

Research Article

Video Style Transfer based on Convolutional Neural Networks

Sun Dong^{1,2}, Youdong Ding,² Yun Qian^{1,3}, and Mengfan Li¹

¹Pan Tianshou College of Architecture, Art and Design, Ningbo University, Ningbo 315211, China

²Shanghai Engineering Research Center of Motion Picture Special Effects, Shanghai 200072, China

³School of Fine Arts, Anhui Normal University, Wuhu 241002, China

Correspondence should be addressed to Yun Qian; qianyun@ahnu.edu.cn

Received 25 November 2021; Revised 16 February 2022; Accepted 28 February 2022; Published 25 March 2022

Academic Editor: Savvas A. Chatzichristofis

Copyright © 2022 Sun Dong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Video style transfer using convolutional neural networks (CNN), a method from the deep learning (DL) field, is described. The CNN model, the style transfer algorithm, and the video transfer process are presented first; then, the feasibility and validity of the proposed CNN-based video transfer method are estimated in a video style transfer experiment on *The Eyes of Van Gogh*. The experimental results show that the proposed approach not only yields video style transfer but also effectively eliminates flickering and other secondary problems in video style transfer.

1. Introduction

In the deep learning field, image style transfer is an important research topic [1]. Some traditional methods for style transfer include texture synthesis, support vector machines, histogram matching, and automatic sample collection [2–4]. Although special effects can be produced, image distortion as well as other prominent problems can also occur, such as the loss of detail, bending and deformation of straight lines, and color change over a large range. In addition, special algorithms are usually needed to further correct mistakes, resulting in low-style transfer efficiency and poor image quality. Recently, convolutional neural network DL models have been successfully applied to image style transfer problems, reigniting interest in this research field [5–8]. In the present study, a style transfer algorithm was developed and tested on *The Eyes of Van Gogh*, an American biography and feature film directed by Alexander Barnett, with the main roles played by Dane Agostini and John Alexander, which narrates the secret story of Van Gogh in St-Remy, Bedlam for 12 months; this film shows the legend of the talented artist who created, loved, and changed

the world through hallucinations, nightmares, and painful memories.

The purpose of this paper is that we tried adopting CNN based style transfer method for the video style transfer experiment. The foregoing CNN-based style transfer method is seldom used in the video field. The proposed style transfer algorithm uses techniques from the DL field and is based on a CNN; the feasibility and validity of the proposed CNN-based video style transfer algorithm are estimated in an experiment on the transfer of the painting style of Van Gogh's *The Starry Night* to the film special effects.

2. Methods

2.1. CNN. A CNN is a DL method developed recently that has attracted considerable attention. In general, a CNN is a multilayered network [9–11]; a typical CNN is shown schematically in Figure 1. A CNN consists of a series of convolution (C) and subsampling (S) layers. Each layer is composed of multiple 2D planes, with each serving as a feature map; the network also includes some fully connected (FC) hidden layers. There is only one input layer in a CNN.

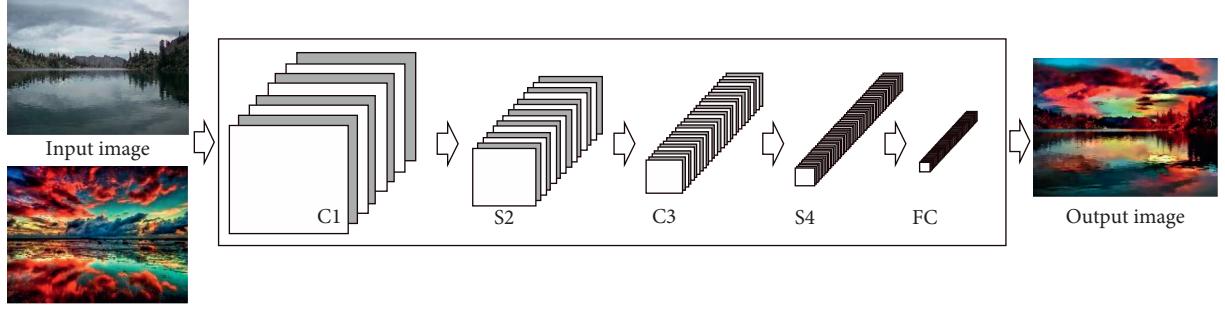


FIGURE 1: CNN structure.

This input layer receives two-dimensional objects directly, and the process of feature extraction into samples is performed by the convolution and sampling layers. Multiple fully connected hidden layers are used mostly used to accomplish specific tasks [12].

