Multi-Attention Network for Arbitrary Style Transfer

Sihui Hua and Dongdong Zhang(✉)

Department of Computer Science and Technology, Tongji University, Shanghai, China
{16hsh,ddzhang}@tongji.edu.cn

Abstract. Arbitrary style transfer task is to synthesize a new image with the content of an image and the style of another image. With the development of deep learning, the effect and efficiency of arbitrary style transfer have been greatly improved. Although the existing methods have made good progress, there are still limitations in the preservation of salient content structure and detailed style patterns. In this paper, we propose Multi-Attention Network for Arbitrary Style Transfer (MANet). In details, we utilize the multi-attention mechanism to extract the salient structure of the content image and the detailed texture of the style image, and transfer the rich style patterns in the art works into the content image. Moreover, we design a novel attention loss to preserve the significant information of the content. The experimental results show that our model can efficiently generate more high-quality stylized images than those generated by the state-of-the-art (SOTA) methods.

Keywords: Style transfer · Attention mechanism · Deep learning

1 Introduction

With the development of deep learning, recent years have witnessed the continuous progress of image style transfer. Especially with the success of convolutional neural network (CNN), it has become the main research method of image style transfer. Gatys et al. [1] first showed that the content information could be extracted from natural images and the style information could be obtained from artworks through a pre-trained VGG network [12], and proposed an image reconstruction algorithm to realize the stylization. However, this algorithm is restricted by low efficiency. In order to reduce the computational cost, several methods [3,6,14] based on feed-forward networks have been developed, which can effectively generate stylized images, but they are limited to a fixed style or lack of visual quality.

For arbitrary style transfer, several approaches have been proposed. For instance, AdaIN [2] is the first method realize effective real-time style transfer, which only adjusts the mean and variance values of content image to match those values of style image. However, this method is too simple to affect the output quality. WCT [7] matches content features with style features through

T. Mantoro et al. (Eds.): ICONIP 2021, LNCS 13111, pp. 390–401, 2021.
https://doi.org/10.1007/978-3-030-92273-3_32

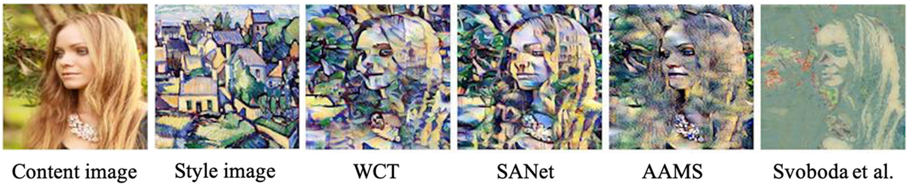

Fig. 1. The existing methods do not well maintain the original semantics of content and transfer the fine texture of style. The output images of WCT [7] and SANet [9] lack part of the woman's outline in the content image. While for AAMS [16] and Svoboda et al. [13], the extraction of style is insufficient.

the whitening and coloring process of covariance replacing variance, and then the stylized features are embedded in the pre-trained encoder-decoder to synthesize stylized images. Nevertheless, the increase of feature dimension will greatly increase the calculation of complexity. Avatar-Net [11] is a style decoration method based on image blocks. It maps the features of content and style while maintaining the content structure to get better quality. Xing et al. [15] proposed a portrait-aware method, which applied fine-grained and coarse-grained style transfer to the background and the entire image, respectively. Jung et al. [4] proposed Graph Instance Normalization (GrIN), which made the style transfer approach more robust by taking into account similar information shared between instances. Svoboda et al. [13] employed a two-stage peer-regularization layer that used graph convolutions to reorganize style and content in latent space.

Recently, with the introduction of the attention mechanism, great breakthroughs have been made in style transfer. Yao et al. [16] applied the self-attention mechanism to style transfer model and added multi-stroke control. In addition, Park et al. [9] proposed a style-attentional network which matched style features with content features.

Although the above methods have achieved good results, there are still some limitations. First of all, these methods do not take into account the original semantic structure of content image comprehensively. Moreover, the existing methods do not reflect the detailed texture of the style image, making the style of output deviate. As shown in Fig. 1, some algorithms lack part of the outline of the content image, while some algorithms do not extract the style patterns enough. These seriously reduce the quality of stylized images.

