A Novel Arbitrary Style Transfer Algorithm via Multi-Order Attention

1st Jiaqi Chang
School of Electrical Engineering and Automation
Anhui University
Hefei, China
e-mail: z22201048@stu.ahu.edu.cn

2nd Dong Sun*
School of Electrical Engineering and Automation
Anhui University
Hefei, China
* Corresponding author's e-mail: sundong@ahu.edu.cn

3rd Qingwei Gao
Anhui University
Hefei, China
e-mail: qingweigao@ahu.edu.cn

4th Yixiang Lu
School of Electrical Engineering and Automation
Anhui University
Hefei, China
e-mail: lyxahu@ahu.edu.cn

5th Muxi Bao
School of Computer Science
Anhui University
Hefei, China
e-mail: 1311349865@qq.com

6th Yong Hu
School of Electrical Engineering and Automation
Anhui University
Hefei, China
e-mail: 286784283@qq.com

***Abstract*—Arbitrary style transfer aims to transfer style images to content images. Existing solutions either directly fuse the style features with the content features or adaptively normalize the content features according to the style features to match the global statistics of the two. However, little attention is paid to the positional relationship of the content information, and the distributional relationship of the style features is not taken into account, so the results often suffer from local distortion and style detail errors. To improve this phenomenon, we propose a new arbitrary style transfer algorithm to solve the existing problem and obtain content and style balanced results, which consists of a Multi-Order Attention, and a Merge module. This algorithm is named SMA (Style Transfer Algorithm via Multi- Order Attention). SMA obtains the content encoding and inputs it together with the style encoding into the MoA module to obtain the weighted distribution of the style encoding as well as the overall distribution, and then adaptively normalizes and fuses the content features through the distribution relationship; then, the Merge module fuses the features at multiple levels. In addition, a new multi-order style loss function is derived based on MoA, which can enhance the learning of style details. Finally, experiments demonstrate that our method achieves good results on several metrics. Our algorithm can balance content and style well, making the results more appealing and exploring new angles of style transfer algorithms.**

Keyword-Arbitrary style transfer; Image processing; Image fusion; Artificial Intelligence

I. INTRODUCTION

The purpose of style transfer is to concentrate on how to apply the artistic style of the style image to the content image. The result has the structural position of the content image and the color texture of the style image. Artistic style transfer is not a new thing, on the contrary, artistic style transfer has been a popular research topic in academia and industry. There have been a variety of approaches to style transfer, which can be broadly categorized into two groups: image-optimization methods and model-optimization methods.

Image-optimization methods were early style transfer methods, pioneered by Gatys et al [1]. However, the Gatys approach is very time-consuming as it extracts relevant content features from a pre-trained deep neural network and then iteratively minimizes the loss of content and style. On the basis of their work, methods [2] obtained better results by changing the loss function, and methods [3] began to consider the use of feed-forward networks for style transfer. All of the above methods offer varying degrees of improvement, but image optimization methods are still limited by their slow online optimization process. The model optimization methods, on the other hand, have improved by training an updated neural network and using a feed-forward network to generate images directly, so the whole process is faster than the image optimization method.

Model-optimization methods: The original model optimization methods used feed-forward networks to generate images directly and learning models that could only be used for a specific kind of style, e.g. [4], so these methods were classified as Per-Style-Per-Model. Methods [5] use different network structures as a way of dealing with a variety of styles, but only with more similar styles. These methods are therefore called Multiple-Style-Per-Mode. After this, methods [6] achieves style transfer for arbitrary styles by adjusting the content feature distribution to match the style feature distribution, which can use arbitrary styles as input after a pre-trained model that produces stylized output image. These methods are therefore called Arbitrary-Style-Per-Model (Arbitrary style transfer). Among the above model optimization-based style transfer methods, the arbitrary style transfer is the most flexible and unrestricted,

979-8-3315-3508-7/25/$31.00 ©2025 IEEE

which can transmit arbitrary styles in a single training, and it is also the method that receives the most attention today.

