Controllable Feature-Preserving Style Transfer

Feichi Chen, Naye Ji(✉), Youbin Zhao, and Fuxing Gao

Communication University of Zhejiang, Hangzhou 310018, China
jinaye@cuz.edu.cn

Abstract. This paper proposes a new style transfer quality assessment approach introducing quantifiable metrics to optimize. First, we utilize a pre-trained DualStyleGAN model to generate multiple stylized portraits in the style vector space. Then, we design a custom scoring mechanism that uses the newly proposed $CSCI$ and $CCVI$ metrics to evaluate the results' structural similarity, color consistency, and edge retention. We select and optimize the top outputs using human aesthetic standards to obtain the most natural, beautiful, and artistic results. Experimental results show that our proposed evaluation pipeline can effectively improve the quality of style transfer.

Keywords: StyleGAN · Style Transfer · Quality Evaluation

1 Introduction

Style transfer, an extensively researched topic in computer vision, aims to transform the artistic style of images while preserving their content. Recent advancements in deep learning have yielded impressive results; however, ensuring the quality of generated stylized images remains challenging.

Although various methods have been developed to enhance stylization quality, each existing method has its own limitations. Examples include StyleGAN3, DualStyleGAN [2], and StableDiffusion [8]. Current style transfer methods excel in specific aspects while lacking in others. For instance, DualStyleGAN struggles with style transfer on Asian faces, while StableDiffusion produces random results that are hard to control.

These challenges stem from the lack of comprehensive considerations for stylization quality. Some methods prioritize results and content effects, while others emphasize style accuracy, but both approaches have inherent limitations. Is there a way to integrate the strengths of these methods and address their shortcomings? The evaluation methods in this paper aim to answer this question.

Existing style transfer methods have made valuable contributions to the field of stylization. However, these methods have some drawbacks, such as not working well with Asians, and some Anime styles are challenging to transfer. These limitations prevent them from achieving comprehensive and high-quality results.

F. Zhao and D. Miao (Eds.): AIGC 2023, CCIS 1946, pp. 95–104, 2024.
https://doi.org/10.1007/978-981-99-7587-7_8

This paper aims to bridge these gaps and merge the strengths of various techniques to overcome their limitations.

This paper presents a detailed analysis of our approach, highlighting the effectiveness of our proposed metrics and their impact on the style transfer process. Furthermore, we conduct experiments on diverse datasets, comparing our results with state-of-the-art methods to demonstrate the superiority of our approach.

Our main contributions lie in the following aspects:

1. We propose a set of quantitative metrics to evaluate style transfer performance, including perceptual loss, structural similarity, edge preservation, color preservation, and visual saliency. A multi-layer weighted scoring approach assesses these factors and filters out low-quality results.
2. We build upon previous style transfer techniques and further optimize the process. We enhance stylized outputs' accuracy and aesthetic appeal by incorporating prior knowledge and fine-tuning details.
3. We conduct extensive experiments on multiple portrait datasets. Comparisons with state-of-the-art methods demonstrate the superiority of our approach in improving stylization quality.
4. We develop an end-to-end pipeline integrating style transfer, quality evaluation, and refinement. This represents an important step towards controllable and high-fidelity artistic stylization.

2 Related Work

2.1 StyleGAN-Related Models

The StyleGAN model, proposed by Karras et al. [1], is a generative adversarial network (GAN) that leverages style transfer techniques to generate high-quality images. This model introduces a novel generator architecture, enabling intuitive control over image synthesis at multiple scales. Compared to traditional GANs, StyleGAN improves distribution quality metrics and interpolation properties. It achieves this by disentangling latent factors of variation, effectively separating different aspects of an image into distinct components that can be manipulated independently. The model is flexible and can be applied to various image synthesis tasks. Notably, it includes the creation of a unique human face dataset called FFHQ and introduces automated methods for quantifying interpolation quality and disentanglement. Overall, the StyleGAN model significantly advances image generation and synthesis.