To address these limitations, we propose Multi-Attention Network for Arbitrary Style Transfer (MANet). It employs multi-attention mechanism to preserve the salient structure of content image and the detailed texture of style image, and renders the rich style patterns of art works into the generated result. The proposed MANet includes unary content attention module (UA), pairwise style attention module (PA) and fusion attention module (FA). The UA module and PA module model the salient boundaries of content and learn within-region relationships of style through the unary self-attention mechanism and the pairwise self-attention mechanism respectively, so as to retain the main content features and vivid style patterns. Then, we integrate the enhanced content and style features through FA module to achieve stylized result. In addition, we propose a novel attention loss. When content image and style image generate stylized result, we design the attention loss to minimize the difference of attention between content image and stylized result, which ensures that the salient semantic information of the content image can be preserved in the process of style transfer. To summarize, our main contributions are as follows:

– We propose an efficient multi-attention network (MANet) for arbitrary style transfer, including UA module, PA module and FA module.
– We propose a novel attention loss function to enhance the salient features of the content, which can retain the salient structure of the original content image.
– Various experiments show that our method can preserve the salient structure of content images and detailed texture of style images, and combine content features and style features flexibly.

2 Related Work

In order to realize arbitrary style transfer efficiently, some methods have been proposed. Huang et al. [2] proposed AdaIN, which transferred the statistics of the mean and variance of the channel in the feature space for style transfer. Li et al. [7] proposed WCT, which integrated whitening and coloring transforms to match the statistical distribution and correlation between content and style features to achieve style transfer. AdaIn and WCT holistically adjust the content features to match the second-order statistics of style features. Svoboda et al. [13] proposed a two-stage peer-regularization layer that used graph convolutions to reorganize style and content in latent space.

Recently, several approaches introduced self-attention mechanism to obtain high-quality stylized images. Park et al. [9] proposed a style-attentional network, which flexibly matched the semantically nearest style features to the content features. It slightly modified the self-attention mechanism (such as the number of input data) to learn the mapping between the content and style features. Yao et al. [16] adapted the self-attention to introduce a residual feature map to catch salient characteristics within content images, and then generated stylized images with multiple stroke patterns.

However, the above methods cannot effectively preserve salient content structure and capture detailed style patterns. The disadvantages of these methods can be observed in Sect. 4.2. To this end, we propose an arbitrary style transfer network, named multi-attention network, and we design a new attention loss to enhance salient content features. In this way, the method can naturally integrate style and content while maintaining the original structure of content and detailed texture of style.

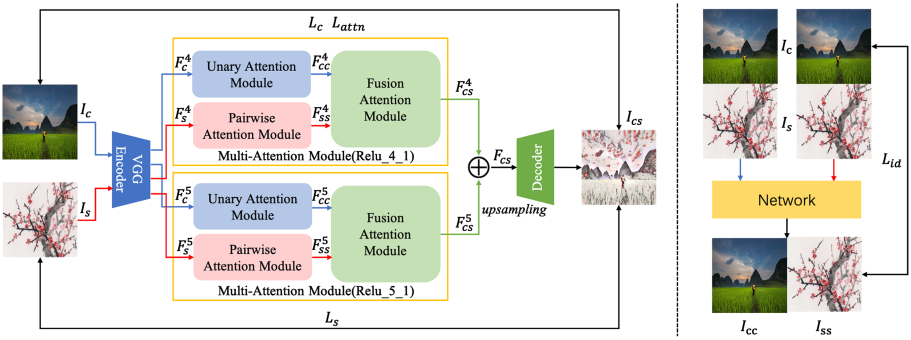

Fig. 2. The overall framework of our network.