In summary, for attention-based style transfer algorithms, there are fewer methods for processing style features that have been weighted by attention. Existing methods do not pay enough attention to the processing of style features weighted by the attention mechanism. Either they do not consider the distribution of the features, and directly fuse them with the content features, e.g., SANet [7], or they pay too much attention to the similar style features and ignore the dissimilar ones, e.g., AdaAttN [8]. This leads to wrong style details in the result. In this paper, we will explore a more reasonable processing algorithm, and at the same time can be well balanced between content and style.

II. Materials and Methods

This paper propose a new Arbitrary Style Transfer Algorithm (SMA), which reweights the style features by means of the Multi-Order Attention Module (MoA) designed in this paper. It then fuses the global distribution of the style features, and the second-order statistical distribution of the style features, respectively, with the adaptive alignment of the content features. Enhancing similar style features while paying attention to dissimilar style features to obtain detailed styles. Based on the MoA module, a new loss function focusing on the multi-order distribution of stylized images is proposed, which can effectively preserve the stylistic features of each weight to capture the stylistic details. The overall framework of the SMA as shown in Fig. 1.

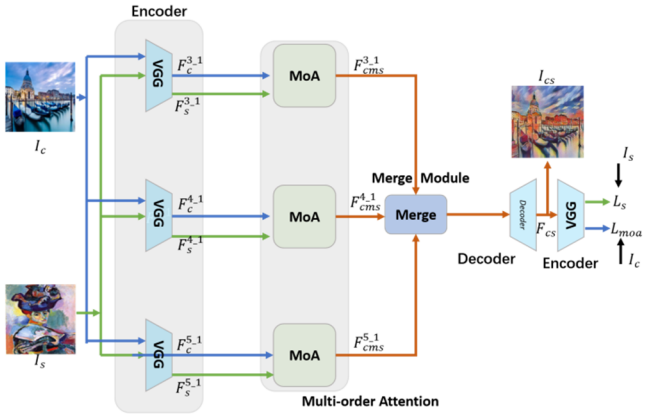

Figure 1. Overview of the SMA architecture

A. Multi-Order Attention Module

This paper compute the multi-order feature distribution after weighting, emphasizing similar features while incorporating information about the remaining features. This improves the stylistic details of the resulting image. We explain this in two steps, the first of which is to compute the distribution of style features after weighting to emphasize similar features. The second is to fuse dissimilar features with low weights. For the former, the existing method AdaIN [9] adjusts the content feature distribution by calculating the style feature variance and mean, but the method is a global distribution and does not consider the local distribution. For the latter, the method SANet [7] fuses all features through the attention mechanism, but the method does not consider the distribution of style features. Similarly, method [8] calculates the standard deviation and mean of style features, but discards features with low weights.

Differences from existing methods, we propose the Multi-Order Attention Module(MoA), which works in the following steps: firstly, calculate the attention graph A of style features and content features; secondly, compute the variance and mean of style features; thirdly, fuse the attention graph with the content features to get the first-order result, and at the same time, align the content features with the distribution of the style features computed in the previous step to get the second-order result; and finally, fuse the two to get the multi-order result.

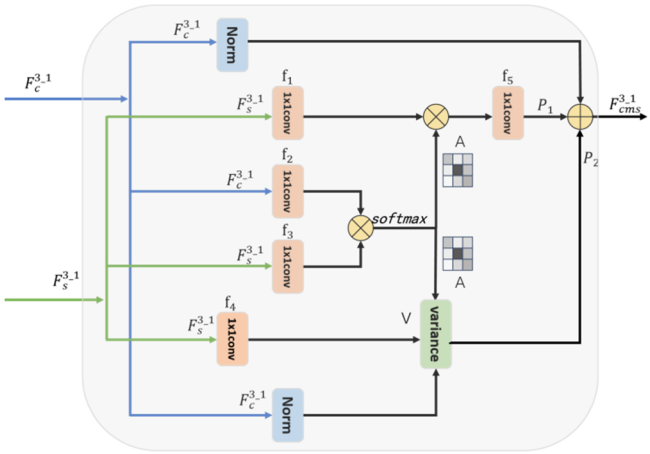

Figure 2. Multi-order Attention Module

The model structure is shown in Fig. 2. First, we compute the attention graph A for content features and style features, which is computed as in method [10]. The expression for A is as follows:

$$A = (F_{cp})f_2 \otimes (F_S)f_3 \tag{1}$$

The F_c,F_s, and A are then taken together to compute the variance and mean to obtain the second-order distribution of the features.