StyleGAN has indeed catalyzed the development of numerous style transfer models, including DualStyleGAN, JoJoGAN [7], and VToonify [13]. Additionally, it has given rise to variant style transfer models that explore different approaches. An example is StyleCLIP [10], which combines the principles of StyleGAN and CLIP models.

The demand for personalized and stylized images has significantly contributed to the popularity of StyleGAN and its derivatives among academic researchers and the general public. DualStyleGAN has introduced a wide range of styles for image style transfer, while JoJoGAN offers a one-shot style transfer capability. These models have garnered considerable attention due to their ability to generate high-quality simulated images. For our research, we selected StyleGAN as our fundamental model due to its proven performance and versatility.

However, StyleGAN and its derivatives still exhibit certain limitations and drawbacks in their performance. While these models have shown the ability to generate impressive results, they often struggle to achieve a consistent and cohesive style and character across generated images. Furthermore, a notable disadvantage is the inadequate performance of DualStyleGAN on Asian faces which is found in our experiments due to its training on FFHQ, a predominantly European and American face dataset. Moreover, in certain styles like Anime, existing models face challenges in effectively transferring those styles to facial images. These shortcomings have posed challenges in ensuring the accuracy and validity of the generated results, prompting the emergence of further studies in this area.

2.2 Evaluation of Style Transfer

To achieve controllable style transfer, previous studies have proposed various evaluation methods. These research efforts aim to quantify the effectiveness of style transfer techniques. One notable approach is the Style-Eval method introduced by Wang et al. [3], which has shown promising results across various style transfers. Style-Eval offers several advantages, including a novel quantitative evaluation framework based on three measurable quality factors. This comprehensive approach thoroughly assesses style transfer quality from multiple perspectives.

Wright and Ommer have introduced a novel method called ArtFID [11] to complement the predominantly qualitative evaluation schemes currently employed. This proposed metric demonstrates a strong correlation with human judgment. One significant advantage of this approach is its ability to facilitate automated comparisons between different style transfer approaches, enabling a comprehensive analysis of their strengths and weaknesses.

Another evaluation procedure, proposed by Mao et al. [12], is known as Quantitative Evaluation (QTEV). It involves plotting the effectiveness, which measures the degree of style transfer, against coherence, which assesses the extent to which the transferred image retains the same object decomposition as the content image. This process generates an EC plot that aids in evaluating the performance of style transfer methods.

Controllable style transfer is an important research direction. Previous studies have proposed various evaluation methods to quantify the effectiveness of style transfer techniques to achieve controllable stylization. While existing methods have achieved specific results, there is room for improving style transfer quality evaluation. Future research can explore new metrics building on current ones. Moreover, combining quality evaluation with style transfer methods to achieve end-to-end quality optimization is also worth exploring. We look forward to seeing new breakthroughs in controllable and high-fidelity style transfer.

2.3 Developments of Style Transfer

A wide range of approaches characterizes recent research on Style Transfer, as theories and methods intersect effectively. Some Style Transfer methods incorporate other models, such as Stable Diffusion, while others explore combinations with other technologies.

InST [4] is an example of a one-shot stylized model that shares similarities with JoJoGAN but is based on Stable Diffusion. In the domain of 3D Reconstruction, PAniC-3D [5] and StyleRF [9] utilize their methods to integrate Style Transfer and 3D Reconstruction techniques. Furthermore, an exciting application of Style Transfer involves generating talking heads using portrait images, as demonstrated by MetaPortraits [6].

As technology and theoretical advancements continue, the utilization of Style Transfer is expected to become more extensive, and we can anticipate the emergence of new products and applications.