3 Methodology

For the purpose of achieving arbitrary style transfer, the proposed network consists of an encoder-decoder module and a multi-attention module. Figure 2 shows the overall framework. A pre-trained VGG19 network [12] is used as an encoder and a symmetric decoder to extract deep features. Inputting a content image I_c and an arbitrary style image I_s, we can extract their respective feature maps $F_c^i = E(I_c)$ and $F_s^i = E(I_s)$ (i is the number of a certain layer). To consider both the overall style distribution and local styles, we extract the features of two layers ($Relu\_4\_1$ and $Relu\_5\_1$) in the encoder, and combine the final results. Meanwhile, considering that the use of a public encoder can only extract features in a few specific fields, it lacks attention to the salient structural parts of the content and the internal pixel relationship of the style. Therefore, we propose a multi-attention module that can learn salient boundaries of content and learn within-region relationship of style separately, and can integrate content and style features appropriately in each position to obtain the stylized features F_{cs}. The multi-attention module is described in detail in Sect. 3.1. The decoder follows the setting of [2] and gets the stylized result $I_{cs} = D(F_{cs})$. We design a new attention loss to preserve the salient structure of the content. Section 3.2 describes the four loss functions we used to train the model.

3.1 Multi-Attention Module

As shown in Fig. 3, the multi-attention module includes three parts: unary content attention module, pairwise style attention module and fusion attention module. With content/style feature $F_c^i(F_s^i)$ through the content/style attention module, we can get $F_{cc}^i(F_{ss}^i)$ that retain salient features. Then the fusion attention module appropriately embeds style features in the content feature maps.

Unary Content Attention Module. The semantic information of content should be preserved during the style transfer to keep the structure consistent

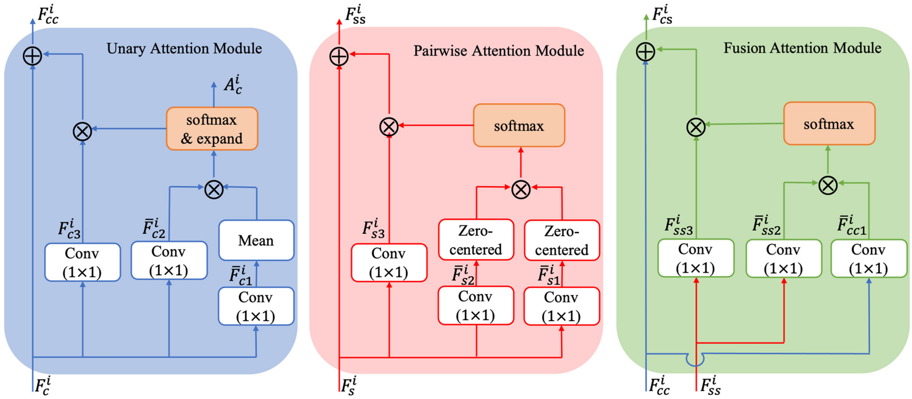

Fig. 3. The framework of Multi-Attention Module.

before and after stylization. Following [17], we design the Unary Content Attention module (UA) to model the salient boundaries. The content feature map F_c^i from the encoder is normalized and fed to two convolutional layers to obtain new feature maps $\bar{F}_{c1}^i$ and $\bar{F}_{c2}^i$. At the same time, we feed the unnormalized F_c^i to another convolutional layer to obtain the feature map F_{c3}^i. Unary content attention map is A_c^i is formulated as follows:

$$A_c^i = expand(softmax(\mu(\bar{F}_{c1}^i)^T \otimes \bar{F}_{c2}^i)) \tag{1}$$

where $expand(\cdot)$ represents to expand the size, $\otimes$ represents the matrix multiplication and $\mu(\cdot)$ denotes the mean of features. Then we get the optimized content features F_{cc}^i:

$$F_{cc}^i = F_{c3}^i \otimes A_c^i + F_c^i \tag{2}$$

Pairwise Style Attention Module. It is necessary to learn the main style patterns of the style image, and render the style appropriately to the content image in the process of style transfer. For a certain style, it contains many elements. So we design the Pairwise Style Attention module (PA) to learn pixel relationships within the same category region in the style image following [17]. Thus to strengthen the components and highlight the expressiveness of the style and to extract the detailed texture. The style feature map F_s^i is normalized and fed to two convolutional layers to obtain new feature maps $\bar{F}_{s1}^i$ and $\bar{F}_{s2}^i$, and then they are subtracted the mean. At the same time, F_s^i is fed to another convolutional layer to obtain the feature map F_{s3}^i. The pairwise style attention map A_s^i is formulated as follows:

$$A_s^i = softmax((\bar{F}_{s1}^i - \mu(\bar{F}_{s1}^i))^T \otimes (\bar{F}_{s2}^i - \mu(\bar{F}_{s2}^i))) \tag{3}$$

