applied sciences

MDPI

Article

Fast Coherent Video Style Transfer via Flow Errors Reduction

Li Wang [1,*], Xiaosong Yang [2] and Jianjun Zhang [2]

[1] School of Computer and Data Engineering, NingboTech University, Ningbo 315100, China
[2] National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, UK; xyang@bournemouth.ac.uk (X.Y.); jzhang@bournemouth.ac.uk (J.Z.)
* Correspondence: liwang@nbt.edu.cn

Abstract: For video style transfer, naively applying still image techniques to process a video frame-by-frame independently often causes flickering artefacts. Some works adopt optical flow into the design of temporal constraint loss to secure temporal consistency. However, these works still suffer from incoherence (including ghosting artefacts) where large motions or occlusions occur, as optical flow fails to detect the boundaries of objects accurately. To address this problem, we propose a novel framework which consists of the following two stages: (1) creating new initialization images from proposed mask techniques, which are able to significantly reduce the flow errors; (2) process these initialized images iteratively with proposed losses to obtain stylized videos which are free of artefacts, which also increases the speed from over 3 min per frame to less than 2 s per frame for the gradient-based optimization methods. To be specific, we propose a multi-scale mask fusion scheme to reduce untraceable flow errors, and obtain an incremental mask to reduce ghosting artefacts. In addition, a multi-frame mask fusion scheme is designed to reduce traceable flow errors. In our proposed losses, the Sharpness Losses are used to deal with the potential image blurriness artefacts over long-range frames, and the Coherent Losses are performed to restrict the temporal consistency at both the multi-frame RGB level and Feature level. Overall, our approach produces stable video stylization outputs even in large motion or occlusion scenarios. The experiments demonstrate that the proposed method outperforms the state-of-the-art video style transfer methods qualitatively and quantitatively on the MPI Sintel dataset.

Keywords: style transfer; video stylization; video stabilization; deep networks

Citation: Wang, L.; Yang, X.; Zhang, J. Fast Coherent Video Style Transfer via Flow Errors Reduction. *Appl. Sci.* **2024**, *14*, 2630. https://doi.org/10.3390/app14062630

Academic Editor: Shiyang Yan

Received: 8 January 2024
Revised: 14 March 2024
Accepted: 16 March 2024
Published: 21 March 2024

1. Introduction

Due to advances in digital technology, it is becoming easier for the public to produce professional multimedia content, such as photos, videos and digital arts, which were once limited by specialists' knowledge and skills. For example, image processing techniques like bilateral filter [1] and style transfer methods like StyleGAN [2] are used to create artistic videos with cartoon-like effects [3,4], and superpixel techniques for image segmentation [5] are applied to produce wonderful photos with stylistic brushes [6]. In addition, there are extended techniques like those in [7,8] which aim to create photorealistic results with reference styles and translate them within image domains. In this work, we pay attention to artistic video stylization which extends artistic style transfer from static images to video applications.

Recently, the success of artistic style transfer for still images [9] has inspired a surge in works [10–17] which tackle the style transfer problem and style classification [18–20] task based on deep correlation features. In their seminal work on artistic style transfer, Gatys et al. [9] sought to transform the artistic style of a painting into another photorealistic image by formulating the task into a gradient-based optimization problem. Starting with random white noise, a new image is evolved to present similar spatial structures to a content image and have stylistic feature correlations with a painting image. The stylized outputs are impressive but the heavy optimization process leads to a slow performance in run time.

To address this issue, Johnson et al. [10] presented a speed-up solution by introducing an offline feed-forward network. Chen et al. [14] proposed another feed-forward network which swaps arbitrary styles between content images and also gives pleasing results. This work introduces a patch-based matching technique that replaces the content image with the style image patch-by-patch through neural activations. Huang et al. [13] proposed to replace Chen et al.'s style swap layer with an adaptive instance normalization layer, which is capable of transferring arbitrary artistic style in real-time.

These aforementioned methods have achieved great success in transforming static photorealistic images into artistic styles. However, they fail in terms of the video style transfer application, which aims to transfer the artistic style of a reference image into a video. For example, processing an input video sequence via per-frame stylization using methods (e.g., [10,13]) often leads to flickering and incoherence problems (e.g., inconsistent texture patterns and colours) between adjacent outputs. To obtain coherent video style transfer, the stylized videos should preserve the consistent temporal features (including artistic texture patterns and colours) between consecutive transferred frames. The reasons behind the temporal incoherence problem may vary with different types of artistic style transfer methods. For gradient-based optimization methods (e.g., [9]), their random initialization and non-convex nature leads to the local minima of the style loss, which causes unstable transferred texture appearances in consecutive frames, and to some extent, ghosting artefacts. For methods based on feed-forward networks [10–14], slight changes in illumination and movements in adjacent frames cause large variations in the stylized results. Therefore, the temporal consistency of consecutive frames in video processing techniques (e.g., [21,22]) should be considered for video stylization.

Within the literature, there are mainly two branches of methods focusing on coherent video style transfer. One of them is using optical flow estimation for temporal consistency, such as the early works [23–25] which built on Gatys' optimization networks as well as recent works [26–34] which have built on Johnson's feed-forward networks. To solve the temporal incoherence problem, the common idea behind these methods is to obtain warped initial images with optical flow estimation, then stylizes them frame-by-frame with learned temporal constraints. For example, Ruder et al. [30] presented an approach based on a feed-forward network, which takes per-frame processed frames as additional inputs and trains their network with incorporating optical flow into temporal consistency loss. Their method needs optical flow estimation in both the training and test stages. However, when motions (or occlusions) are too large, these feed-forward based methods (e.g., [29,30]) still have incoherence problems due to flow errors as optical flow fails to track the boundaries of objects correctly. Methods from the other branch tackle the temporal incoherence problem without optical flow estimation, and instead they try to learn temporal features directly from input videos before combining them with learned style representations to stylize video frames. For example, Li et al. [35] discovered that the normalized affinity of generated features is the same as that for content features, thus it is suitable for obtaining stable frame-based video style transfers without optical flow estimation, as the normalized affinity of stylized features naturally contains temporal information like content features do. Wang et al. [36] and Wu et al. [37] proposed relaxed constraints for their objective functions for feature-level patterns to achieve temporal consistency because they reduce the effect of local noise from illumination and movements on variations in the stylized results. More recently, Gu et al. [38] proposed a joint learning framework for both image and video domains using a transformer network, which involves a domain interaction transformer to make video style transfer more efficient computationally and propagates the temporal information from input video sequences to output video frames. However, these methods from the second branch achieve temporal consistency by sacrificing faithful style transformation on videos, as their learned temporal information is focused on content structures between consecutive frames but not the stylistic effects of artistic images (including texture patterns and colours), and their relaxed constraints on style transformation leads to unfaithfully stylized results.

It is difficult for previous methods to handle the balance of style effects and temporal consistency. The methods from the first branch have the advantage of preserving artistic styles well but have less temporal consistency as optical flow methods fail to detect the boundaries of objects accurately, while the approaches from the second branch are able to keep the temporal consistency but there are unfaithful style effects in outputs as their relaxed constraints on style transformation lead to unfaithfully stylized results. To obtain the balance, we revisit the gradient-based optimization method due to its nature of effectively preserving style effects and easily obtaining temporal consistency with extra relaxed constraints. In this work, we propose a novel framework to perform video style transfer from a new perspective which consists of the following two stages: 1. new mask techniques are proposed to create new initialization images; 2. these initialized images are iteratively processed with proposed losses to obtain stable stylized videos. To be specific, our approach proposes a set of new mask techniques such as a multi-scale scheme, incremental mask and multi-frame mask fusion to prevent the temporal incoherence problem (including ghosting artefacts). And we are able to significantly reduce flow errors even for large motion or strong occlusion cases. Each initialized image obtained via our mask techniques needs much fewer iterations to produce a consistent style transformation over a certain number of frames, while at the same time the speed increases from over 3 min per frame to less than 2 s per frame. In addition, we discover that taking into account both the multi-frame RGB-level and Feature-level Coherent Losses will result in a better temporal consistency. During the experiments, we also discovered that the image quality of initialized images degrades over long-range frames. To retain the quality, we propose Sharpness Losses to deal with potential image blurriness artefacts.

Overall, the main contributions of the approach proposed in this paper are the following:

1. Novel mask techniques for new initialized images, which are capable of reducing flow errors even for large motions or strong occlusion cases.
2. Extra constraints like Coherent Losses and Sharpness Losses, which help to obtain a better temporal consistency and retain image quality over long-range frames.
3. The speed of the gradient-based optimization methods is increased from minutes per frame to less than 2 s per frame.

In this way, our gradient-based optimization method produces coherent video outputs which preserve the faithful style effects and the temporal consistency of the original images as well even for large motion or strong occlusion cases.

The organization of this paper is as follows: Section 2 will give a review of the literature on image style transfer and video style transfer; Section 3 will illustrate the motivation behind this work, the proposed system overview, the novel initialization via proposed mask techniques, loss functions for retaining image quality and loss functions on both RGB-level and Feature-level constraints for temporal consistency; Section 4 will give the implementation details for reproducibility; Section 5 will demonstrate ablation studies of the proposed components and comparisons to state-of-the-art methods with qualitative and quantitative evaluations; and Section 6 will offer some conclusions and ideas for further works.

2. Related Work

In this section, we will give a detailed literature review of some representative approaches for image style transfer and video style transfer, which illustrates their methodologies in brief and points out their limitations. To help readers to clearly understand the research history and relationships between the studies presented here, this section will be followed first by Section 2.1 Image Style Transfer and then by Section 2.2 Video Style Transfer.

2.1. Image Style Transfer

In their exploration of deep neural networks, Gatys et al. [39] qwew the first to show that the feature maps learnt by hidden units can be applied to synthesize texture when fusing the feature correlations of an image. Based on this, Gatys et al. [40] extended their method to image transformation task, which sought to generate a new image that succeeds in preserving the spatial structure content from one real world photograph and while preserving the artistic style (including texture, tone and colour) of a painting. The appealing visual appealing effects attracted a number of subsequent works [10–14,25,35,41–49] on artistic style transfer for still images based on deep neural networks.