As in method [7-8], by operating the attention graph A with the style features, the style features output through attention can be considered as a distribution weighted by attention. So $(F_S)f_4 \otimes A^T$ can be thought of as the expectation of the distribution of style traits.

$$E(F_S) = (F_S)f_4 \otimes A^T \tag{2}$$

Knowing the expectation of the distribution, and wanting to find the second-order distribution, then you should also find the variance V, which can be found by using the variance formula ($V^2 = E(X^2) - E(X)^2$). The expression is given below:

$$V = [(F_s)f_4]^2 \otimes A^T - [(F_s)f_4 \otimes A^T]^2 \tag{3}$$

After obtaining the second-order distribution of the style features, the content features are adaptively normalized and then aligned with the style features.

$$P_2 = N(F_c) \cdot V + (F_s)f_4 \otimes A^T \tag{4}$$

Similarly, P_1 is equivalent to finding the distribution of style features after attention.

$$P_1 = [(F_s)f_1 \otimes A^T]f_5 \quad (5)$$

In the above equation, the N is the target normalized channel-by-channel by mean-variance. f_1, f_2, f_3, f_4, f_5 are the learnable weight matrix, which is the 1×1 convolution. In summary, our MoA synthesizes distributions of multiple orders, which can be considered to account for the original distribution while taking into account the characteristic second-order distribution. Therefore, MoA can emphasize similar features while incorporating the remaining feature information to improve the details of the resultant images.

B. *Loss functions*

Our overall loss function is:

$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_{cmoa} \mathcal{L}_{cmoa} \quad (6)$$

Where λ_s, λ_{cmoa} are the hyperparameter that controls the weight of the corresponding loss, which are set to 1, 5.

First, for style loss, we use $\mathcal{L}_s$, following existing work [7], we use the same approach to control style passing.

Our proposed new stylized positional structural loss constrains the structural positional information of the stylized image. Also, it constrains the feature map obtained from the multi-order attention operation of the SMA network to match the stylized image with the following equation:

$$\mathcal{L}_{cmoa} = \lambda_{cmoa} \sum_{x=3}^{5} (\| E(N(I_{cs})^{x\_1}) - N(F_c^{X\_1}) \|_2 + \| E(N(I_{cs})^{x\_1}) - N(F_{cms}^{X\_1}) \|_2) \quad (7)$$

Where N is the target normalized channel-by-channel by mean-variance and denotes that it has been decoded by VGG. The operations I_{cs} and F_c in the summation formula are calculated to compare the stylized image with the implicit coding encoded by the image position module to control the loss of positional structural information. I_{cs} and F_{cms} are calculated to compare the stylized image with the implicit coding of the fused images that have passed through the various levels of the subject network, which is a part of the fusion coding that incorporates the content image encoded by the position module and the implicit coding of the stylized image encoded by VGG. This part of the fusion coding is a fusion of the implicit coding of the content images encoded through the position module and the implicit coding of the stylized images encoded through the VGG to control the network loss.

III. Results & Discussion

In this section, to evaluate our algorithm, this paper compare our SMA with 10 state-of-the-art style transfer methods, including 2 baseline algorithms (AdaIN [9], SANet [7]), 1 attention based methods(AdaAttN [8]), 1 text-guided style transfer method(TxST[11]), 1 transformer based methods(StFr [12]), 1 global transformation based methods(EFDM [13]), 1 contrastive learning based method(IEAST [14]), 1 linear interpolation based method: STVAE [17], 1 attention based methods: MANet [18], and this paper's SMA algorithm, for a comparative qualitative as well as quantitative comparison of a total of 10 algorithms for arbitrary style transfer transformation.