Then, we get the enhanced style features F_{ss}^i:

$$F_{ss}^i = F_{s3}^i \otimes {A_s^i}^T + F_s^i \tag{4}$$

Fusion Attention Module. Through UA and PA modules, we obtain the enhanced content and style features. Similar to [9], we add FA module to appropriately integrate content and style features. Figure 3 shows the FA process. The enhanced content features F_{cc}^i and style features F_{ss}^i are normalized to obtain $\bar{F}_{cc}^i$ and $\bar{F}_{ss}^i$. Then we feed $\bar{F}_{cc}^i$, $\bar{F}_{ss}^i$ and F_{ss}^i to three convolutional layers to generate three new feature maps $\bar{F}_{cc1}^i$, $\bar{F}_{ss2}^i$ and F_{ss3}^i. The correlation map A_{cs}^i is formulated as follows:

$$A_{cs}^i = softmax({\bar{F}_{cc1}^i}^T \otimes \bar{F}_{ss2}^i) \tag{5}$$

where A_{cs}^i maps the correspondence between the enhanced content features F_{cc}^i and the enhanced style features F_{ss}^i. Then, we can calculate the stylized feature map of each layer ($Relu\_4\_1$ and $Relu\_5\_1$) as follows:

$$F_{cs}^i = F_{ss3}^i \otimes {A_{cs}^i}^T + F_{cc}^i \tag{6}$$

Finally, we get the final stylized feature map F_{cs} from the two MANets:

$$F_{cs} = F_{cs}^4 + upsampling(F_{cs}^5) \tag{7}$$

3.2 Loss Function

Our model contains 4 loss functions during training.

Content Loss. Similar to AdaIN [2], the content loss is computed using the pre-trained encoder. The content loss L_c is utilized to make the stylized image close to the content image in content, as follows:

$$L_c = \sum_{i=4}^{5} \|\phi_i(I_{cs}) - \phi_i(I_c)\|_2 \tag{8}$$

where ϕ_i represents the feature map extracted from i-th layer in the encoder.

Attention Loss. The stylized image should preserve the salient characteristics of the original content image. Therefore, we propose a new attention loss to minimize the difference in attention between content image and stylized result, taking into account the insufficiency that the salient information in content may be distorted. In addition, it helps to make the visual effect of the generated image better. The attention loss L_{attn} is as follows:

$$L_{attn} = \sum_{i=4}^{5} \left\|A_c^i(I_{cs}) - A_c^i(I_c)\right\|_2 \tag{9}$$

where $A_c^i(\cdot)$ represents the attention map obtained by feeding the features extracted from i-th layer in the encoder to the content attention module.

Style Loss. We apply the style loss in AdaIN [2], and L_s is used to make the generated image close to the style image in style:

$$L_s = \sum_{i=1}^{5} \|\mu(\phi_i(I_{cs})) - \mu(\phi_i(I_s))\|_2 + \|\sigma(\phi_i(I_{cs})) - \sigma(\phi_i(I_s))\|_2 \tag{10}$$

where $\sigma(\cdot)$ represents the variance of features.

Identity Loss. Similar to [9], we introduce the identity loss to consider both the global statistics and the local mapping relation between content and style features. The identity loss is formulated as follows:

$$\begin{aligned} L_{id} =& \lambda_{id1}(\|I_{cc} - I_c\|_2 + \|I_{ss} - I_s\|_2) \\ &+ \lambda_{id2} \sum_{i=1}^{5}(\|\phi_i(I_{cc}) - \phi_i(I_c)\|_2 + \|\phi_i(I_{ss}) - \phi_i(I_s)\|_2) \end{aligned} \tag{11}$$

where $I_{cc}(I_{ss})$ denotes the generated result by using two same content (style) images as content and style images simultaneously. λ_{id1} and λ_{id2} denote the weight of identity loss. Our total loss function formula is as follows:

$$L = \lambda_c L_c + \lambda_{attn} L_{attn} + \lambda_s L_s + L_{id} \tag{12}$$

where λ_c, λ_{attn} and λ_s are weighting parameters.