The initial approach proposed by Gatys et al. [9] used the network for image classification purposes, like VGG-19 [50], to complete the style transformation task. They introduced two loss functions and formulated an optimization problem. The optimization process generates a new image from white noise which contains both a similar spatial structure to the content image and similar stylistic features to the style image. This online gradient-based optimization method is burdened by a heavy iteration computation process which leads to a very slow execution time. To address this problem, Johnson et al. [10] proposed a fast offline network to approximate the optimum of Gatys et al.'s loss functions. They used perceptual loss functions [10] to train feed-forward convolutional neural networks [51,52] and complete image transformation task in real time. Inspired by this successful approach, techniques (e.g., [12,13]) based on feed-forward networks have been adopted by popular APPs such as Prisma and DeepArt. Recently, the following works [53–59] introduced the self-attention mechanism to an encode–transfer–decode framework for better style transfer. To date, researchers [16,34] have attempted to perform image style transfer via diffusion models with a style prompt. However, such an attention mechanism needs much more computational resources to complete the task, which is not practical in real-world applications.

2.2. Video Style Transfer

There are two main branches of approaches proposed for video style transfer, and they are categorised into two types: video style transfer with optical flow estimation and video style transfer without optical flow estimation. Early works [23,24] firstly introduced optical flow estimation to enhance the temporal consistency between adjacent stylized frames, which were built on original gradient-based optimization methods (e.g., [9]). However, such approaches are less practical as they need minutes to stylize one single frame. To speed up the style transformation process, more works such as [27–29,60] proposed feed-forward-based video style transfer methods which use optical flow estimation during the training stage but can stylize videos without them in the test stage. These methods leverage a running performance of up to three orders of magnitude faster speed while obtaining coherent stylized videos.

However, the methods mentioned above all need optical flow with extra complex computation, which makes it impractical to stylize high-resolution or long videos. To avoid such extra computation, Lai et al. [61] proposed a blind video post-processing technique which is capable of preserving temporal consistency. Their method processes the per-frame stylized results and achieves a real-time execution time as no optical flow is needed in the test stage, but their results still suffer from temporal incoherences as no prior knowledge about style effects is considered. Li et al. [35] proposed a new linear transformation matrix to minimize the difference between the covariance of content features and style features, which maintains temporal consistency by propagating the computed matrix from the beginning frame to the rest of them. Wang et al. [36] proposed a novel method to make the stylization loss term more robust to motions without optical flow estimation. Deng et al. [62] proposed a multi-channel correlation network which is able to fuse style features and input content features for maintaining the coherence of input videos and stylized videos. More recently, Wu et al. [37] proposed a universal versatile style transfer method with a novel Contrastive Coherent Preserving Loss, which is capable of preserving the coherence of the content source but without degrading the stylization performance. Gu et al. [38] proposed

a Unified Style Transfer framework with a domain interaction transformer, which enables temporal information from input videos with rich texture appearance to be integrated with the style reference image for stylized videos. However, these methods without optical flow estimation fail to find the balance between stylization effect and temporal consistency.

3. Methods

To better understand the proposed methodology, this section will present the method with two parts: the concept behind the idea and details of the proposed framework. The first part aims to describe how the idea for this paper appeared, what problems we will deal with when following the idea to process video style transfer, and what potential solutions we can propose based on the discoveries behind problems. The second part aims to illustrate the proposed framework from top levels (e.g., the outline of framework and architecture of proposed network) to bottom levels (e.g., initialization with proposed mask techniques and loss functions). Specifically, Section 3.1 describes the idea for the composition of stylized frames, the intuition behind the mask designs, and points out the problems (e.g., texture discontinuity and image blurriness artefacts) to be solved and the potential solutions in this work; Section 3.2 illustrates the proposed framework outline and the architecture of the proposed gradient-based optimization network; Section 3.3 demonstrates the new initialization with the proposed mask techniques (for texture discontinuity); Sections 3.4 and 3.5 give details about the loss functions including Sharpness Losses (for image blurriness artefacts) and Coherent Losses (for enhanced temporal consistency).

3.1. Motivation

For artistic video stylization, the unexpected flickering problem hinders satisfactory results when still image style transfer methods (e.g., [10,13]) are applied to process frames independently. As shown in Figure 1, the adjacent per-frame stylized results exhibit some colour and texture incoherence (e.g., middle columns in zoom-ins). To preserve the coherency for video style transfer, this work starts with a straightforward idea.

To simplify, we start with two consecutive original video frames f_v^{t-1} and f_v^t, and their corresponding per-frame stylized results f_s^{t-1} and f_s^t, and their corresponding optical flow F^t from f_v^t to f_v^{t-1}; then, we produce a warped image $w^t = \mathcal{W}(f_s^{t-1}, F^t)$ by warping f_s^{t-1} with F^t, and a mask M^t containing per-pixel flow traceable (e.g., values tend to be 1) and untraceable regions (e.g., values tend to be 0). (A warp operation with optical flow means that pixels in the image will be moved to the locations in the next frame by multiplying pixel-level motion vectors. The flow contains these pixel-level motion vectors.) To obtain a stable consecutive stylized result, a straightforward idea which came up was to compose the warped image w^t and f_s^t in the traceable and untraceable flow regions, respectively. In this way, the we are capable of preserving the coherency of the composition result as much as possible. However, there is one prerequisite that the pasted contents at untraceable regions must have the exact original content details especially in occlusion regions. Otherwise, a heavy image compensation/optimization process is needed during video stabilization. Fortunately, artistic style transfer methods for still images (e.g., [10]) satisfy this prerequisite as they may change the textures or colours in the occluded regions but they indeed do not damage the consistency of the original content's details. For example, in Figure 2, compared to the original frame f_v^t, the content structures (in red and orange rectangles) of the composition result preserve the consistency better than the warped image w^t, as the warped image w^t has unexpected content structures in the red box and duplicated ones in the orange box. Hence, this straightforward idea is worth carefully investigating to obtain stabilized video outputs.

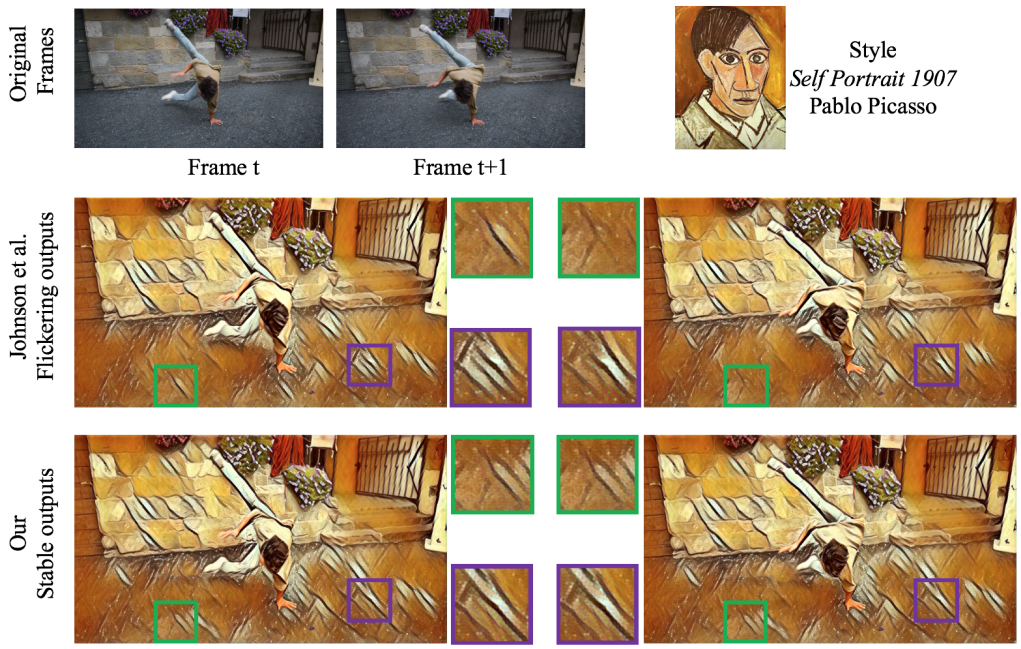

Figure 1. Flickering artefacts in video style transfer. The first row shows two original consecutive video frames (**left**) and the style image (**right**). The second row shows the temporal incoherence artefact by Johnson et al. [10]. The green and purple rectangles indicate the different appearances (texture patterns and colours) between these two stylized outputs, which exhibit flickering artefacts. The third row shows the stable results produced by our method, where the outputs preserve consistent texture appearances.

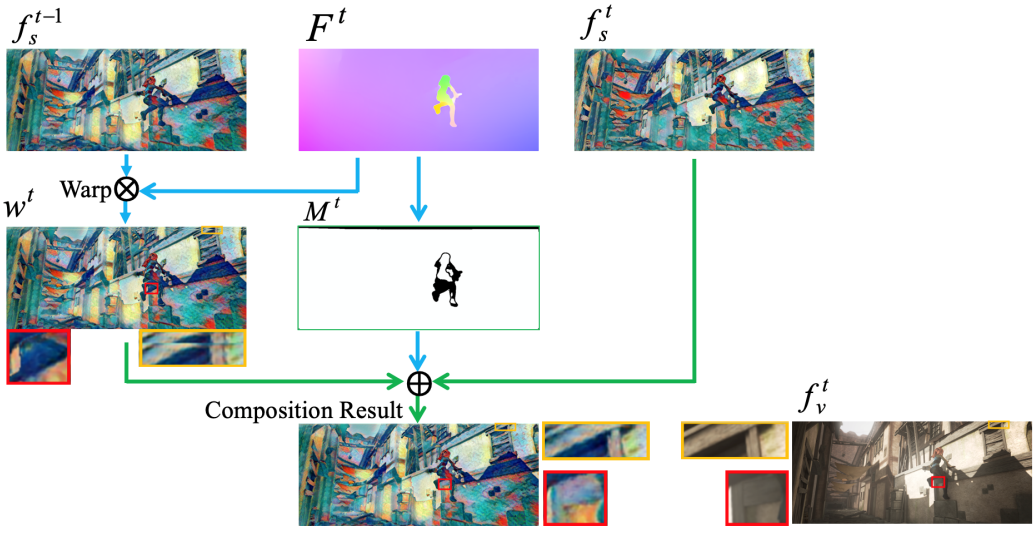

Figure 2. Prerequisite. We test the straightforward idea by recomposing pasted stylized content in untraceable flow regions (please see zoom-in rectangles), which shows that the composition result preserves the consistency of content structures better than the warped image w^t, as the warped image w^t has unexpected content structures in the red box and duplicated ones in the orange box. $\otimes$ denotes the warp operation, which warps f_s^{t-1} into w^t with F^t, and $\oplus$ denotes element-wise addition in this paper. Please zoom in to see the details.

