fig. 7. Either direct use of mainstream or the latest stylized and harmonized combination models to stylize and harmonize the foreground can handle the harmonization task of style images very well. And our model blends global and local styles, smooths the boundary, and achieves the best results. (See the supplementary materials for a clearer version.)

User Study 1: Whether to Edit We invited 30 users, and randomly selected 20 images generated by four different methods (sec. 4.2), and unprocessed images. The user needs to answer whether it has been edited and click on the edited part. We recorded the response time of the user's answer. We asked the user in advance if they were familiar with the image in case of the impact of prior knowledge. And we only think that the correct answer is the sample if both the edit and the part click are correct, in case other parts that may be edited will cause interference. We counted the response time and error rate of each image being answered, and the unedited image statistics were answered correctly for ease of comparison. As shown in Fig. 8, it is clear that our method has a higher answer error rate and a longer user response time than other methods, which is closest to the unedited image. The high rate of answer errors indicates that most users believe that this is an unedited image. The longer the user's reaction time, indicating that the user's observation time is longer, the more difficult it is to distinguish. The larger the area that intersects the coordinate axis, the better the result obtained by the method in general.

User Study 2: Quality Comparison We invited 30 users, and randomly selected 10 groups of images, and each group was processed by four different methods. The user needed to choose the one with the best effect in each group, and we recorded the reaction time selected by the user. The faster the reaction time, the better the method is than the others. As shown in Fig. 8, ours are considered by most users to be the best, and some images are far better than others.

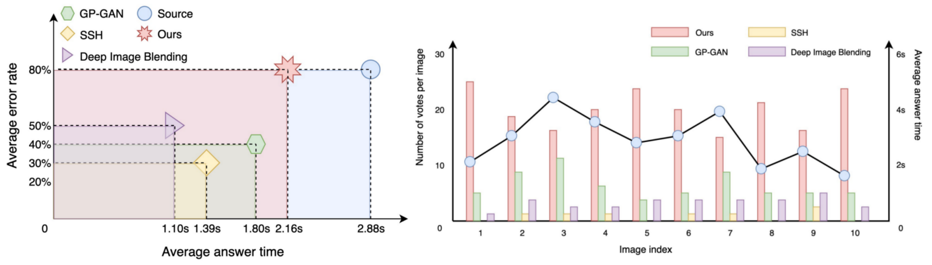

Fig. 8. Statistical results of "User Study 1 - Whether to Edit" and "User Study 2 - Quality Comparison". The higher the answer error rate, the longer the average reaction time, and the larger the area that intersects the coordinate axis, the better the result. Our method has the best results and is closer to the unprocessed image. And ours is the most selected by the user, and some images are far better than others.

5 Conclusion

Style image harmonization is the process of pasting the cropped areas from the source image into the background style image and harmonizing two as a whole. We propose a style image harmonization in which global and local information guide mutually, and solve the problems of foreground distortion, content loss, and semantic inconsistencies caused by the excessive transfer of local styles in the existing methods. Firstly, global and local styles are extracted by the transformer and CNNs separately, and an adaptive multi-scale fusion structure bridges the transformer decoder and CNNs decoder to fuse global and local styles. Secondly, the blending network GradGAN smooths the image gradient. Finally, the fusion style and gradient result in a harmonious deep image.

To evaluate the method presented, we made quantitative and qualitative comparisons. Compared to the existing methods, our model shows the best results, balancing the global and local styles on the foreground stylization, maintaining the original information of the object, and smoother with the surrounding gradient. User studies have shown that the images processed by our model are often considered unedited ones, which is superior to the results of other methods. We believe that our approach assists artists in editing their work, providing more possibilities for users to create works of art.

References

1. S. Xue, A. Agarwala, J.D.: Understanding and improving the realism of image composites. ACM Transactions on Graphics **31** (2012) 1–10
2. H. Wu, S. Zheng, J.Z.: Gp-gan: Towards realistic high-resolution image blending. ACM International Conference on Multimedia (2019) 2487–2495
3. I. Goodfellow, J. Pouget-Abadie, M.M.: Generative adversarial nets. Advances in Neural Information Processing Systems **27** (2014)
4. L. Zhang, T. Wen, J.S.: Deep image blending. IEEE/CVF Winter Conference on Applications of Computer Vision (2020) 231–240