4 Experiments

4.1 Implementation Details

When we trained the network, we used MS-COCO [8] as content dataset, and WikiArt [10] as style dataset, both of which contain approximately $80,000$ training images. In the training process, we used the Adam optimizer [5] with a learning rate of 0.0001 and a batch size of five content-style image pairs, and we randomly cropped the 256×256 pixels area of both images. In the testing stage, any input size can be supported. The weights λ_c, λ_{attn}, λ_s, λ_{id1}, and λ_{id2} are set to 3, 5, 3, 50, and 1 to balance each loss.

4.2 Comparison with Prior Methods

Qualitative Evaluation. We show the style transfer results of five SOTA technologies: AdaIN [2], WCT [7], SANet [9], AAMS [16] and Svoboda et al. [13], as shown in Fig. 4. AdaIN only needs to adjust the mean and variance of content features. However, due to the oversimplification of this method, its output quality is affected, and the color distribution of some content is often preserved (e.g., the preserved complexion of content in row 1 and the missing eyes and eyebrows in row 5 in Fig. 4). WCT synthesizes stylized images through whitening and coloring

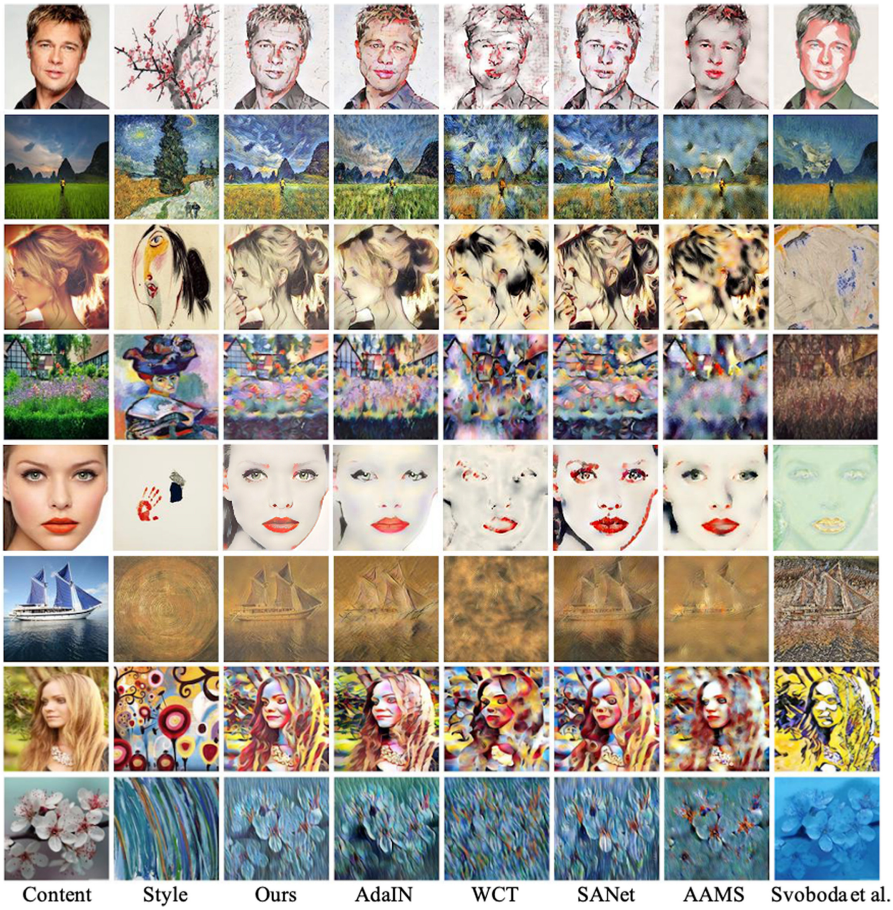

Fig. 4. Example results for comparisons

transformations, but sometimes has content distortion (e.g., the distorted faces and objects in Fig. 4). SANet uses the style-attentional network to match the semantically closest style feature to the content feature, retaining the global style and local style. However, sometimes the salient features of the content image are distorted (e.g., the blurred face in rows 3, 7 and the unclear main structure in rows 4,6). AAMS also introduces a self-attention mechanism to increase multi-stroke control. However, the content and style are not well integrated and the stylized image is blurry. The content structure is not obvious, and the texture information of the style is not displayed (rows 2, 3, 4, 6, and 8 in Fig. 4). Svoboda et al. introduces a fixpoint triplet style loss and a two-stage peer-regularization layer. But because it learns the overall artist style, the correlation between the stylization result and the style image is very limited, and it cannot be controlled and reflect the main mode of the style (rows 1, 3, 4, 5, 6, and 7). In contrast, our

Table 1. Quantitative comparison over different methods.

	AdaIN	WCT	SANet	AAMS	Svoboda et al.	Ours
Preference/%	15.1	10.6	19.7	13.8	7.9	32.9
L_c (content)	10.96	13.29	14.41	12.26	12.54	11.93
L_s (style)	14.92	15.30	14.90	15.19	15.74	12.57