Based on this observation, we investigate this idea, and find out that naively applying this straightforward composition may produce two new issues. Firstly, it clearly may produce discontinuous transferred textures in the composition result. For example, in Figure 3, the errors caused by the optical flow method lead to unexpected flow errors in the mask, which directly cause discontinuous textures or colours. Secondly, naively copying and pasting pixels from a warped image w^t into corresponding flow traceable regions degenerates image quality and produces blurriness artefacts. For example, in Figure 4, the copied and pasted results will accumulate degeneration errors and produce image blurriness artefacts in eye regions (red rectangles) over a long period.

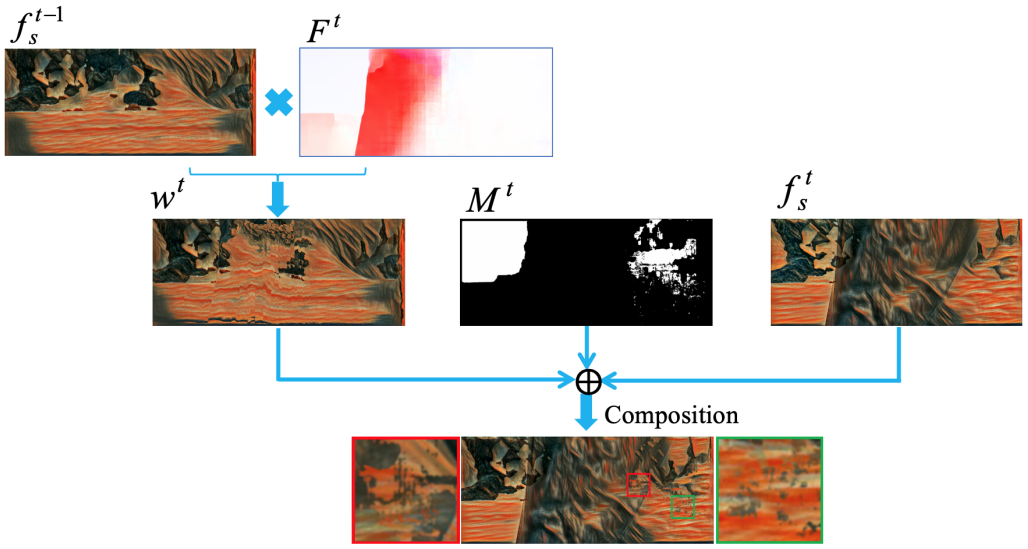

Figure 3. Texture discontinuity problem. Naively combining w^t and f_s^t via M^t into flow regions causes texture discontinuity problem. For example, in the green rectangle, the gray colours preserved from w^t lose the consistency of texture context (in red colours) which look like noise artefacts.

Figure 4. Image blurriness artefacts. Images in the upper rows are original video frames. The blurriness artefacts become more obvious with time step.

To address the ***texture discontinuity problem***, we propose a set of new mask techniques which include multi-scale mask fusion, incremental mask and multi-frame mask fusion. The reason behind the proposed mask design choices is to reduce the errors of untraceable (shown in black colour in mask images) and traceable flow regions (shown in white colour in mask images) in the mask obtained from the optical flow method (e.g., [63]). We discovered that the flow errors mostly occur at the boundaries of objects between adjacent frames as the optical flow method fails to accurately detect them. For example, the errors in untraceable flow regions are mainly caused by the low sensitivity of still objects in two adjacent frames, and the errors of traceable flow regions are mainly caused by the lack of

long temporal information (e.g., considering multiple consecutive frames at the same time). To reduce the errors in untraceable flow regions, we propose a multi-scale mask fusion technique which will only preserve untraceable flow regions at all scale levels and remove the untraceable ones in few levels. In this way, the sensitivity of the optical flow method is leveraged to still objects. However, this multi-scale scheme could also make the untraceable flow regions become much thinner than before, especially at the boundaries of moving objects, which introduces ghosting artefacts. To solve this problem, an incremental mask is proposed to make the boundaries thick again. To reduce the errors of flow traceable regions, we propose fusing the mask results from multiple adjacent frames, which only preserves the traceable flow regions (white colour in mask images) in these consecutive frames. By this means, the proposed mask techniques are able to handle large motions in multiple consecutive frames as the flow errors are effectively mitigated. In the end, we obtain a mask with much fewer errors and compose the warped image w^t and per-frame stylized result f_s^t with it. The composition image will be the new initialization for the gradient-based optimization method to preserve consistency. To reduce the ***image blurriness artefacts***, we adopt Perceptual Losses from [10] and Pixel Loss as the Sharpness Losses to update the pixel values iteratively.

The aforementioned techniques and losses only preserve two-frames' coherency. To ensure coherency at entire video level, we introduce both multi-frame RGB-level and Feature-level Coherent Losses which contribute lower average stability errors than a single one of them independently [64]. In addition, we adopt the recurrent convolutional network strategy [65] so that our network takes the current stabilized warped frame $w^t = \mathcal{W}(\widehat{x}_{out}^{t-1}, F^t)$ and the current per-frame stylized frame f_s^t as inputs, and then produces a stabilized output $\widehat{x}_{out}^t$. During the optimization process, we enforce Coherent Losses and Sharpness Losses to ensure coherency and image quality between the generated image $\widehat{x}^t$ and the previous output image $\widehat{x}_{out}^{t-1}$. In this manner, our method propagates all the flow traceable points as far as possible during the entire video style transfer process.

3.2. Fast Coherent Video Style Transfer

3.2.1. System Outline

Figure 5 shows an overview of the proposed framework. Our method takes the original video frames f_v^{t-2}, f_v^{t-1},f_v^t and per-frame stylized results f_s^t and previous output $\widehat{x}_{out}^{t-1}$ as inputs, and produces coherent output video frames $\widehat{x}_{out}^t$ where $t \in \{1, \ldots, N\}$ and N denotes the total number of frames. We develop a mask generation method consisting of the set of techniques mentioned in Section 3.1 to reduce the flow errors, and an initialization generation method to output a new initial image which is much closer to the final coherent result which speeds up the gradient-based optimization method. Specifically, starting with original the video framesf_v^{t-2}, f_v^{t-1} and f_v^t in time step t, our method generates a backward flow F^t from time t to $t-1$ using Flownet2 [63], an image w^t that warps previous output image $\widehat{x}_{out}^{t-1}$ and F^t and a mask M^t. Then we compose these three images to obtain an initial image $\widehat{x}_{init}^t$ which is fed into the network along with the three images above. The optimization process needs much less iterations than previous methods (e.g., [23,24]) as flow errors have been reduced significantly in the initial image. In order to obtain a long-term coherency, we adopt a recurrent strategy so that the output result $\widehat{x}_{out}^t$ will be fed as input into the next time step. Figure 6 shows the recurrent strategy. The short-term coherency between adjacent outputs is propagated into a long-term temporal consistency during the entire video style transfer process. In this way, our method is capable of propagating all the flow traceable points as far as possible. The details of mask generation, initialization generation and gradient-based optimization network will be discussed in following subsections.

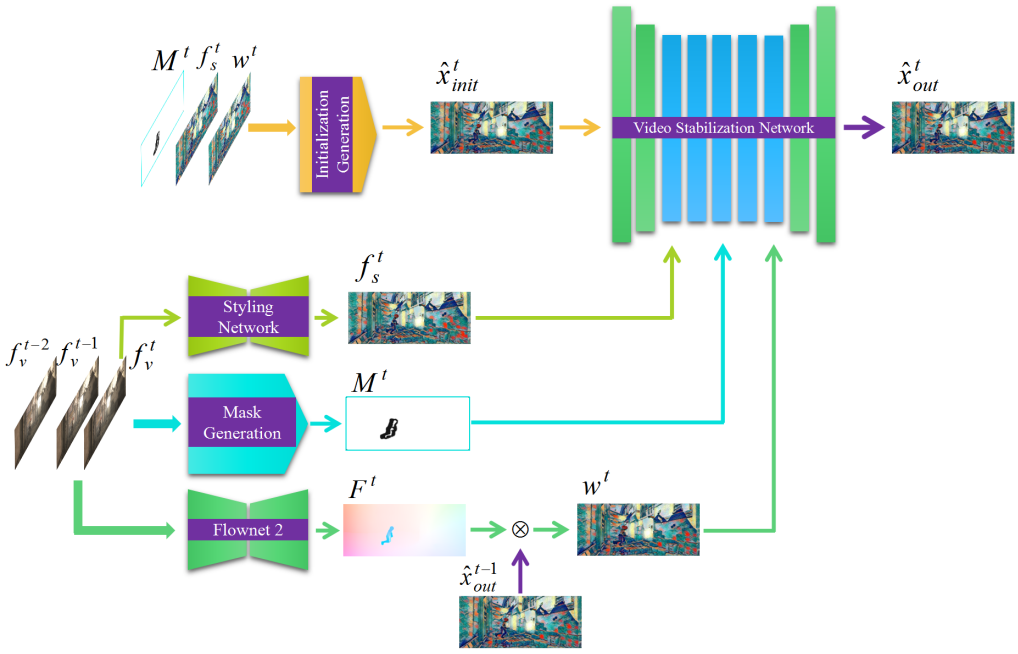

Figure 5. System overview. Starting from three consecutive frames, our system takes corresponding per-frame stylized f_s^t, mask M^t and warped image w^t as inputs, then computes initialization $\hat{x}_{init}^t$ for gradient-based optimization video stabilization network.