2.2. Style Transfer Algorithm. The methodology of Gatys et al. [13, 14] is reviewed. On this basis, the feature extraction and storage of style images and content images (single frames of video) are proposed. A style image \vec{a} is transmitted through the network (expressed as A^l), and the styles included in all layers are computed and stored ($A^l \in R^{N_l \times N_l}$). A content image \vec{P} is transmitted through the network (expressed as P^l) and stored in layer l ($P^l \in R^{N_l \times D_l}$). Therein, N_l represents the number of filters in that layer; and D_l is the spatial dimension of the feature map, namely the product of width and height. Then, a random white noise image \vec{x} is transmitted through the network; both the content feature F^l and the style feature G^l are computed. $F^l[\cdot] \in R^{N_l \times D_l}$ and F_{ij}^l are the activations of the i -th filter at j in layer l ; $G^l[\cdot] \in R^{N_l \times N_l}$ and G_{ij}^l are the vectorization results of layer l in i and j feature maps; the feature correlation is obtained as $G_{ij}^l = \sum_{k=1}^{D_l} F_{ik}^l F_{jk}^l$.

For each layer of the style image, the mean quadratic deviation of elements between G^l and A^l is computed, and the style loss $\mathcal{L}_{\text{style}}$ is computed using equation (1).

$$\mathcal{L}_{\text{style}}(\vec{a}, \vec{x}) = \sum_{l \in L_{\text{style}}} \frac{1}{N_l^2 D_l^2} \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2. \quad (1)$$

The mean quadratic deviation between F^l and P^l is computed, and the content image loss $\mathcal{L}_{\text{content}}$ is computed from equation (2).

$$\mathcal{L}_{\text{content}}(\vec{P}, \vec{x}) = \sum_{l \in L_{\text{content}}} \frac{1}{N_l D_l} \sum_{i,j} (F_{ij}^l - P_{ij}^l)^2. \quad (2)$$

The total loss $\mathcal{L}_{\text{singleframe}}$ is a linear combination of the style and content loss functions; it could propagate computation reversibly with errors, pertaining to the derivatives of pixel values; a gradient is used to iterate and upgrade the image \vec{x} until it matches the style feature of the style image \vec{a} and the content feature of the content image \vec{P} simultaneously; weight factors, including both α and β , determine

the importance of the two components, content and style, which is calculated from equation (3).

$$\mathcal{L}_{\text{singleframe}}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{\text{content}}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{\text{style}}(\vec{a}, \vec{x}). \quad (3)$$

2.3. Elimination of Flickering in Video Style Transfer. At present, most restoration methods require modeling to account for flickering in the image sequence; the flicker parameters of the model are first estimated, and then color correction and restoration are performed. However, the existing methods cannot treat flicker problems arising from video style transfer. A color transfer algorithm proposed by Reinhard et al. [15] is put forward in this paper, based on which the steps of video frame color correction and sequence restoration are simplified; the steps are specific to flickering after video style transfer, and the interframe color transfer is used directly. Thus, the data processing load and computational complexity are reduced, yet the restoration efficiency is increased. Video color transfer is an algorithm that changes the frame color [16]. A synthetic frame with the form of the original frame and reference frame color can be obtained by defining a reference frame that provides the original frame for both the structure and color layout, which is especially suitable for continuous video processing. The specific algorithm is as follows:

- (1) Both the original frame and the reference frame of the video are converted to the l, α , and β color space from the RGB color space, and the correlation between the two frames is removed.
- (2) The mean and the standard deviations of the original frame and the reference frame in each channel of the l, α , and β color space are computed, respectively. The mean values of the three channels of the original frame are m_S^l, m_S^α , and m_S^β ; the standard deviations of the original frame are $\sigma_S^l, \sigma_S^\alpha$, and σ_S^β ; the mean values of the reference frame are m_R^l, m_R^α , and m_R^β ; and the standard deviations of the reference frame are $\sigma_R^l, \sigma_R^\alpha$, and σ_R^β .
- (3) In accordance with equation (4), the overall color information of the mean value of all pixel values weakening the original frame is subtracted from all pixel values for each channel.

$$\begin{cases} l'_S = l_S - m_S^l, \\ \alpha'_S = \alpha_S - m_S^\alpha, \\ \beta'_S = \beta_S - m_S^\beta. \end{cases} \quad (4)$$

Here, l_S , α_S , and β_S are all pixel values for the three channels of the original frame, and l'_S , α'_S , and β'_S are the pixel values for the three channels of the original frame after weakening.

- (4) The standard deviation ratio of the original frame and reference frame is taken as the coefficient of channel value offset; the detailed information of the reference frame is mapped to the original frame in accordance with the following equation:

$$\begin{cases} l' = \frac{\sigma_R^l}{\sigma_S^l} * l'_S, \\ \alpha' = \frac{\sigma_R^\alpha}{\sigma_S^\alpha} * \alpha'_S, \\ \beta' = \frac{\sigma_R^\beta}{\sigma_S^\beta} * \beta'_S. \end{cases} \quad (5)$$

Here, l' , α' , and β' are all pixel values for the synthetic frame in the three channels of the l , α , and β color space.