3 Method

We propose a novel method that incorporates a comprehensive evaluation of style transfer specifically designed for human faces. This approach aims to generate improved results by leveraging effective evaluation techniques. Our method combines multiple standards and technologies, including a unique evaluation metric called STE(Style Transfer Evaluation) which includes $CSCI$ and $CCVI$, along with a human-led aesthetic evaluation. Furthermore, our approach can accommodate various styles such as cartoon and anime, making it applicable to diverse studies.

As shown in Fig. 1 our approach consists of three stages. In the first stage, traditional StyleGAN models like DualStyleGAN are utilized to generate multiple batches of style images, which serve as inputs for the subsequent steps. The second stage employs Style Transfer Evaluation to perform preliminary screening, selecting the optimal results from each batch. Finally, in the third stage, an aesthetic evaluation is conducted to provide the final assessment and generate the output results.

3.1 Image Style Transfer

We are using DualStyleGAN as our basic model and we train it on its pre-train model. This method proposed a novel approach for high-resolution portrait style transfer training with a few hundred examples. The main contribution of this work is the characterization and disentanglement of facial identity versus artistic styles. To conclude, modeling the portrait synthesis with a dual style transfer process which can control both facial identity and style degree.

The model is trained using a dataset comprising exemplar artistic portraits and target faces. Given a target face, this model generates a random intrinsic style code (z) and an extrinsic style code (w). By applying style loss and identity

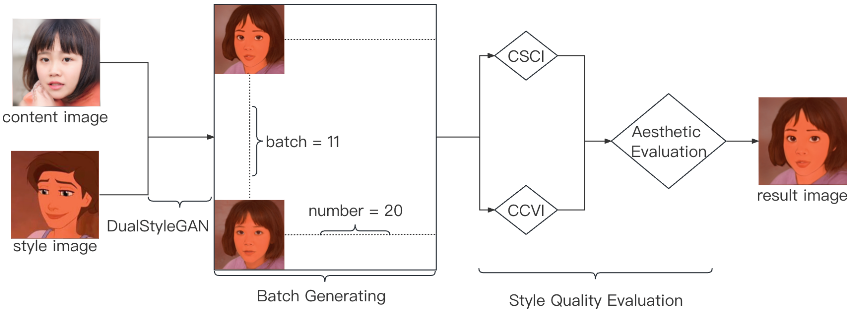

Fig. 1. The general workflow of Style Transfer Evaluation.

loss, it transfers the style of an exemplar artistic portrait onto the target face using the intrinsic and extrinsic style codes.

In the generative process, we set up a batch size and a range of weights. This process generates multiple batches of different results for the next stage of our work.

3.2 Style Transfer Evaluation

To evaluate the quality of style transfer, we employed the following methods. Firstly, we extracted the vector representation of the style and output image and encoded the images using advanced image coding techniques for subsequent analysis and calculations.

Next, we introduced the formula $CSCI$ (Combined Structural Content Similarity Index). This formula effectively combines two metrics, $SSIM$ (Structural Similarity Index) and $PSNR$ (Peak Signal-to-Noise Ratio), by utilizing a sliding window to traverse the images and calculate the similarity between the two images in terms of structure and content. The formulation of the $CSCI$ is as follows:

$$CSCI(\mathbf{s}, \mathbf{c}) = \frac{1}{N} \sum_{i=1}^{n} \frac{ssim(\mathbf{s}, \mathbf{c}) + psnr(\mathbf{s}, \mathbf{c})}{\mid ssim(\mathbf{s}, \mathbf{c}) \mid + \mid psnr(\mathbf{s}, \mathbf{c}) \mid} \tag{1}$$

Here, s and c are the style image and content image respectively, i represents the sliding window index, N denotes the total number of windows. The $CSCI$ score ranges from 0 to positive infinity, with 0 indicating perfect no structural and content similarity between the style and output images.