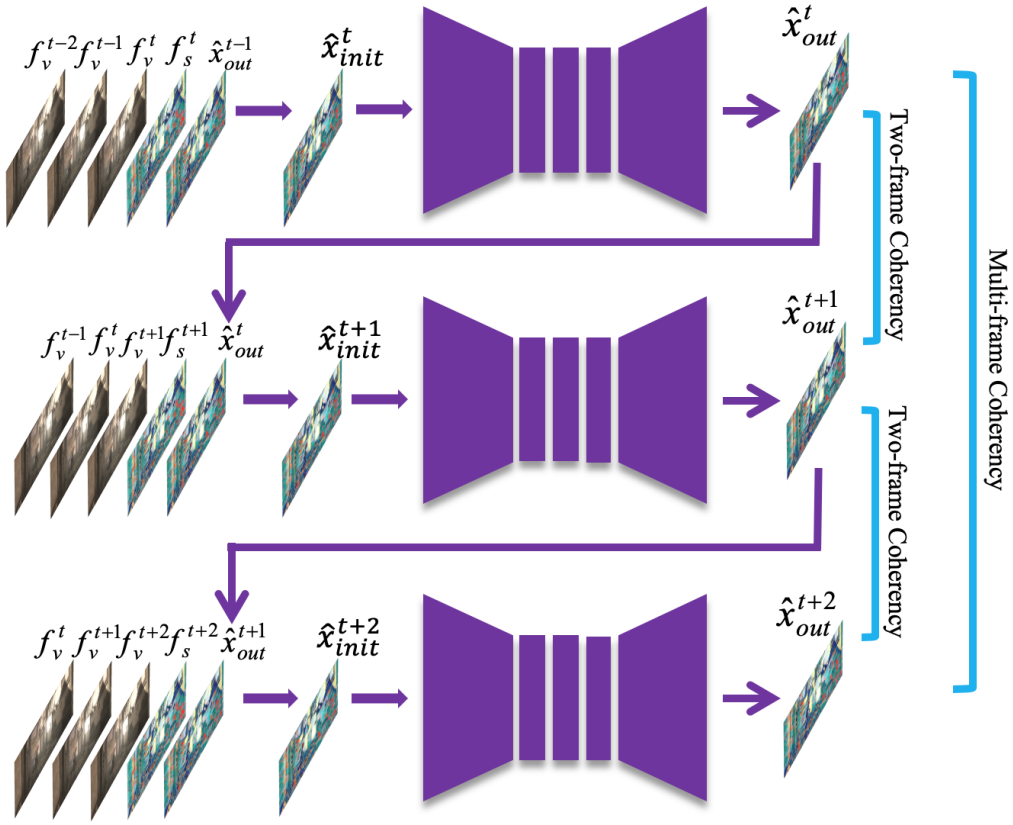

Figure 6. Recurrent strategy for video style transfer.

3.2.2. Network Architecture Overview

Figure 7 shows the details of the proposed gradient-based optimization network. In time step t, there are four images in total which are used as inputs passed into the network which are per-frame stylized result f_s^t, mask M^t, warped image w^t and initial image $\widehat{x}_{init}^t$. Coherent Losses force temporal consistency between adjacent outputs, and Perceptual Losses and Pixel Loss ensures that the image blurriness artefacts are reduced. The Coherent Losses contain an RGB-level loss and a Feature-level loss where the first one constrains the mean square error between RGB values of $\widehat{x}^t$ and w^t and the second one restricts the mean square error between feature representations of them. The Perceptual Losses force differences between $\widehat{x}^t$ and f_s^t inside the network to be reduced, and the Pixel Loss intends to keep the RGB values between $\widehat{x}^t$ and f_s^t. During each iteration, the generated image $\widehat{x}^t$ gradually compensates for discontinuous texture points and updates features into the entire image. Specifically, the gradients computed from Total Loss are back propagated into the network, and the updated weights and biases inside of each CNN layer push $\widehat{x}^t$ to grow into an image with more similarity to the inputs. In addition, the proposed gradient-based optimization network is much faster than previous methods found in the literature [23,24] as it uses a new initial image $\widehat{x}_{init}^t$. The reason for this is that the proposed optimization process needs much less iterations than previous methods [23,24] using w^t, since the initial image $\widehat{x}_{init}^t$ has significantly reduced flow errors while w^t does not.

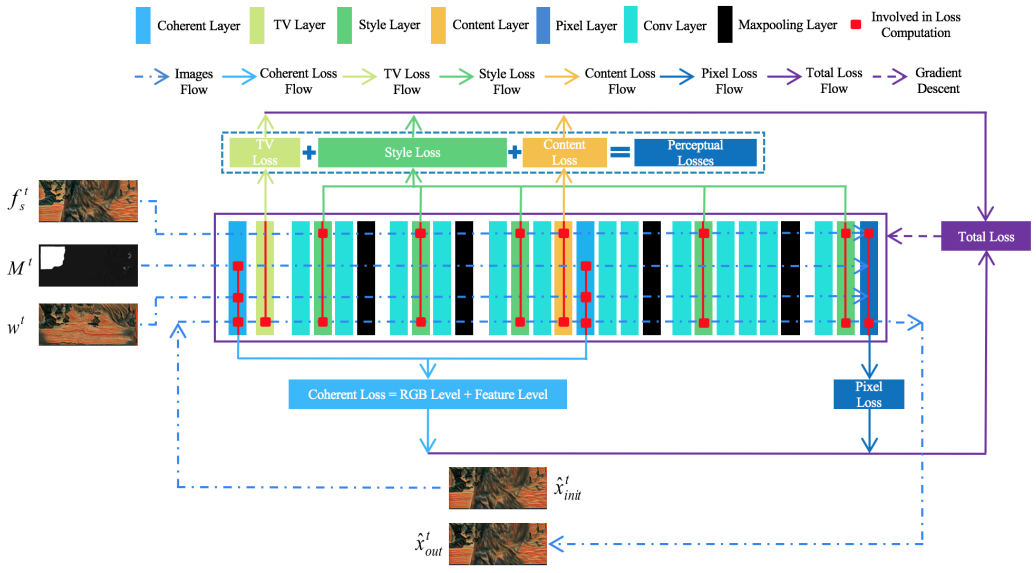

Figure 7. Network architecture overview. During optimization, the network takes $\widehat{x}_{init}^t$ obtained from initial generation, current per-frame stylized result f_s^t, mask M^t and a warped image w^t as inputs, and gradually optimizes initial image $\widehat{x}_{init}^t$ into $\widehat{x}_{out}^t$ based on gradients computed from losses.

3.3. A New Initialization for Gradient-Based Optimization Network

Based on the observation mentioned in Section 3.1, the most important part of the initial generation is to create a reliable flow mask to reduce flow errors. To this end, we propose a mask generation method to deal with it. At the beginning, we start with the following items: original adjacent video frames f_v^{t-2}, f_v^{t-1} and f_v^t, per-frame stylized results f_s^{t-2}, f_s^{t-1} and f_s^t. Then we rescale the original video frames into multiple resolutions. For the MPI Sintel dataset [66], we consider two scales $r \in \mathcal{R}$ where $\mathcal{R} = \{\sigma, \frac{1}{2}\sigma\}$ denotes the set of resolutions and σ denotes the original video resolution. And we utilize optical flow methods (e.g., [63]) to compute the corresponding forward flow F_{rf}^t and backward flow F_{rb}^t. At this time, we obtain a warped image $w^t = \mathcal{W}(f_s^{t-1}, F_{\sigma b}^t)$ at the original video resolution by warping the previous per-frame stylized result f_s^{t-1} and flow $F_{\sigma b}^t$. Next, we calculate multi-scale per-pixel flow masks which are given by a forward–backward

consistency check. The values at points of flow masks tend to be 1 in traceable flow regions where both forward and backward direction estimation agrees. On the contrary, the values at positions of flow masks tend to be 0 at disagreeing points. Then, we rescale the flow masks into the original video resolution and compose them into one mask in a value maximum manner which remains the maximum values from those flow masks at each pixel location. This step is able to fix unexpected untraceable flow errors. For example, the untraceable flow errors (black regions) in the red rectangle are fixed by multi-scale fusion in Figure 8. In addition, we find out that copying the current per-frame stylized results into the corresponding untraceable flow regions may cause worse ghosting artefacts as the flow untraceable regions become much thinner than before multi-scale fusion. Hence, we propose an incremental mask $M_{\theta}^{t=>t-1}$ which generates an incremental circle along with untraceable flow regions. Specifically, the points in the circle closer to untraceable regions has lower values. In this work, the circle width is set by default to 3 pixels and the gradient is 0.2. To further reduce traceable flow errors where large motions occur, we consider combining multiple incremental masks $M_{\theta}^{t=>t-i}$, $i \in \mathcal{T}$ where $\mathcal{T}$ denotes the set of indices of adjacent video frames. We combine $M_{\theta}^{t=>t-i}$ $(i \in \mathcal{T})$ in a value minimum manner to correct errors. In general, the fusion in a maximum manner reduces untraceable flow errors (black regions, see Figure 8) and the fusion in a minimum manner reduces traceable flow errors (white regions). in Section 5.1.