- (5) The overall information about the reference frame is added to the synthetic frame; that is, this information is added to the mean value for each channel of the reference frame as shown in the following equation, and the final synthetic frame is thereby obtained.

$$\begin{cases} l = l' + m_R^l, \\ \alpha = \alpha' + m_R^\alpha, \\ \beta = \beta' + m_R^\beta. \end{cases} \quad (6)$$

Here, l , α , and β are all pixel values for the three channels finally obtained by the synthetic frame.

- (6) After color transfer, the synthetic frame from the l , α , and β color space is converted to the RGB color space.

2.4. Video Style Transfer Process

- (1) The video is converted into a single frame. Preprocessing should be performed on the transferred video; single-frame processing is performed for continuous videos, and the preprocessed video is saved as a JPG file. MATLAB software is applied for processing, and classified conservation is conducted based on the shot sequence.
- (2) *Video style transfer.* The style transfer algorithm is used to perform video frame style transfer using a CNN. The same group of shot circles is usually

selected in the transfer process to conduct the video style transfer experiment.

- (3) *Video color transfer.* Secondary flicker problems frequently occur in video style transfer. The essence of flicker is that adjacent frames vary significantly in brightness or hue, and the visual perception of flickering also appears when the video is played continuously. The flicker problem in video transfer is treated using the color transfer algorithm.
- (4) *Single-frame synthetic video.* After continuous video transfer and treatment of flickering, MATLAB is used for the single-frame synthesis of AVI-formatted videos to evaluate the results.

3. Experimental Work

3.1. Model Parameters. Model selection and parameter optimization are key steps in video style transfer, and a proper model and parameters can significantly enhance the transfer of high-quality artistic videos. First, four artificial intelligence models, namely CaffeNet, GoogLeNet, VGG16, and VGG19, were selected, all of which have their own unique advantages [17–20]. CaffeNet is a classical DL model; its advantages include network expansion and the ability to solve fitting problems. It is also the simplest network among the four models; since these models have been proposed, several deeper network structures have been proposed. GoogLeNet utilizes the concept of an inception module, aiming at strengthening the function of basic feature extraction modules. It considerably enhances the feature extraction ability of a single layer, but does not significantly increase the computed amount. Although VGG-Net has inherited some network frameworks from LeNet and AlexNet, the former is not identical to the latter ones; VGG-Net uses more layers, usually from 16 to 19. VGG-Net mainly increases the network structure depth, while reducing parameter configuration. The model suitable for the style transfer of the video *The Eyes of Van Gogh* must be analyzed in advance. Second, the style/content conversion rates (10^{-1} , 10^{-2} , 10^{-3} , 10^{-4} , and 10^{-5}) are key parameters for transfer, and a preexperiment is also required for parameter selection. Therefore, the video clip of *The Eyes of Van Gogh* was selected for the style transfer experimental analysis.

Figure 2 shows the experimental results obtained when the style/content conversion rates of the four models were set to 10^{-1} , 10^{-2} , and 10^{-3} ; the video frame style transfer was not sufficient because the images in the original video remained, owing to the excessively low conversion rate. When the conversion rate was 10^{-5} , the style transfer was excessive and some important information, such as the form and structure of the original video frame, was lost. The CaffeNet-based style transfer revealed several serious errors, such as distortion, making it not suitable for this video transfer; although GoogLeNet performed slightly better than CaffeNet, there were still many errors; the performances of VGG16 and VGG19 were better, and these methods achieved optimal results, especially for the conversion rate of 10^{-4} . The experimental results for the VGG16/19 model were further compared, and the results of the VGG19-based transfer were considered to have higher fidelity.



FIGURE 2: Comparison of models and parameters.

more plentiful layers, and a high transfer efficiency (marked by a red frame in Figure 2). Figure 3 shows the computational times for the different models and parameters.

As a result, the VGG19 model and the style/content conversion rate of 10^{-4} were selected for the style transfer in the video transfer experiment on *The Eyes of Van Gogh*.