algorithm can adapt to multiple style and preserve the structural information of the content, as shown in Fig. 4.

Different from the above methods, our multi-attention network can further retain the salient information of content and the detailed texture of style. In addition, our method can effectively integrate content and style by learning the relationship between content and style features, so that the generated result not only retains the semantic information of the content, but also contains rich colors and textures of the style.

Quantitative Evaluation. To compare our visual performance with the above-mentioned SOTA methods further, we conducted a user study. For each method, we used 18 content images and 20 style images to generate 360 results. We selected 20 content and style combinations, showing the generated images of six methods to 50 participants to choose their favorite result for each combination. We collect $1,000$ votes in total, and the preference results are shown in the second row of Table 1. The results show that our method can obtain preferred results.

Evaluating the results of style transfer is very subjective. For quantitative evaluation, we also made two comparisons. They are reported in the last two rows of Table 1. We compared different methods through content and style loss. The evaluation metrics include content and style terms used in previous methods (AdaIn [2]). The results were obtained by calculating the average content and style loss of 100 images in 512×512 scale. It can be seen that our method does not directly minimize content and style loss, because we use a total of four loss types. Nevertheless, the style loss obtained by our method is the lowest among all methods, and the content loss is only slightly higher than that of AdaIN. It indicates that our method favors fully stylized results rather than results with high content fidelity.

4.3 Ablation Study

Attention Loss Analysis. We compare the styled results with and without attention loss to verify the effect of attention loss in this section. As shown in Fig. 5, compared with the stylized results without attention loss, using attention loss can retain the salient features of the original content image and more visible content structure. With attention loss, the stylized results take into account that

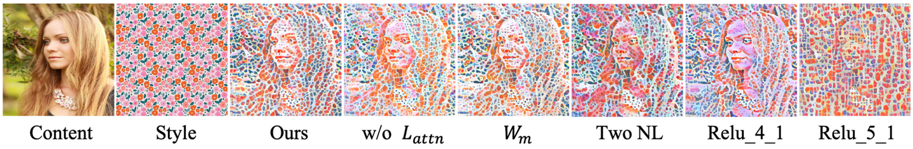

Fig. 5. Ablation study.

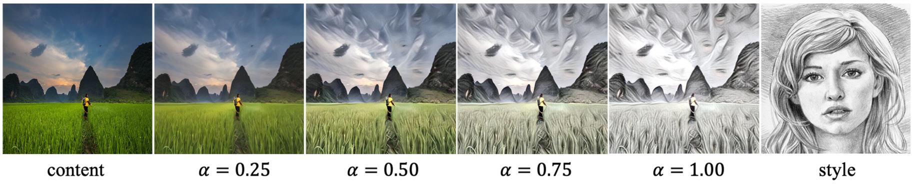

Fig. 6. Content-style trade-off.

the salient information in the content image may be distorted, and preserve the significant information of content.