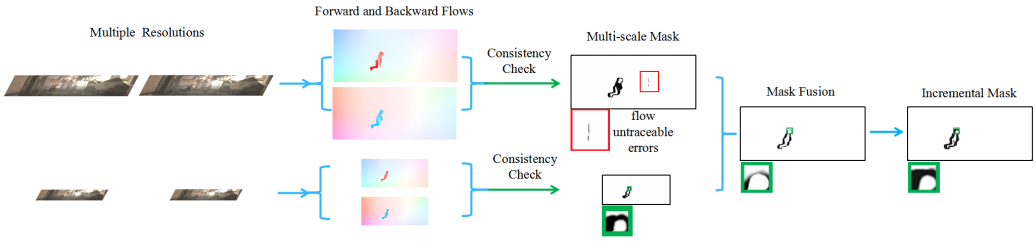

Figure 8. The process of multi-scale mask fusion and incremental mask. The unexpected flow untraceable errors (see red rectangles) are fixed in this step. The fused mask after the multi-scale scheme may cause worse ghosting artefacts as the flow untraceable regions become thinner than before; thus, the incremental mask is proposed to thicken the boundaries (see green rectangles).

Eventually, we obtain a flow mask $M^t = \min(M_{\theta}^{t=>t-i}), (i \in \mathcal{T})$ which is used for initialization generation by composing the warped image w^t and per-frame stylized result f_s^t. The generation of the initial image is shown in Figure 9. The proposed initial image is defined as:

$$\hat{x}_{init}^t = M^t \otimes w^t + (1 - M^t) \otimes f_s^t \tag{1}$$

where $\otimes$ denotes element-wise multiplication. M^t is a single channel mask. $\hat{x}_{init}^1$ is the first per-frame stylized result f_s^1 when $t = 1$.

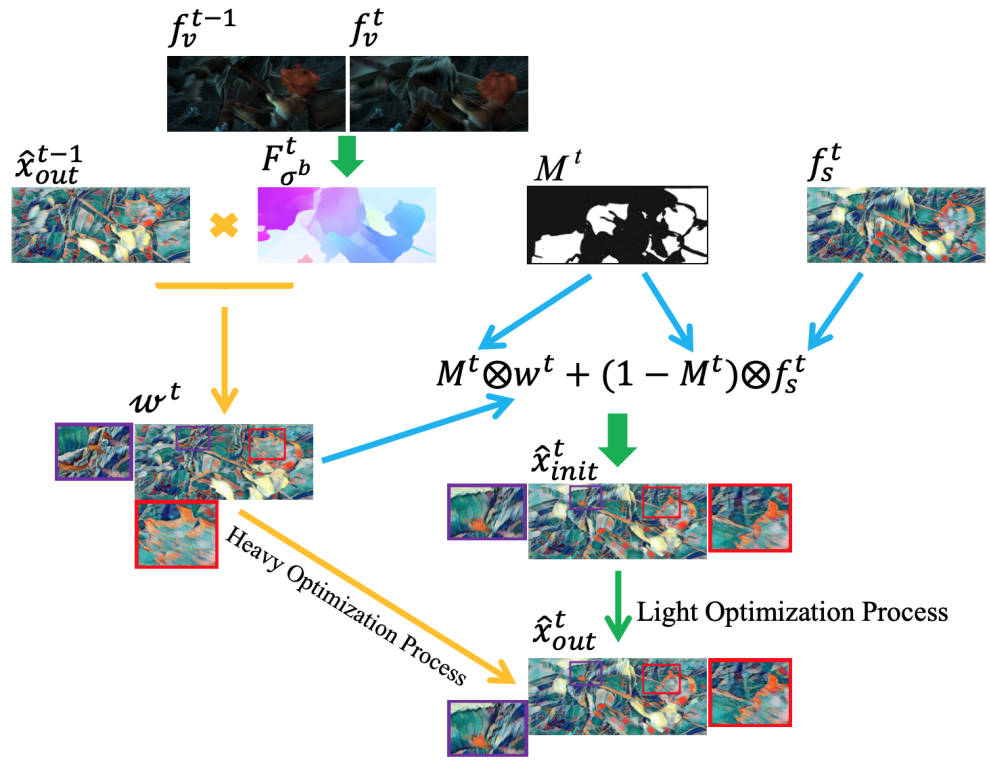

Figure 9. Initialization generation. M^t is a single channel per-pixel mask which is obtained from mask generation. Note that the generated $\hat{x}_{init}^t$ contains much fewer errors than the warped image w^t in purple and red rectangles, which leads to much fewer iterations needed to compensate for correct pixel values.

3.4. Loss Functions for Image Sharpness

Over a long period of video processing, especially for time-lapse videos, some points in frames are propagated from the beginning to the end and the copied pixel values gradually lose their quality, which results in the loss of image quality. To prevent image degeneration, we adopt the Perceptual Losses [10] and a Pixel Loss into the proposed network. The Perceptual Losses constrain the differences between high-level feature representations of the generated image $\hat{x}^t$ and the current per-frame stylized result f_s^t. The Pixel Loss preserves the pixel values between the generated image $\hat{x}^t$ and f_s^t in the RGB domain.

The Perceptual Losses contain a Content Loss $\mathcal{L}_{con}(\hat{x}^t, f_s^t)$, Style Loss $\mathcal{L}_{sty}(\hat{x}^t, f_s^t)$ and Total Variation Regularization $\mathcal{L}_{tv}(\hat{x}^t)$, which can be formulated as follows:

$$\mathcal{L}_{perce}(\hat{x}^t, f_s^t) = \alpha \mathcal{L}_{con}(\hat{x}^t, f_s^t) + \beta \mathcal{L}_{sty}(\hat{x}^t, f_s^t) + \gamma \mathcal{L}_{tv} \tag{2}$$

where α, β and γ are the weights of three loss terms, respectively. In our experiments, the ratio of α/β close to 0.3 produces better image quality, and we set $\gamma = 1e-3$ (default in [10]) in all experiments. The Pixel Loss is defined as the mean square error between the generated image $\hat{x}^t$ and current stylized result f_s^t:

$$\mathcal{L}_{pixel}(\hat{x}^t, f_s^t) = \frac{1}{D} \sum_{i,j}^{D} (\hat{x}_{(i,j)}^t - f_{s(i,j)}^t)^2 \tag{3}$$

where $D = 3 \times H \times W$ denotes the total number of pixels in the input image and $H \times W$ denotes the height times width. In this work, we refer to Perceptual Losses and Pixel Loss as our Sharpness Losses which ensure the image's sharpness during the entire video style transfer process. The Sharpness Losses are the combination of Perceptual Losses and Pixel Loss, which is defined as:

$$\mathcal{L}_{sharpness} = \mathcal{L}_{perce} + \kappa \mathcal{L}_{pixel} \tag{4}$$

where κ denotes the weight for Pixel Loss.

3.5. Loss Functions for Temporal Consistency

3.5.1. Rgb-Level Coherent Loss

The flickering artefact is actually represented by texture and colour discontinuities in RGB-level regions between consecutive frames, such as disoccluded regions and motion boundaries. Pixel values in these areas change in adjacent frames, and the optimizer [9] or feed-forward network [10] transforms them differently in a particular style as well. To detect these disoccluded regions and motion boundaries, we apply optical flow methods (e.g., Flownet2 [63]) to estimate these flow traceable areas between coherent frames. Let f_s^{t-1} and f_s^t denote two adjacent per-frame stylized results, $\widehat{x}_{out}^{t-1}$ denote the previous output, $\mathcal{W}(\cdot)$ denote the function to warp image and w^t denote the warped image using the previous output image $\widehat{x}_{out}^{t-1}$ and the optical flow F^t from f_s^t to f_s^{t-1} (backward direction). The warped image w^t is then given by:

$$w^t = \mathcal{W}(\widehat{x}_{out}^{t-1}, F^t) \tag{5}$$

In [23,24], the Coherent Loss function is supposed to preserve the pixel values of the traceable flow regions in the stabilized outputs, and the flow errors in w^t are then rebuilt through the style transfer process. The straightforward two-frame temporal coherency loss considers the consistency between two adjacent frames; thus, the ***two-frame RGB-level Coherent Loss*** is denoted as the mean squared error between the generated image $\widehat{x}^t$ and w^t:

$$\mathcal{L}_{two}^{RGB}(\widehat{x}^t, w^t, M^t) = \frac{1}{D}\sum_{i=1}^{D} M_i^t \cdot (\widehat{x}_i^t - w_i^t)^2 \tag{6}$$

where $D = N \times H \times W$ denotes the dimensionality of $\widehat{x}^t$ and w^t. N denotes the number of image channel and $H \times W$ is height times width, and M^t denotes the per-pixel flow mask with weights of the coherent loss. This $\mathcal{L}_{two}^{RGB}$ only considers consistency between two adjacent frames, which causes small errors as the proposed initial image utilizes a mask operating on multiple consecutive frames. To further enhance the coherency, we take consistency between more adjacent frames into account. Let us consider a multi-frame coherency between several adjacent frames, and let $\mathcal{T}$ (same as Section 3.3) denote the set of indices for video frames which are considered as relative frames. For instance, $\mathcal{T} = \{1, 2, 3\}$ denotes that processing frame $\widehat{x}^t$ considers coherency between frame f_s^t and frame f_s^{t-1}, frame f_s^t and frame f_s^{t-2}, frame f_s^t and frame f_s^{t-3}. Then the ***multi-frame RGB-level Coherent Loss*** is defined as the combination of three two-frame RGB-level coherent loss $\mathcal{L}_{two}$s:

$$\mathcal{L}_{mul}^{RGB}(\widehat{x}^t, w^{t-\mathcal{T}}) = \sum_{i \in \mathcal{T}: t > i} \mathcal{L}_{two}^{RGB}(\widehat{x}^t, w^{t-i}, M^{t-i}) \tag{7}$$