Indeed, the $CSCI$ formula comprehensively considers the structure and content information of the images, calculating the similarity based on weighted coefficients. Additionally, we introduced the $CCVI$ (Color Consistency Visual Index) formula, which evaluates the visual and color similarity between two images. By combining metrics such as Edge Preservation (EP), Color Preservation (CP), and Visual Significance (VS), the $CCVI$ formula accurately measures color retention and visual consistency. The formula for $CCVI$ is as follows:

$$CCVI = (\mathbf{s}, \mathbf{c}) = \frac{1}{N} \sum_{i=1}^{n} \frac{ep(\mathbf{s}, \mathbf{c}) + cp(\mathbf{s}, \mathbf{c}) + vs(\mathbf{s}, \mathbf{c})}{\mid ep(\mathbf{s}, \mathbf{c}) \mid + \mid cp(\mathbf{s}, \mathbf{c}) \mid + \mid vs(\mathbf{s}, \mathbf{c}) \mid} \tag{2}$$

Here, ep, cp, and vs represent edge preservation, color preservation, and visual significance, respectively. The values of these metrics are normalized to ensure they range between 0 and 1.

Based on the results of style transfer, we constructed two scoring sequences and assessed each output using the $CSCI$ and $CCVI$ formulas, while higher index values from the $CSCI$ and $CCVI$ formulas correlate with superior style transfer quality. Once the scoring was completed, we sorted the results in descending order, prioritizing the $CSCI$ score over the $CCVI$ score. Ultimately, the output with the highest combined score was selected as the best result for style transfer. This approach allowed us to consider multiple factors, including the structure, content, color, and visual aspects of the image, to evaluate the style transfer's quality comprehensively. By employing this scoring and evaluation method, we ensured a comprehensive assessment of style transfer quality and identified the most successful transfer result based on multiple criteria.

3.3 Aesthetics Evaluation

To balance artistry and aesthetics, we introduced a set of aesthetic criteria for evaluation. One such criterion is the "Three Courts and Five Eyes", which analyzes the proportion and distribution of facial features to define a standard face. However, it is essential to consider the diversity of face shapes, such as oval, Chinese, round, and others, to maintain individuality. We use the proportion characteristics of each face shape as quantitative evaluation criteria for face aesthetics. Based on these criteria, we design an aesthetic scoring function that incorporates indicators of naturalness, aesthetics, and artistry. This function comprehensively evaluates the style transfer outputs while considering the specific characteristics of different face shapes.

Based on the quantitative evaluation criteria for facial aesthetics, this paper aims to develop an adaptive collaborative exploration reinforcement learning model. This model is designed to function autonomously and self-learn the facial shape structure by considering facial aesthetics, as well as the naturalness, aesthetics, artistry, vividness, and emotional expression of the generated results. Specifically, the naturalness, aesthetics, and artistry indicators are reflected in the artistic style pen touch, picture cleanliness, and degree of defects, denoted as s_b, s_c, and s_a, respectively. These indicators are recorded and utilized in the learning process of the model. However, human aesthetic evaluation risks introducing cultural biases, so we give specific meanings for the degree.

The s_b indicator quantifies the uniqueness of the painting style in the generated portrait. It is computed as the distance between the style reference image and the content image, with a lower distance corresponding to a higher s_b score. This signifies how closely the synthesized portrait reflects the artistic style. The s_c indicator evaluates the capability of the generated portrait in suppressing low-quality effects like noise, blurring, and jagged edges. Moreover, the s_a indicator measures the degree to which discernible flaws or distortions in facial characteristics or outlines are averted. Integrating these indicators facilitates a quantitative appraisal of style representation, quality enhancement, and fidelity of the synthesized portraits.

4 Experiments

4.1 Datasets

In our experiments, we utilized three datasets to evaluate the performance of our method in cartoon stylization. For the Caricature dataset, we collected 199 images from WebCaricature, curated explicitly for studying face caricature synthesis. Additionally, we obtained an Anime dataset from Danbooru Portraits, consisting of 140 pairs of style-corresponding portraits. Furthermore, our cartoon stylization experiments involved a cartoon dataset comprising 317 cartoon face images sourced from Toonify [14]. These diverse datasets enable us to evaluate the effectiveness and versatility of our method across different stylization tasks, including sketch stylization, caricature synthesis, and cartoon stylization.