5. L. A. Gates, A. S. Ecker, M.B.: Image style transfer using convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2414–2423
6. X. Huang, S.B.: Arbitrary style transfer in real-time with adaptive instance normalization. IEEE International Conference on Computer Vision (ICCV) (2017) 1501–1510
7. Y. Li, C. Fang, J.Y.: Universal style transfer via feature transforms. Advances in Neural Information Processing Systems **30** (2017)
8. S. Liu, T. Lin, D.H.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. IEEE International Conference on Computer Vision (ICCV) (2021) 6649–6658
9. Y. Deng, F. Tang, X.P.: StyTr^2: Image style transfer with transformers. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 11326–11336
10. W. Cong, L. Niu, J.Z.: Bargainnet: Background-guided domain translation for image harmonization. IEEE International Conference on Multimedia and Expo (ICME) (2021) 1–6
11. W. Cong, L. Niu, J.Z.: Dovenet: Deep image harmonization via domain verification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 8394–8403
12. K. Sofiiuk, P. Popenova, A.K.: Foreground-aware semantic representations for image harmonization. IEEE/CVF Winter Conference on Applications of Computer Vision (2021) 1620–1629
13. S. Liu, D. Huang, Y.W.: Learning spatial fusion for single-shot object detection. arXiv preprint (2019)
14. Y. Jing, X. Liu, Y.D.: Dynamic instance normalization for arbitrary style transfer. AAAI Conference on Artificial Intelligence **34** (2020) 4369–4376
15. J. An, S. Huang, Y.S.: Unbiased image style transfer via reversible neural flows. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 862–871
16. F. Pitie F, A. C. Kokaram, R.D.: N-dimensional probability density function transfer and its application to color transfer. IEEE International Conference on Computer Vision (ICCV) (2005) 1434–1439
17. E. Reinhard, M. Adhikhmin, B.G.: Color transfer between images. IEEE Computer graphics and applications **21** (2001) 34–41
18. A. Sengupta, Y. Ye, R.W.: Going deeper in spiking neural networks: Vgg and residual architectures. Frontiers in Neuroscience **13** (2019) 95
19. X. Xia, M. Zhang, T.X.: Joint bilateral learning for real-time universal photorealistic style transfer. European Conference on Computer Vision (ECCV) (2020) 327–342
20. J. Gu, J.C.Y.: Adain-based tunable cyclegan for efficient unsupervised low-dose ct denoising. IEEE Transactions on Computational Imaging **7** (2021) 73–85
21. T. Karras, S. Laine, M.A.: Analyzing and improving the image quality of stylegan. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 8110–8119
22. A. Dosovitskiy, L. Beyer, A.K.: An image is worth 16x16 words: Transformers for image recognition at scal. arXiv preprint (2020)
23. L. Yuan, Y. Chen, T.W.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. IEEE International Conference on Computer Vision (ICCV) (2021) 558–567
24. A. Arnab, M. Dehghani, G.H.: Vivit: A video vision transformer. IEEE International Conference on Computer Vision (ICCV) (2021) 6836–6846

25. W. Wang, E. Xie, X.L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media (2022) 1–10
26. P. Zhang, X. Dai, J.Y.: Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. IEEE International Conference on Computer Vision (ICCV) (2021) 2998–3008
27. M. Grundland, R. Vohra, G.P.W.: Cross dissolve without cross fade: Preserving contrast, color and salience in image compositing. Computer Graphics Forum **25** (2006) 557–586
28. K. Sunkavalli, M. K. Johnson, W.M.: Multi-scale image harmonization. ACM Transactions on Graphics (TOG) **29** (2010) 1–10
29. M. W. Tao, M. K. Johnson, S.P.: Error-tolerant image compositing. European Conference on Computer Vision (ECCV) (2010) 31–44
30. J. Jia, J. Sun, C.K.T.: Drag-and-drop pasting. ACM Transactions on graphics (TOG) **25** (2006) 631–637
31. T. Porter, T.D.: Compositing digital images. Annual Conference on Computer Graphics and Interactive Techniques (1984) 253–259
32. R. Fattal, D. Lischinski, M.W.: Gradient domain high dynamic range compression. Annual Conference on Computer Graphics and Interactive Techniques (2002) 249–256
33. A. Levin, A. Zomet, S.P.: Seamless image stitching in the gradient domain. European Conference on Computer Vision (ECCV) (2004) 377–389
34. R. Szeliski, M. Uyttendaele, D.S.: Fast poisson blending using multi-splines. IEEE International Conference on Computational Photography (ICCP) (2011) 1–8
35. P. Pérez, M. Gangnet, A.B.: Poisson image editing. ACM SIGGRAPH 2003 Papers (2003) 313–318
36. J. Ling, H. Xue, L.S.: Region-aware adaptive instance normalization for image harmonization. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 9361–9370
37. Z. Guo, H. Zheng, Y.J.: Intrinsic image harmonization. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 16367–16376
38. F. Luan, S. Paris, E.S.: Deep painterly harmonization. Computer Graphics Forum **37** (2018) 95–106
39. Y. Jiang, H. Zhang, J.Z.: Ssh: A self-supervised framework for image harmonization. IEEE International Conference on Computer Vision (ICCV) (2021) 4832–4841
40. L. Jing, Y.T.: Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **43** (2020) 4037–40581
41. J. Johnson, A. Alahi, L.F.F.: Perceptual losses for real-time style transfer and super-resolution. European Conference on Computer Vision (ECCV) (2016) 694–711
42. H. Zhao, O. Gallo, I.F.: Loss functions for image restoration with neural networks[j]. ieee transactions on computational imaging. IEEE transactions on pattern analysis and machine intelligence **3** (2016) 47–40581
43. D. P. Kingma, J.B.: Adam: A method for stochastic optimization. arXiv preprint (2014)
44. R. Xiong, Y. Yang, D.H.: On layer normalization in the transformer architecture. International Conference on Machine Learning (PMLR) (2020) 10524–10533