3.5.2. Feature-Level Coherent Loss

In previous methods [27,28,30], RGB-level coherent loss was considered to constrain the consistency between pixel values of the warped w^t and generated $\widehat{x}^t$ images. However, it may not be accurate in preserving stylized texture consistency since their methods do not take into account the temporal consistency of feature representations in CNN layers. Hence, we adopt a feature-level Coherent Loss [64] for preserving texture consistency which is capable of constraining the feature consistency in high-level CNN layers. Let $\psi^l(\widehat{x}^t) \in \mathbb{R}^{N_l \times H_l \times W_l}$ and $\psi^l(w^t) \in \mathbb{R}^{N_l \times H_l \times W_l}$ denote the feature representations of generated images $\widehat{x}^t$ and warped images w^t at layer l, respectively, and M^t denote the per-pixel mask, then the ***two-frame Feature-level Coherent Loss*** for two frames is defined as the mean squared error between $\psi^l(\widehat{x}^t)$ and $\psi^l(w^t)$:

$$\mathcal{L}_{two}^{fea}(\widehat{x}^t, w^t, M^t) = \sum_{l \in L_{coh}^{fea}} \frac{1}{N_l} \sum_{i}^{N_l} M_i^t \cdot (\psi_i^l(\widehat{x}^t) - \psi_i^l(w^t))^2 \tag{8}$$

where L_{coh}^{fea} denotes the set of layers computing Feature-level Coherent Loss and N_l denotes the dimensionality of feature representations $\psi^l(\cdot)$. Similar to RGB-level Coherent Loss, we also consider the feature-level coherent loss between more adjacent frames (same $\mathcal{T}$ in $\mathcal{L}_{mul}^{RGB}$), and propose the ***multi-frame Feature-level Coherent Loss*** term:

$$\mathcal{L}_{mul}^{fea}(\widehat{x}^t, w^{t-\mathcal{T}}) = \sum_{i \in \mathcal{T}: t>i} \mathcal{L}_{two}^{fea}(\widehat{x}^t, w^{t-i}, M^{t-i}) \tag{9}$$

The total multi-frame Coherent Losses are defined as the combination of $\mathcal{L}_{mul}^{RGB}$ and $\mathcal{L}_{mul}^{fea}$:

$$\mathcal{L}_{coherent} = \lambda_{coh}^{RGB} \mathcal{L}_{mul}^{RGB} + \lambda_{coh}^{fea} \mathcal{L}_{mul}^{fea} \tag{10}$$

where λ_{coh}^{RGB} and λ_{coh}^{fea} are the weights to corresponding terms. We find that the ratio of $\lambda_{coh}^{RGB} / \lambda_{coh}^{fea}$ close to 2.5 makes a better temporal consistency preservation.

Overall The total loss term for the optimization process in each time step is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{sharpness} + \mathcal{L}_{coherent} \tag{11}$$

4. Implementation Details

Code Development: our method was developed on a Torch implementation called artistic video style transfer [24] in Ubuntu 18.04 LTS, which uses the Lua script language and a pretrained-VGG19 as the backbone neural network. The code was built with CUDA 10.8 on a single NVIDIA GeForce 1080Ti card. The optimizer for the iterations is L-BFGS. The stopping criterion: we consider the optimization to be converged when the total loss does not change by more than one during ten iterations.

Choices of Layers for Losses: the $\{relu1\_1, relu2\_1, relu3\_1, relu4\_1, relu5\_1\}$ layers are chosen for the Style Loss, and the $\{relu3\_2\}$ layer is chosen for the Content Loss within the Perceptual Losses. We choose $\{relu3\_2\}$ for the Feature-Level Coherent Loss.

Inputs and Outputs: The following inputs are fed into our method: a per-frame stylized image f_s^t as the feature and style target, a warped image w_{t-i}^t $(i \in \mathcal{T})$ as the temporal consistency target, a mask M^t as the per-pixel flow weight and the $\widehat{x}_{init}^t$ generated from Equation (1) as the initial image. Outputs: the stylized result $\widehat{x}_{out}^t$. In all the experiments, we used two scales (original resolution and half of the original resolution) of frame resolution for the mask fusion mentioned in Figure 8.

Choices of Hyperparameters in Equations (2), (4), (7), (9) and (10): For videos at 1024×436 (MPI Sintel dataset) and 854×480 (Davis 2017 dataset) resolution, the hyperparameters were chosen as follows: $\alpha = 3 \times 10^1$, $\beta = 9 \times 10^1$ and $\gamma = 1 \times 10^{-3}$ in Equation (2), $\kappa = 9 \times 10^{-7}$ in Equation (4), $\mathcal{T} = \{1, 2\}$ in Equations (7) and (9), $\lambda_{coh}^{RGB} = 5 \times 10^1$ and $\lambda_{coh}^{fea} = 2 \times 10^1$ in Equation (10). Discussions about these choices: in our experiments, a ratio of α / β close to 0.3 preserves the stylistic texture appearance of the per-frame stylized image f_s in the outputs better. A higher ratio will generate results with more sharpness but fewer stylistic textures; on the contrary, a lower ratio degenerates image quality due to excessive stylization. A coherent ratio of $\lambda_{coh}^{RGB} / \lambda_{coh}^{fea}$ close to 2.5 preserves the balance between execution time and temporal consistency better. A higher coherent ratio of warps f_s^t on RGB-level may cause pixel errors to accumulate along with propagation, which costs more time as the network has to correct them. While a lower coherent ratio tends to reduce the stylization consistency of the same object among consecutive frames. In addition, a κ close to 9×10^{-7} maintains a better balance between image quality and temporal consistency. For example, larger κ values tend to produce outputs with better image quality (e.g., pixels updated more frequently) but poor temporal consistency as pixels propagated

from the beginning are lost with time. Smaller κ values, to some extent, fail to effectively mitigate blurriness artefacts but preserve a better temporal consistency. In this paper we chose the aforementioned hyperparameters values for all the testing videos. The optical flow used in this paper was Flownet2 [63], but Deepflow2 [67] is also supported.

Speed Compared to the gradient-based optimization methods in the literature [23,24] (3–5 min per frame), the proposed optimization process takes around **1.8 s** per frame for a resolution of 1024 × 436 and around **1.6 s** per frame for a resolution of 854 × 480 on a single NVIDIA GTX 1080 Ti graphics card. The reason for the fast speed is that the optimizer (e.g., L-BFGS) in our work needs much fewer iterations as there are already enough for temporal consistency and image sharpness.

5. Experiments

5.1. Qualitative Evaluation

5.1.1. Analysis of Initialization

Figure 8 in Section 3.3 shows that the mask generation is capable of reducing untraceable flow errors (see rectangle). In this section, we analyze the capability of reducing flow traceable errors (including ghosting artefacts), which mainly degrades the visual image quality. Figure 10 shows the effect of the proposed initialization on the reduction of flow traceable errors. Starting from three adjacent original video frames, the mask generation outputs a mask with much fewer traceable flow errors (white regions in the right side) than previous single scale mask. This directly helps the composed initial image contain more consistent textures (see rectangles) than x^t_{init} in Figure 3 in Section 3.1. The small discontinuous textures on the left are then fixed by the gradient-based optimization network with Sharpness Losses.

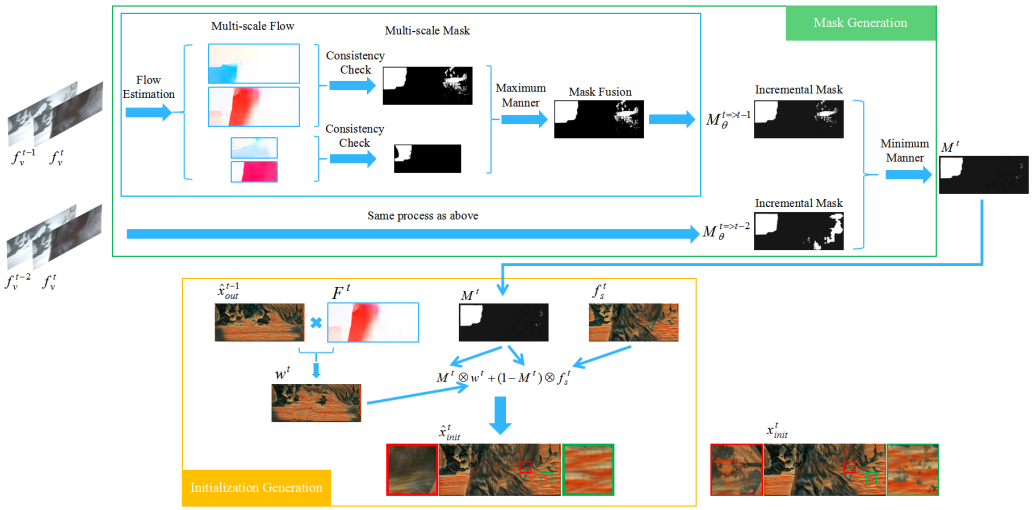

Figure 10. The decrease in traceable flow errors (white regions in the right side) by using the proposed initialization. The rectangles indicate the error difference between the initialization $\hat{x}^t_{init}$ and x^t_{init} without our mask generation. The fusion in a maximum/minimum value manner indicates that we remain the maximum/minimum values from those masks at each pixel location.

Ablation study on proposed mask techniques. As mentioned in Section 3.3, we propose a multi-scale scheme, incremental mask and multi-frame mask fusion for initialization generation. To analyze these techniques fairly, we divided the proposed mask techniques into four different groups: naive method (without any proposed techniques), w/ multi-scale scheme, w/ multi-scale + incremental and w/ multi-scale + incremental + multi-frame. The outputs of the four groups are shown in Figure 11. As can be seen, the zoomed-in rectangles on naive method indicate that the general flow mask proposed by [24] causes the ghosting artefacts. Adding a multi-scale scheme into the naive method is

able to reduce the untraceable flow errors (see Figure 8) while also causing worse ghosting artefacts. We gradually add incremental mask (bottom-left) and multi-frame mask fusion (bottom-right) techniques into the w/ multi-scale scheme method, which finally produces results without ghosting artefacts.