Multi-Attention Module Analysis. In this section, we will try to reveal the effectiveness of the multi-attention module. In [17], the transformation in unary non-local (NL) neural networks was changed to use independent 1×1 convolution transformation W_m. But only one convolution is not enough to extract the salient information of the content in style transfer. Figure 5 shows the result of using W_m. It can be seen that the salient structure of the content is not clear enough. We also compared the results of using two NL [17] instead of UA and PA modules. As illustrated, this method cannot extract effective features for content and style separately, and the result is not good enough in reflecting content and style.

In addition, Fig. 5 shows two stylized outputs obtained from *Relu*_4_1 and *Relu*_5_1, respectively. The content structure is good when only *Relu*_4_1 is used for style transfer. But the partial styles are not displayed well. On the contrary, *Relu*_5_1 obtains a detailed style mode, but the content structure is distorted. In our work, we integrates two layers to obtain a completely stylized result while preserving the salient information of the content.

4.4 Runtime Controls

Content-Style Trade-Off. During training, to control the degree of stylization, we can adjust the style weight λ_s in Eq. 12. During test time, the degree of stylization can be controlled by changing α:

$$I_{cs} = D(\alpha F_{cs} + (1 - \alpha) F_c) \tag{13}$$

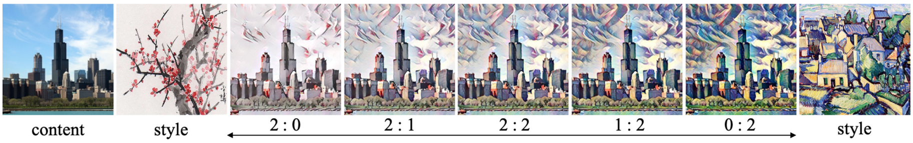

Fig. 7. Style interpolation.

α can be adjusted between 0 and 1. When $\alpha = 0$, the original content image is obtained, and when $\alpha = 1$, we obtain the most stylized image. Figure 6 presents the examples.

Style Interpolation. In order to obtain multi-style results, feature maps from different styles can be fed into the decoder, thereby combining multiple style images into one generated result. Figure 7 shows the results.

5 Conclusions

In this paper, we propose a multi-attention network to extract the salient structure of the content image and the detailed texture of the style image, and appropriately combine the content and style. Furthermore, we propose a new attention loss to consider the deficiencies that the salient information in the content image may be distorted, and to ensure that the significant semantic information of the content image is preserved in the style transfer process. Sufficient experiments show that our network can consider the salient content structure and detailed style patterns to synthesize better results.

References

1. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
2. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
3. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
4. Jung, D., Yang, S., Choi, J., Kim, C.: Arbitrary style transfer using graph instance normalization. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 1596–1600. IEEE (2020)
5. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

6. Li, C., Wand, M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 702–716. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_43
7. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. arXiv preprint arXiv:1705.08086 (2017)
8. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
9. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5880–5888 (2019)
10. Phillips, F., Mackintosh, B.: Wiki art gallery, inc.: a case for critical thinking. Issues Account. Educ. **26**(3), 593–608 (2011)
11. Sheng, L., Lin, Z., Shao, J., Wang, X.: Avatar-net: multi-scale zero-shot style transfer by feature decoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8242–8250 (2018)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
13. Svoboda, J., Anoosheh, A., Osendorfer, C., Masci, J.: Two-stage peer-regularized feature recombination for arbitrary image style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13816–13825 (2020)
14. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Texture networks: feed-forward synthesis of textures and stylized images. In: ICML, vol. 1, p. 4 (2016)
15. Xing, Y., Li, J., Dai, T., Tang, Q., Niu, L., Xia, S.T.: Portrait-aware artistic style transfer. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 2117–2121. IEEE (2018)
16. Yao, Y., Ren, J., Xie, X., Liu, W., Liu, Y.J., Wang, J.: Attention-aware multi-stroke style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1467–1475 (2019)
17. Yin, M., et al.: Disentangled non-local neural networks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12360, pp. 191–207. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_12