3.2. Video Style Transfer Experiment. Based on the CNN [21–24], the style transfer algorithm based on the Caffe platform was used to input the video frame to be transferred. The VGG-19 network trained in advance was used for computing the loss; conv4_2 in this network represents the content; conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1 represent the style; the loss of the parameter weight used was not more than 0.02%; and the number of iterations was 512. The hardware device used was a high-performance workstation (HPZ840 workstation, parameter configuration: Intel Xeon E5 Eight-core, two GPUs, Nvidia TITAN Xp. memory: 64 GB). The following experimental steps were performed: first, we imported 24 frames (800×480 per frame in size) at a time for continuous style transfer. The model was VGG19, and the

conversion rate was 10^{-4} ; second, Van Gogh's representative work, *The Starry Night*, was selected as the source style image; its short yet thick brushwork and fiery color are filled with personality characteristics and artistic charm; finally, CUDA was used for parallel computing and the outputting of the JPG video frame (the maximal length was 1024).

Figure 4 shows the results of the style transfer experiment for a continuous video. Figures 4(a) and 4(c) were selected from the original frames in the experimental video *The Eyes of Van Gogh*; we chose two groups of shots, with low indoor brightness and high outdoor brightness, respectively, for the video style transfer experiment. This group of video shots was relatively fixed, and the characters had no large displacements. Figures 4(b) and 4(d) show the realization of the video style transfer using equations (1)–(3) that were presented in Section 2 of this paper. The experimental results show that, on the one hand, the style transfer retains the form structure of the original video frame and other information, and when mixed with the brushwork texture and color elements of *The Starry Night*, their combination produces a unique visual effect; on the other hand, the video frame details obtained in the style transfer are excellent with rich colors, without any

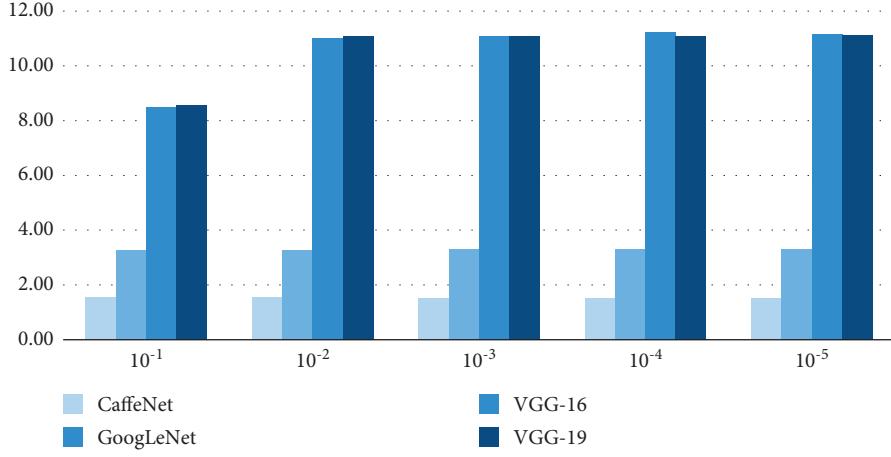


FIGURE 3: Comparison of time spent.