4.2 Implementation Details

We trained the DualStyleGAN model by fine-tuning its pre-trained model, adjusting the training iterations to 3000 with a batch size of 32. The training process took approximately 12 h and utilized two 3090 GPUs. During the stylized portrait generation stage, we fine-tuned the 18-bit weights and explored the vector space around the default value. This process generates 220 stylized results in 11 batches, with each batch containing 20 images.

We use $[n_1 * v_1, 1 * v_i, n_2 * v_2, n_3 * v_3]$ to indicate the vector w. the first n_1 weights in vector w are set to the value of v_1, the next one weights are set to the value of v_i which is changeable, the following n_2 is weighted as v_2, the last n_3 weights are v3. w_c, w_i, w_r and w_f denote the controllable weight vector (the first i weights of w), variable weight vector (the weights of i) ready weight vector(the weights from i +1 to 11) and the final weight vector(the last 9 weights) respectively. By default, we set w_c to 0.75, w_i is 0 to 1, the step size is 0.5, w_r to 0.75, w_f to 1. For testing for cartoon, caricature, anime, respectively.

After obtaining the results, we performed evaluations using the $CSCI$ and $CCVI$ metrics. The results were then sorted in descending order based on the evaluation scores. Subsequently, we conducted an aesthetic evaluation to assess the outputs further. Ultimately, we selected the result that ranked first in evaluation scores as the final output.

4.3 Comparable Experiments

We compare our method with JoJoGAN and DualStyleGAN in Cartoon, Anime, and Caricature styles. From Fig. 2 it can be seen that our method profitably cartoonized subjects with better results. Because of our evaluation process, the qualities of our results are better in color preservation, structure feature retention, and image fidelity. Specifically, our method can find the best results to preserve details such as hair, eyes, and mouth.

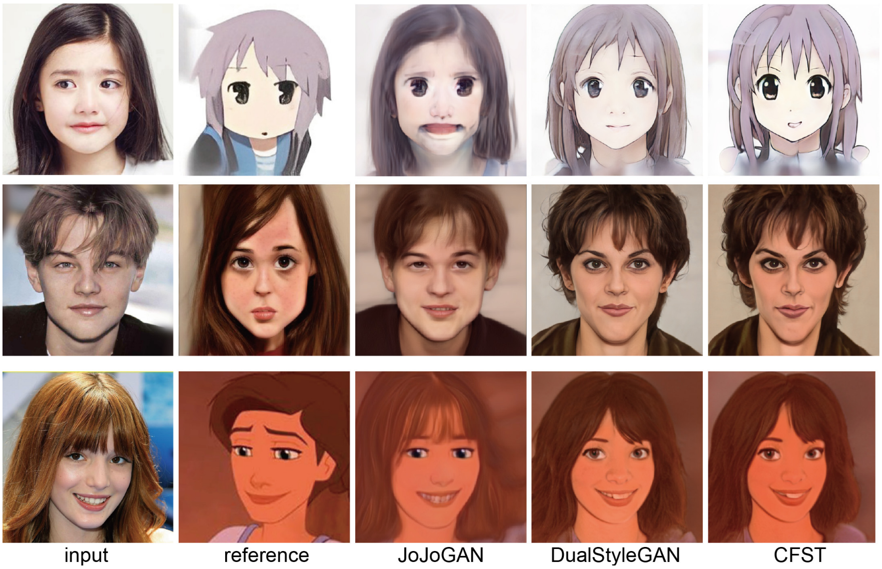

Fig. 2. Comparison with other methods

Table 1. Comparison of Style Transfer Methods

Style	Cartoon			Caricature			Anime		
Method	JoJoGAN	DualStyleGAN	Ours	JoJoGAN	DualStyleGAN	Ours	JoJoGAN	DualStyleGAN	Ours
$CSCI$	3.59	13.56	17.89	5.34	15.73	17.01	1.28	6.9	8.42
$CCVI$	0.37	0.78	0.86	0.22	0.67	0.72	0.23	0.56	0.63