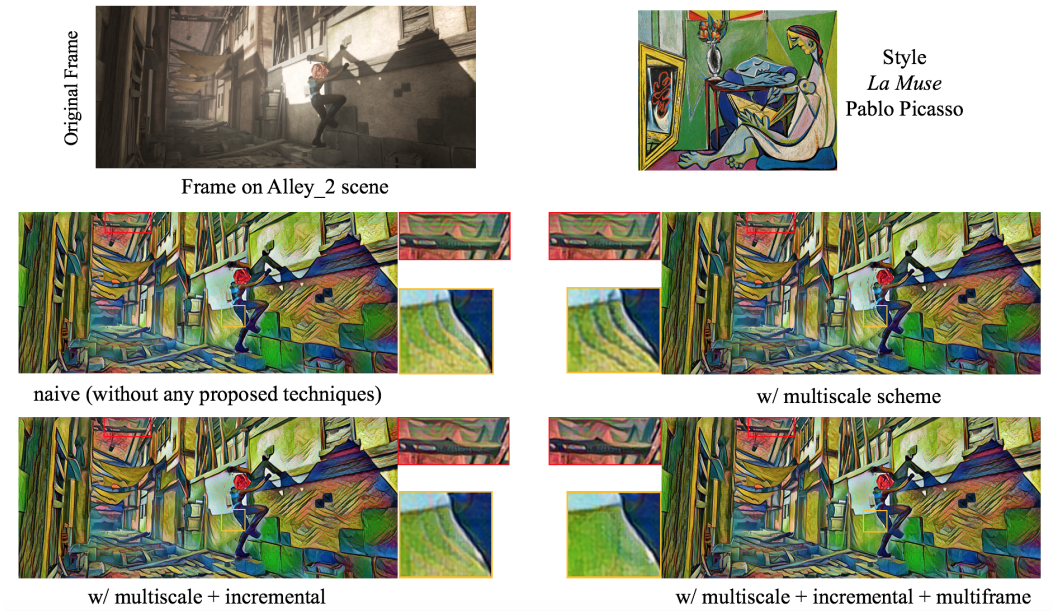

Figure 11. Qualitative ablation study on proposed mask techniques of Alley_2 scene from MPI Sintel dataset [66]. The naive method using the flow mask produced by [24] causes ghosting artefacts (see unexpected grids and curves in red and orange rectangles). The multi-scale scheme causes worse ghosting artefacts. By gradually adding the incremental mask and multi-frame mask fusion techniques, the unexpected grids and curves are effectively mitigated which produces better visual quality without ghosting artefacts.

5.1.2. Analysis of Loss Functions

Sharpness Losses. Figure 12 shows the effect of Sharpness Losses in our approach. Without the Sharpness Losses, the straightforward idea that copying and pasting pixels from the per-frame stylized result f_s^t and the warped image w^t into corresponding regions through the mask M^t causes the pixel loss, and this loss accumulates along the entire video process, which results in a significant number of image blurriness artefacts. By adding the Sharpness Losses, our approach ensures that the pixel values are compensated for from the beginning to the end which prevents image blurriness artefacts in the video outputs.

Coherent Losses. Figure 13 shows the effect of Coherent Losses in our method. Per-frame processing methods like [10,13] produce flickering artefacts in adjacent frames (see the zoom-ins) without any consideration of temporal consistency. The proposed method takes the per-frame stylized frames as inputs and produces the texture consistent consecutive outputs; for example, the rectangle areas.

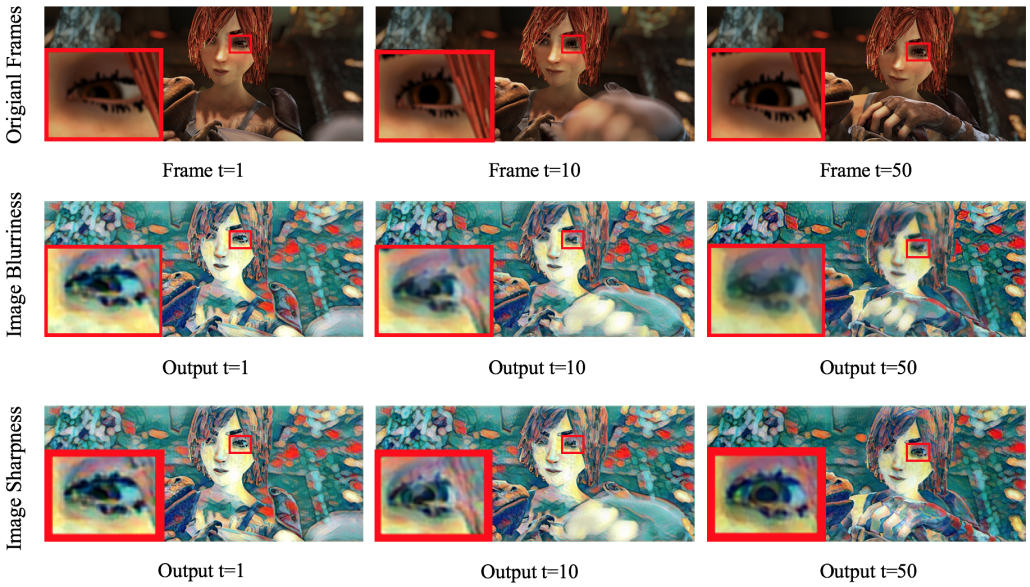

Figure 12. The effect of image sharpness. The top rows are original video frames, the middle rows are outputs without Sharpness Losses, and the bottom rows are outputs with Sharpness Losses. The red rectangles indicate the difference of image sharpness.

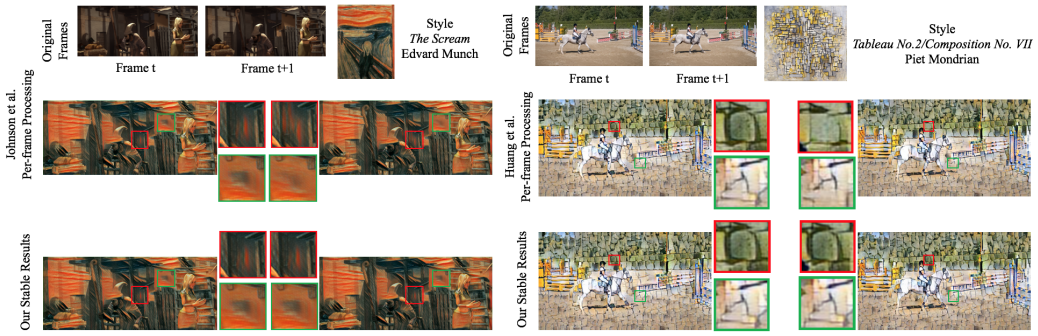

Figure 13. The effect of temporal consistency. The per-frame processing methods are Johnson et al. [10] and Huang et al. [13]. The red and green rectangles indicate the discontinuous texture appearances.

5.1.3. Comparisons to Methods Found in the Literature

Figure 14 shows a qualitative comparison between the proposed approach and more recent state-of-the-art methods [36–38,62]. As the figure shows, all five of the methods give visually satisfying transfer results in terms of long-term temporal consistency. However, the representation of texture pattern from the style reference image varies differently, and the details (e.g., object contours) are preserved with artefacts. For example, the representative mosaic texture patterns from the style image basically do not appear in any of the results of previous four methods [36–38,62], but the colour information does. In addition, Deng et al. [62] and Wu et al. [37] produced distorted results with severe artefacts in which object contours are distorted in the zoom-ins (please see the zoom-ins in the third and fourth row). Gu et al. [38] introduced strange vertical artefacts (please see zoom-ins in the fifth row) into their results. Wang et al. [36] produced unfaithful colour information in their results compared to the other methods. In contrast, our results preserve the mosaic texture patterns better and object contours do not have distortions.

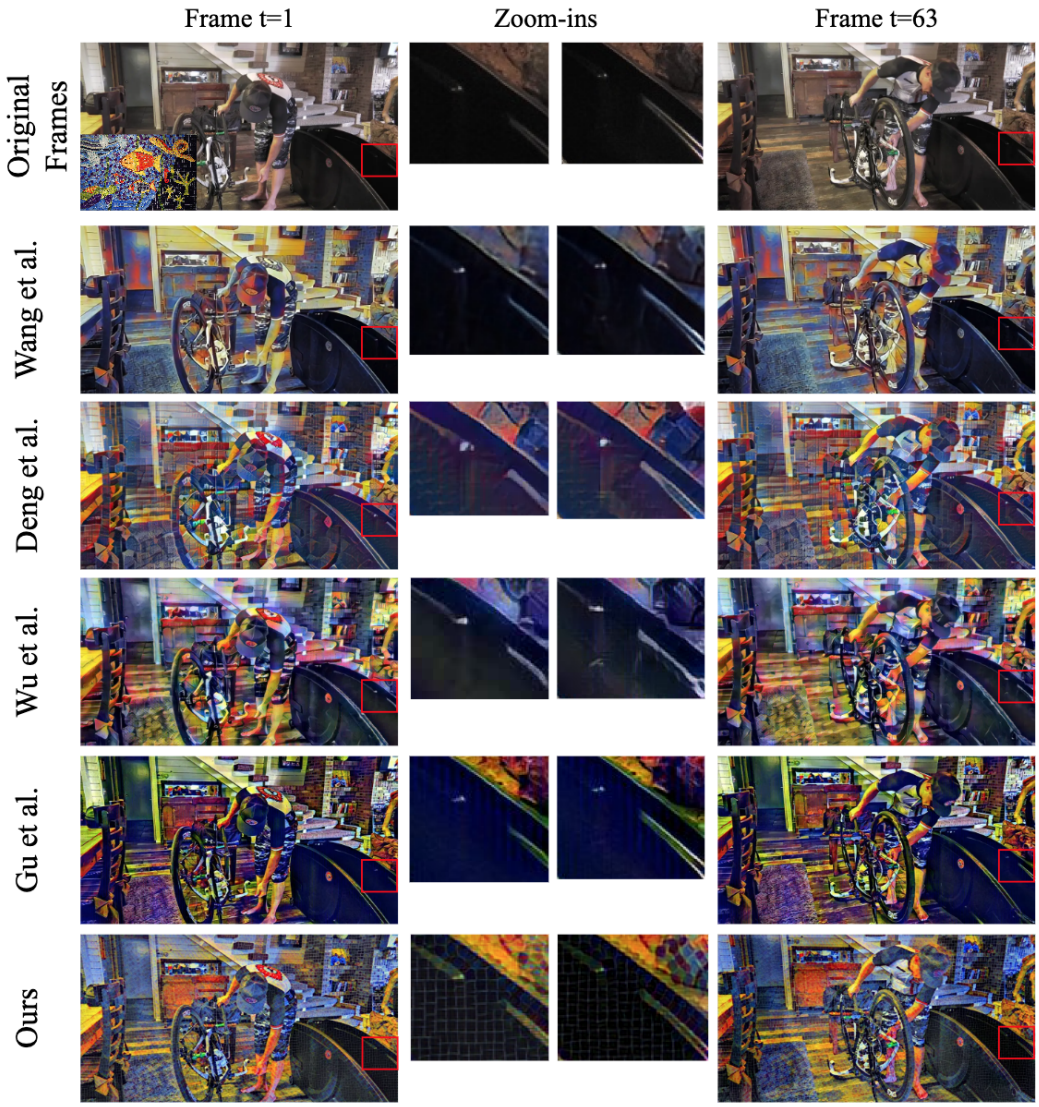

Figure 14. Comparison to latest state-of-the-art methods [36–38,62] on bike-packing scene from DAVIS 2017 dataset [66]. The red rectangles indicate the differences in long-term temporal consistency results.