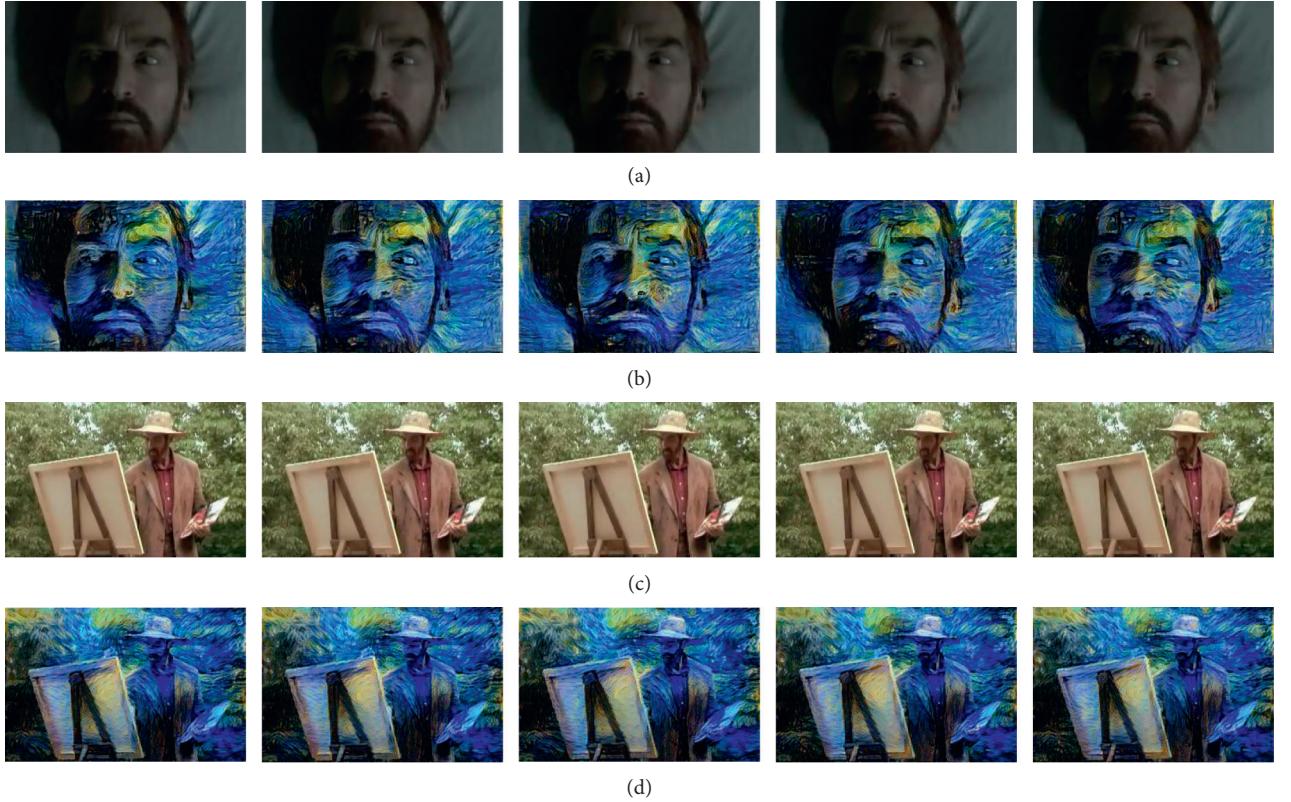


FIGURE 4: Result of the video style transfer experiment (a-d).

style transfer errors such as pseudoscopic images or fuzziness, and without interframe flickering.

Considering the real and effective experimental results, we selected a set of continuous frames in which the characters had large deviations from the video for the style transfer experiment, to estimate the reliability and validity of the style transfer algorithm for video style transfer applications. The experimental steps and model parameters were the same as those mentioned previously.

Figure 5 shows the results of this video style transfer. Although the form structure information, mixture with

textures and colors of the source style image, and other elements of the target video frame were reserved to produce visual effects, we also noted some mistakes in the details of the video style transfer. A prominent problem was inter-frame flickering, as shown in the part marked with a red frame in Figure 5(b). Therein, some parts of several single frames exhibited hue and brightness deviations, which caused flickering as secondary damage during continuous play. Thus, we used the color transfer algorithm to further process and eliminate flickering, thereby attaining optimal video transfer.

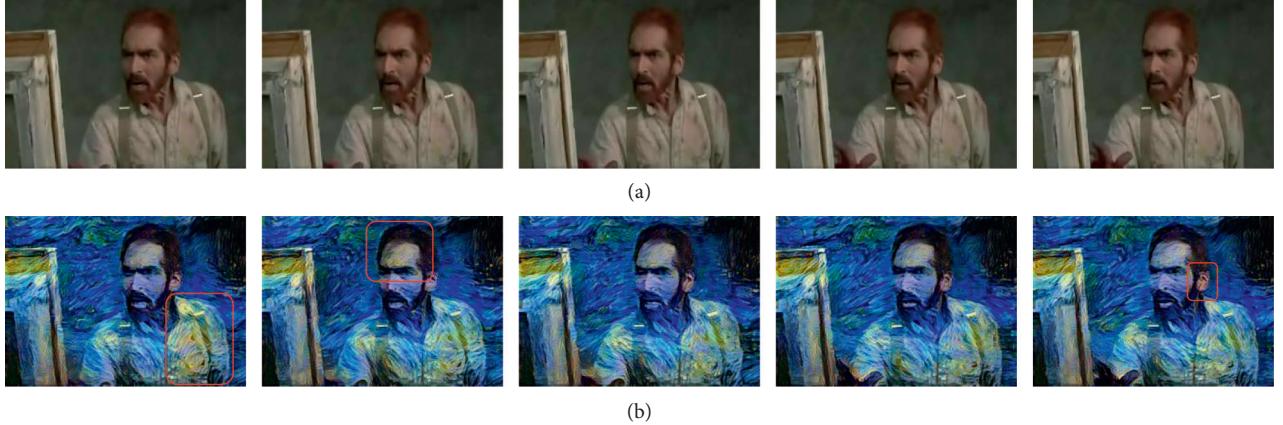


FIGURE 5: Results of the video style transfer experiment (a, b).