A. *Qualitative Comparison*

The representative style transfer results generated by our algorithm and the baseline algorithms are provided in Fig. 3.

Figure 3. Comparison with other methods in arbitrary image style transfe

SANet applies the cross-attention weighted adjustment of features, and we can see that its style transfer is successful, but the content structure information is seriously lost, the building in the picture is full of twisted styles, as in the second column. AdaIN directly adjusts the second-order statistics of the content features with global adjustment, which results in the fact that the content structure information is still seriously lost, and local style distortion occasionally occurs, as in four row, fourth column.

AdaAttN adopts the local attention and obtains the second-order statistics of the features. It is obvious that for the content structure, information is retained intact. Still, there is some loss of information about some styles, approximating color transfer,

and losing image style on some images, such as in fifth row, the fifth and sixth column. The transfer result of the fifth column is completely unbalanced, and the colors and styles of the style pictures are all lost. The sixth column has only color transfer, and the circular structure of the style picture is completely lost. The transfer results of AdaAttN tend to retain content information regardless of style information.

StFr is a feed-forward style, a global style transfer method. It is perfect for image style learning, but it is too fine-grained and loses the original structure information, as shown in sixth row, fifth column, and some styles only have an overall color transfer. EFDM is completely faithful to the color transfer and does not pay attention to the overall structure content information.

IEAST and TxST tend to be more sensory intuitive: IEAST learns the colors of the stylized images, but not the stylistic features, as shown in nineth row, second and fifth column. TxST does not learn style features at all in the eighth and ninth line, although some of its transfer results are very beautiful.

B. Quantitative Results

In this section, in order to evaluate our method quantitatively, we choose several metrics, SSIM [15], LPIPS [16], for quantitative evaluation, as shown in Table I. LPIPS is a widely used metric in measuring diversity.

TABLE I. COMPARISON WITH OTHER METHODS IN ARBITRARY IMAGE STYLE TRANSFER (OVERBOLD: BEST)

Method	SSIM	LPIPS
SANet	0.218	0.635
AdaIN	0.272	0.595
StFr	0.218	0.578
EFDM	0.237	0.579
TxST	0.263	0.554
IEAST	0.284	0.530
SMA	0.290	0.528

C. Efficiency Analysis

We show in Table II the runtime performance of our method at two common resolutions (256×256, 512×512). All experiments were performed using a single Nvidia 3090Ti GPU. Although our method employs multiple levels of multi-order attention and positional edges, the runtime overhead does not increase much. From the runtime results, SMA is still in the same class as methods such as SANet. So, our proposed network can be practically applied to synthesized images.

TABLE II. EXECUTION TIME COMPARISON (IN SECONDS).

Method	256×256	512×512
SANet	0.011	0.025
AdaIN	0.004	0.013
StFr	0.013	0.026
EFDM	0.011	0.039
TxST	0.107	0.394
IEAST	0.065	0.092
SMA	0.026	0.053

D. Training process and parameter setting

We use Adam as an optimizer with the learning rate initially set to 0.0001 and the batch size set to 5 content style pairs. λ_s, λ_{cmoa}which are set to 1, 5. The training process has 100,000 iterations. During the training process, we scaled all the images read in at a resolution of 512×512, kept the aspect ratio, and then cropped them randomly into regions of 256×256 size. The training lasts for 10K iterations on a single Nvidia 4090 GPU. We set the batch size to four content and style image pairs.

E. Dataset

To train our model, we need a single-style image and a collection of content images. In this work, following conventions, we use MS-COCO(about 328k images) as content images and select style images from WikiArt(about 42k images). The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of 328K images. The WikiArt contains painting from 195 different artists. The dataset has 42129 images for training and 10628 images for testing.