Table 1 compares the $CSCI$ and $CCVI$ metrics achieved by different style transfer methods on three portrait styles. Our proposed method obtains the highest scores for both metrics across all styles compared to JoJoGAN and DualStyleGAN.

Our proposed method outperforms others in achieving controllable, high-fidelity artistic stylization on diverse datasets. The higher $CSCI$ and $CCVI$

values indicate that our method better preserves structural and content similarity as well as color consistency between the style image and the transferred result.

5 Conclusion

In conclusion, by utilizing DualStyleGAN with various weights, we can generate portraits. These generated portraits can then be ranked and sorted using our evaluation methods, namely the $CSCI$, $CCVI$, and aesthetics evaluation. Through this ranking process, we can optimize the original results and identify the most desirable outputs. This method holds potential for generating portraits using Style Transfer and filtering datasets, providing a valuable approach for enhancing the quality and selection of stylized portraits.

Future research will focus on enhancing evaluation formula robustness, model generalization, and style transfer for Asian faces. We will also try to refine the quality of non-realistic styles like anime, and explore model ensemble methods to leverage different algorithms' strengths. Moreover, it also can be used in Text-to-Image models like Stable Diffusion. Overall, future work will center on improving controllability and quality to achieve controllable and high-fidelity style transfer. Key directions include boosting quantification, expanding versatility, combining approaches, and enabling customization. We look forward to future innovations that will unlock the full potential of this technology.

Acknowledgements. We would like to thank the "Pioneer" and "Leading Goose" R&D Program of Zhejiang(No.2023C01212) as well as the Public Welfare Technology Application Research Project of Zhejiang (No.LGF22F020008).

References

1. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
2. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: Pastiche master: exemplar-based high-resolution portrait style transfer. In: CVPR (2022)
3. Wang, Z., et al.: Evaluate and improve the quality of neural style transfer. In: CVIU (2021)
4. Zhang, Y., et al.: Inversion-based style transfer with diffusion models. In: CVPR (2023)
5. Chen, S., et al.: PAniC-3D: stylized single-view 3D reconstruction from portraits of anime characters. In: CVPR (2023)
6. Zhang, B., et al.: MetaPortrait: identity-preserving talking head generation with fast personalized adaptation. In: CVPR (2023)
7. Chong, M.J., Forsyth, D.A.: JoJoGAN: one shot face stylization. arXiv preprint arXiv:2112.11641 (2021)
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
9. Liu, K., et al.: StyleRF: zero-shot 3D style transfer of neural radiance fields. In: CVPR (2023)

10. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: StyleCLIP: text-driven manipulation of StyleGAN imagery. In: ICCV (2021)
11. Wright, M., Ommer, B.: ArtFID: quantitative evaluation of neural style transfer. In: Andres, B., Bernard, F., Cremers, D., Frintrop, S., Goldlücke, B., Ihrke, I. (eds.) DAGM GCPR 2022. LNCS, vol. 13485, pp. 560–576. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16788-1_34
12. Yeh, M.-C., Tang, S., Bhattad, A., Forsyth, D.A.: Quantitative evaluation of style transfer. (2018). https://doi.org/10.48550/arXiv.1804.00118
13. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: VToonify: controllable high-resolution portrait video style transfer. In: ACM TOG (Proceedings of SIGGRAPH Asia) (2022)
14. Pinkney, J.N., Adler, D.: Resolution dependent GAN interpolation for controllable image synthesis between domains. arXiv preprint arXiv:2010.05334 (2020)