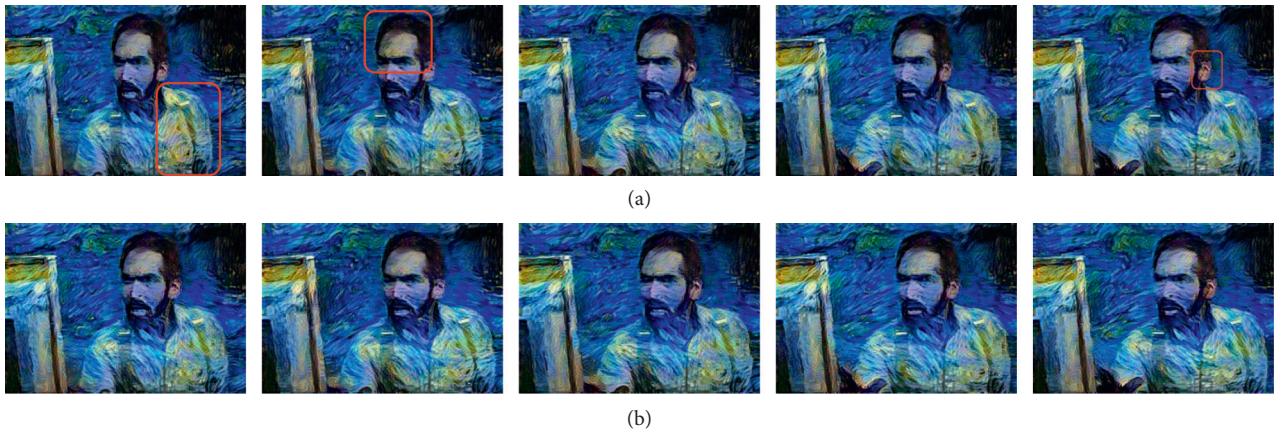


FIGURE 6: Results of the video flickering elimination experiment: (a) before and (b) after.

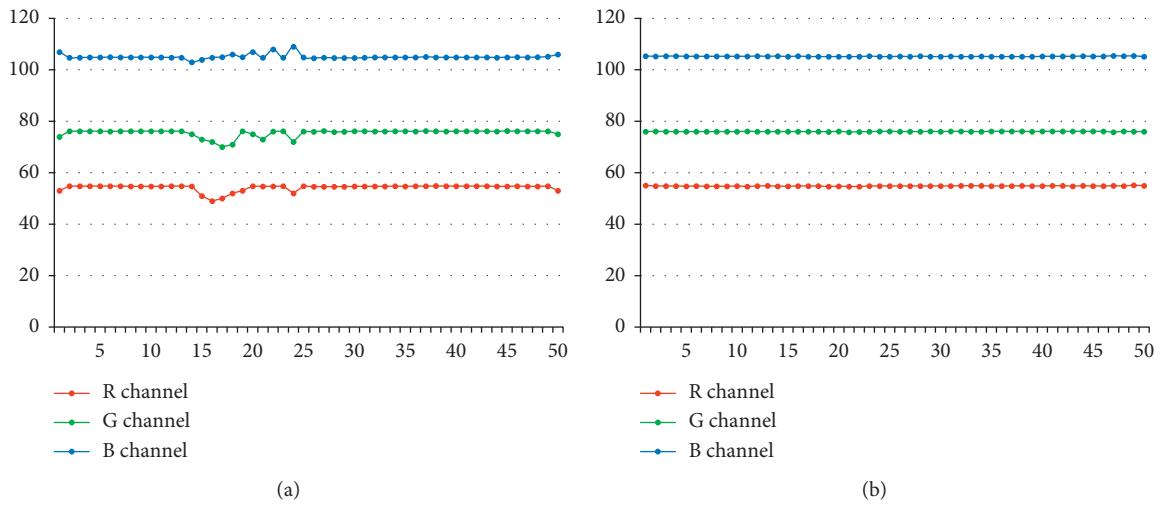


FIGURE 7: Comparison of the mean statistics of videos: (a) before and (b) after.

3.3. Elimination of Flickering. The shooting process was performed for various videos imported in accordance with equations (4)–(6) that were presented in Section 2 of this

paper. A proper frame in the same shot was selected as a reference frame (here, we chose the middle frame as the reference frame), and the color feature of the reference frame



FIGURE 8: Results of the video style transfer experiment.