F. Ablation Study

In this subsection, we show the impact of our proposed new multi-order statistics loss and the transfer effect of the model with the multi-order attention module. As shown in Table III and Table IV. When $\mathcal{L}_{cmoa}$ is removed, this paper uses the baseline model SANet instead of the corresponding losses and modules. It can be seen that the loss function and MoA module proposed in this paper improve the image quality to some extent.

TABLE III. EFFECTIVENESS OF LOSS FUNCTION (OVERBOLD: BEST)

Metric	Full mode	Without $\mathcal{L}_{cmoa}$
SSIM	0.345	0.272
LPIPS	0.608	0.673

TABLE IV. EFFECTIVENESS OF MODULE (OVERBOLD: BEST)

Metric	Full mode	Without MoA
SSIM	0.272	0.194
LPIPS	0.538	0.640

IV. CONCLUSIONS

Existing style transfer methods do not balance the relationship between content and style well and do not make good use of the distribution of features. To address the above problems, this paper proposes a new style transfer algorithm via Multi-Order Attention, which we call SMA. SMA utilizes the multi-order distribution of features to improve the style details while processing the content images to obtain a more complete structure. SMA consists of the following: The Multi-Order Attention(MoA) module computes weighted statistics by treating the style feature as a weighted multi-order distribution statistic of all the style feature and matches it to the content feature; The Merge module adaptively fuses the multilevel structures; Our algorithm achieves the best results on both the traditional metric SSIM, as well as the deep learning metric LIPIS, thus proving that it can effectively improve the quality of the resulting images.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 62071001), the Nature Science Foundation of Anhui (Nos. 2008085MF192, 2008085MF183, 2208085QF206 and 2308085QF224), the Key Science Project of Anhui Education Department of China (Nos. KJ2018A0012, KJ2019A0022, and KJ2019A0023) and the China Postdoctoral Science Foundation (2023M730009).

REFERENCES

[1] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).

[2] Kolkin, N., Salavon, J., & Shakhnarovich, G. (2019). Style transfer by relaxed optimal transport and self-similarity. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10051-10060).

[3] Chen, D., Yuan, L., Liao, J., Yu, N., & Hua, G. (2017). Stylebank: An explicit representation for neural image style transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1897-1906).

[4] Li, C., & Wand, M. (2016). Precomputed real-time texture synthesis with markovian generative adversarial networks. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14 (pp. 702-716). Springer International Publishing.

[5] Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2017). Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6924-6932).

[6] Zhang, Y., Li, M., Li, R., Jia, K., & Zhang, L. (2022). Exact feature distribution matching for arbitrary style transfer and domain generalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8035-8045).

[7] Park, D. Y., & Lee, K. H. (2019). Arbitrary style transfer with style-attentional networks. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5880-5888).

[8] Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., ... & Ding, E. (2021). Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6649-6658).

[9] Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE international conference on computer vision (pp. 1501-1510).

[10] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7794-7803).

[11] Liu, Z. S., Wang, L. W., Siu, W. C., & Kalogeiton, V. (2022). Name your style: An arbitrary artist-aware image style transfer. arXiv preprint arXiv:2202.13562.

[12] Wu, X., Hu, Z., Sheng, L., & Xu, D. (2021). Styleformer: Real-time arbitrary style transfer via parametric style composition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14618-14627).

[13] Zhang, Y., Li, M., Li, R., Jia, K., & Zhang, L. (2022). Exact feature distribution matching for arbitrary style transfer and domain generalization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8035-8045).

[14] Chen, H., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., & Lu, D. (2021). Artistic style transfer with internal-external learning and contrastive learning. Advances in Neural Information Processing Systems, 34, 26561-26573.

[15] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

[16] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586-595).

[17] Liu, Z. S., Kalogeiton, V., & Cani, M. P. (2021, September). Multiple style transfer via variational autoencoder. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 2413-2417). IEEE.

[18] Deng, Y., Tang, F., Dong, W., Sun, W., Huang, F., & Xu, C. (2020, October). Arbitrary style transfer via multi-adaptation network. In Proceedings of the 28th ACM international conference on multimedia (pp. 2719-2727).