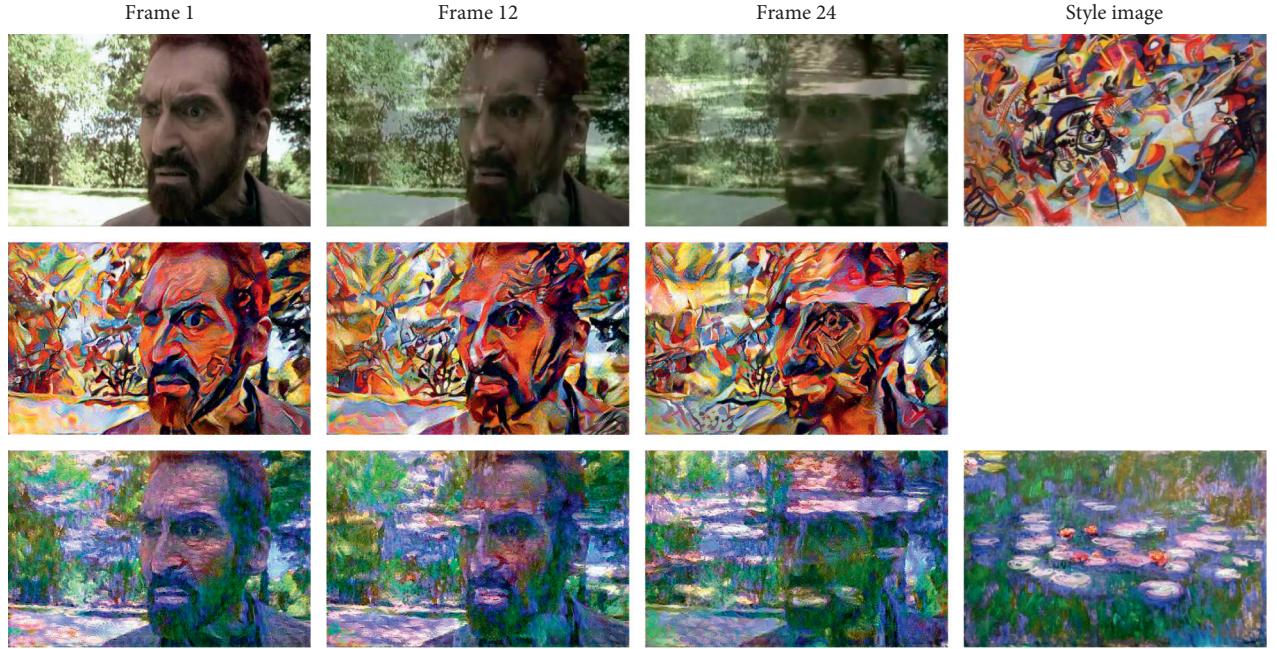


FIGURE 9: Result of different video style transfer experiment.

was transferred to each frame in the same group of shots successively; after the video processing of the same group, the above steps were repeated until all of the frames imported into the video were processed. Figure 6(a) shows the video frames after the elimination of flickering using the color transfer algorithm; we observe that compared with the areas marked with red frames in Figure 6(a) (the character's chest, forehead, and other body parts), the style transfer errors were effectively eliminated.

Figure 7 shows the mean statistics of videos before and after the elimination of flickering; Figure 7(a) shows the statistics before the flickering elimination process, and

Figure 7(b) depicts the statistics after the flickering elimination process.

Figure 8 shows a scene from *The Eyes of Van Gogh* where two people walk through the scene. Without the color transfer algorithm, the stylized videos demonstrate flickering between adjacent frames after the people pass by. Figure 9 shows another scene from *The Eyes of Van Gogh* with fast camera motion. The color transfer algorithm eliminated the interframe flickering. The experimental results show that the color transfer algorithm can effectively eliminate secondary flickering arising from video style transfer, and the resulting video is full of colors and exhibits a uniform hue.

4. Conclusion

The CNN-based style transfer algorithm quickly and effectively generates diverse and stylized videos, as well as unique visual effects. The experiment proved that the video style transfer method proposed herein is feasible and effective. In terms of parameter optimization of the video style transfer model, we found that the style transfer results are strongly determined by the style/content conversion rate and model selection. The experiment also showed that for the film, *The Eyes of Van Gogh*, the optimal model was VGG19 and the optimal conversion rate was 10^{-4} . It should be noted that model parameters should have been selected in combination with different videos; a sample analysis experiment should be conducted in advance to obtain the best results. In addition, as flickering and other secondary problems often occur in video style transfers, the video after style transfer requires further processing using the color transfer algorithm to obtain high-quality experimental results.

In future work, we hope to explore the use of the proposed CNN-based style transfer algorithm for other video transformation tasks, such as the production of stable and visually appealing stylized videos even in the presence of fast motion and strong occlusion. Owing to the subjectivity of video quality evaluation, we also plan to establish a subjective evaluation index system for better evaluation of style transfer video quality. Video style transfer is a common problem like loss of details, bending and deformation, or color change over a large range, which is to cause secondary video damage like a flicker. Subjective evaluation is the most commonly used method in video quality evaluation. However, subjective evaluation is a time-consuming task. For this, we plan to employ a forced-choice evaluation on Amazon Mechanical Turk (AMT) with 200 different users to evaluate our experimental results. This is a part of further research. In addition, we plan to extend the dataset to include more videos, which would make our approach more generalizable.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

All authors declare that there are no conflicts of interest with this study.

Acknowledgments

This research was supported by the National Natural Science Foundation of China [Grant no.61402278, 61303093], the Teaching and Research Project of Ningbo University [Grant no. JYXMXZD2022019], the Social Science Foundation of Anhui Province [Grant no. AHSKY2018D74], and the Outstanding Young Talents Foundation by the Ministry of Education of Anhui Province [Grant no. gxyq2018002].

References

- [1] M. Elad and P. Milanfar, "Style transfer via texture synthesis," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2338–2351, 2017.
- [2] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*, pp. 694–711, Amsterdam, The Netherlands, October 2016.
- [3] J. Liao, Y. Yao, L. Yuan, H. Gang, and B. K. Sing, "Visual attribute transfer through deep image analogy," *ACM Transactions on Graphics*, vol. 36, 2017.
- [4] T. Dutta and H. P. Gupta, "Leveraging smart devices for automatic mood-transferring in real-time oil painting," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 2, pp. 1581–1588, 2017.
- [5] H. S. Faridul, T. Pouli, C. Chamaret et al., "Colour mapping: a review of recent methods, extensions and applications," *Computer Graphics Forum*, vol. 35, 2016.
- [6] W. H. Bangyal, R. Qasim, Z. Ahmad et al., "Detection of fake news text classification on COVID-19 using deep learning approaches," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 5514220, 14 pages, 2021.
- [7] W. H. Bangyal, A. Hameed, J. Ahmad, and R. Etengu, "New modified controlled bat algorithm for numerical optimization problem," *Computers, Materials & Continua*, vol. 70, pp. 2241–2259, 2021.
- [8] W. H. Bangyal, K. Nisar, A. A. B. Ibrahim, M. R. Haque, J. J. P. C. Rodrigues, and D. B. Rawat, "Comparative analysis of low discrepancy sequence-based initialization approaches using population-based algorithms for solving the global optimization problems," *Applied Sciences*, vol. 11, no. 16, Article ID 7591, 2021.
- [9] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep photo style transfer," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, July 2017.
- [10] Y. Nikulin and R. Novak, "Exploring the neural algorithm of artistic style," 2016, <https://arxiv.org/abs/1602.07188>.
- [11] S. Hicsonmez, N. Samet, F. Sener, and P. Duygulu, "DRAW: deep networks for recognizing styles of artists who illustrate children's books," in *Proceedings of the 2017 ACM on international conference on multimedia*, Yokohama Japan, June 2017.
- [12] Y. Jing, Y. Yang, Z. Feng, Y. Yizhou, and S. Mingli, "Neural style transfer: a review," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, 2017.
- [13] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.
- [14] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proceedings of the Computer Vision and Pattern Recognition*, Hannover, Germany, September 2016.
- [15] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer Graphics and Applications*, vol. 21, 2001.
- [16] X. Huang, Y. D. Ding, and B. Wu, "Implementation and design of old film global flicker restoration system," *Video Engineering*, vol. 40, no. 12, pp. 125–129, 2016.
- [17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Vision and Pattern Recognition*, 2014.
- [18] M. S. Ryoo and L. Matthies, "Video-based convolutional neural networks for activity recognition from robot-centric

- videos,” in *Proceedings of the SPIE Defense Security. International Society for Optics and Photonics*, pp. 98370R–98376R, Bellingham, WA, USA, September 2016.
- [19] A. Zhai, D. Kislyuk, Y. Jing et al., “Visual discovery at pin-interest,” in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 515–524, Perth, Australia, April 2017.
 - [20] F. Okura, K. Vanhoey, A. Bousseau, A. A. Efros, and G. Drettakis, “Unifying color and texture transfer for predictive appearance manipulation,” *Computer Graphics Forum*, vol. 34, no. 4, pp. 53–63, 2015.
 - [21] D. Zheng, “A novel method for fabric color transfer,” *Color Research & Application*, vol. 40, no. 3, pp. 304–310, 2015.
 - [22] O. Frigo, N. Sabater, J. Delon, and H. Pierre, “Video style transfer by consistent adaptive patch sampling,” *The Visual Computer*, vol. 35, 2019.
 - [23] Y. Liu, M. Cohen, M. Uyttendaele, and S. Rusinkiewicz, “AutoStyle: automatic style transfer from image collections to users’ images,” *Computer Graphics Forum*, vol. 33, no. 4, pp. 21–31, 2014.
 - [24] A. Gupta, J. Johnson, A. Alahi, and L. Fei Fei, “Characterizing and improving stability in neural style transfer,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.