Figure 15 shows a comparison between our approach and Li et al. [35]. The method proposed by Li et al. [35] learns a linear transformation matrix to minimize the difference between the covariance of the transformed content features and style features, which serves as a second-order statistics transformation of the reference image onto the content image in prior methods [9,10]. The linear transformation is highly efficient but causes less stylistic texture to be present in transformed videos. For instance, the mosaic texture patterns in the style reference image in Figure 14 do not appear in their transferred video frames, which increases the temporal consistency among adjacent frames but leads to the appearance of less artistic texture. However, our method preserves temporal consistency (c.f. last table) and texture patterns better than theirs (c.f. zoom-ins).

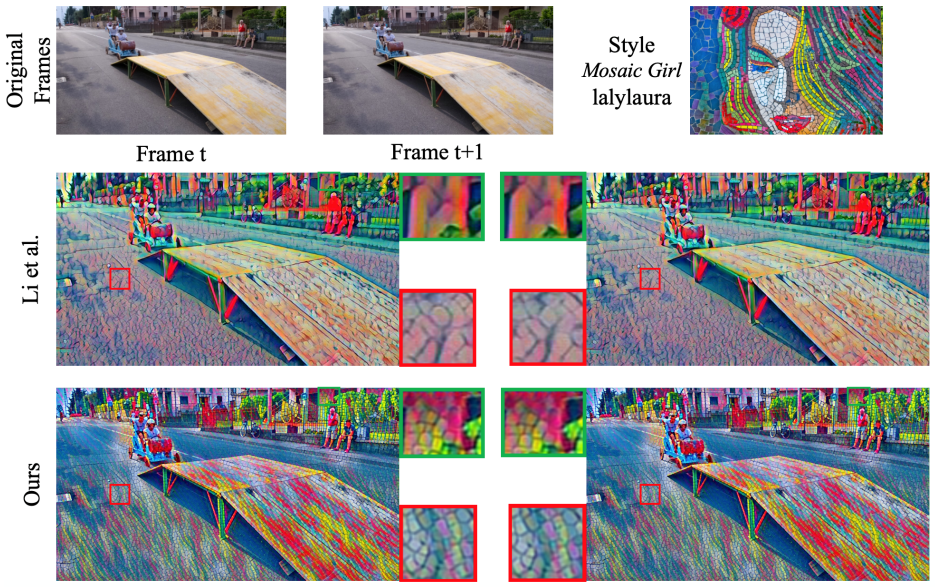

Figure 15. Comparison to Li et al. [35] on Soapbox scene from DAVIS 2017 dataset [66]. Both the red and green rectangles indicate the differences in two adjacent stabilization results. Please view our Supplementary Video for better observations.

Figure 16 shows the comparison between our approach and Ruder et al. [30] in a large motion and strong occlusion case. The network-based video style transfer approach proposed by Ruder et al. [30] fails to produce consistent texture appearances around traceable flow areas (white areas) as their method is not able to correct such flow errors. while our approach corrects these traceable flow errors via the proposed mask techniques, and generates consistent texture appearances in the given contexts.

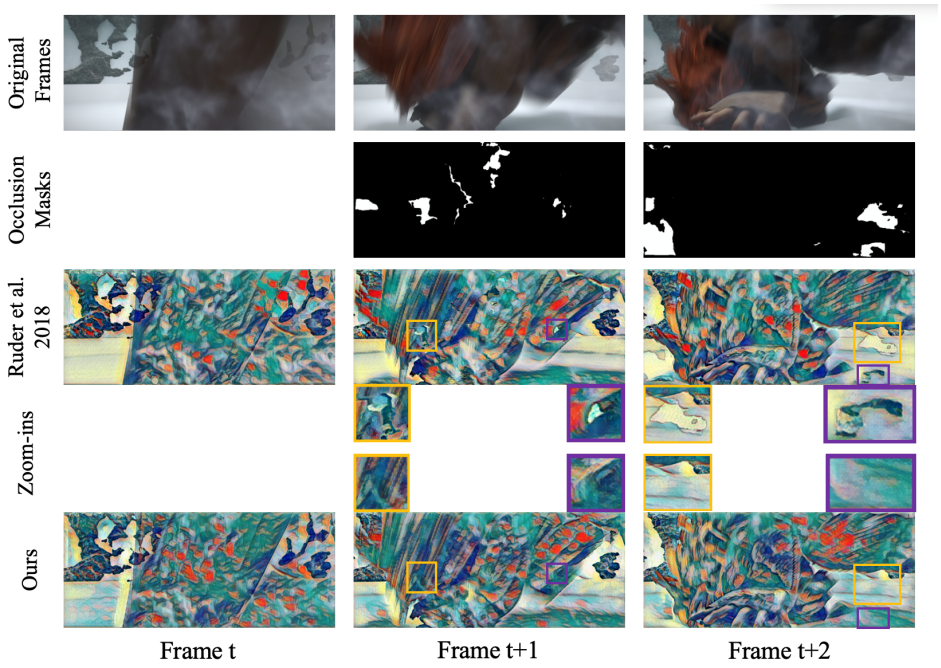

Figure 16. Comparison to Ruder et al. [30] on Ambush_4 scene from MPI Sintel dataset [66]. The per-frame processing method for both methods is Johnson et al. [10]. The rectangles indicate the difference in the consistent texture appearances in a few adjacent stabilization results. Orange rectangles show the texture appearances around large occlusion boundaries (including flow traceable errors) by Ruder et al. [30] are discontinuous with context, while purple boxes demonstrate that our textures are consistent with context. Please view our Supplementary Video.

5.2. Quantitative Evaluation

We verified our method on the MPI Sintel dataset [66] and Davis 2017 dataset [68], and tested our approach on more than 40 videos including animation and real-world videos. To fairly evaluate the performance of different approaches, we adopted two metrics to assess the transferred videos on stylization effects and temporal consistency. Here, the SIFID score is adopted to measure the style distribution distance between the generated image and its style input. And a lower SIFID value indicates the closer style distribution of a pair. In addition, the stability error is adopted to measure temporal consistency, and a lower stability error value indicates the better temporal consistency. In this section, we firstly detail the ablation studies for image sharpness, temporal consistency and proposed mask techniques in initialization to verify our Sharpness Loss, Coherent Losses and initialization. Then, we quantitatively compare our method with state-of-the-art methods ([10,13,30,35–38,62]) by using the term ***stability error*** e_{stab} which calculates the temporal errors between pairs of adjacent frames in an output video. The ***stability error*** e_{stab} is defined as:

$$e_{stab} = \sqrt{\frac{1}{(N-1) \times D} \sum_{t=2}^{N} \sum_{i=1}^{D} m_i^t (x_i^t - w_i^t)^2} \tag{12}$$

where N denotes the total frame number of a video output, D denotes the total number of pixels in one video frame and m^t denotes the per-pixel flow weight. This formulation is similar to Equation (6), except that we sum up all the temporal loss for pairs of consecutive frames in a video. The warped image w^t warps the = x_{out}^{t-1} at $t-1$ which is time forward to t. The ground truth of optical flow and flow weight can be obtained from Flownet2 [63] or the MPI Sintel dataset [66].

5.2.1. Ablation Study on Loss Functions

Sharpness Losses. To quantitatively verify the Sharpness Losses, we chose a No-reference autoregressive(AR)-based Image Sharpness Metric (ARISM) which is proposed particularly for assessing image sharpness [69]. The ARISM is established on the hypothesis that AR model parameters estimated from eight connected neighbourhoods of one image pixel tend to be very close to each other when this pixel is located in a comparatively smooth region; otherwise, these parameters are clearly distinct when this pixel is in a sharp region. The ARISM sharpness score is formulated as:

$$\rho = \sum_{k \in \Omega} \theta_k \rho_k \tag{13}$$

where $\Omega = \{E, C, E^{bb}, C^{bb}\}$ and θ_k are the weights of each component. E and C are two classical metrics used to define the difference between the maximum and minimum values of AR parameters at point (i, j) of the input image. E^{bb} and C^{bb} are the block-based pooling [70] of E and C, respectively.

ARISM has been proved to be robust in assessing colour images with no reference, which is suitable for our case since the outputs of the video style transfer are colourful and also have no reference images. Figure 17 shows the scores of each frame in the Alley_2 scene (MPI Sintel dataset) which are obtained from w/ Perceptual Losses + Pixel Loss (aka, Sharpness Losses blue line) and w/ Perceptual Losses (red line) and w/o Sharpness Losses (magenta line), and **the higher ARISM score is better**. The scores of w/ Perceptual Losses + Pixel Loss are steady around the average score 2.674, while the average score of w/ Perceptual Losses tend to decrease from 2.674 to 2.66, and those of w/o Sharpness Losses are close to 2.651. Note that the scores of w/ Sharpness Losses are always higher than those of w/o Pixel Loss, which indicates that outputs with both Sharpness Losses contain much fewer blurriness artefacts than those without Pixel Loss.

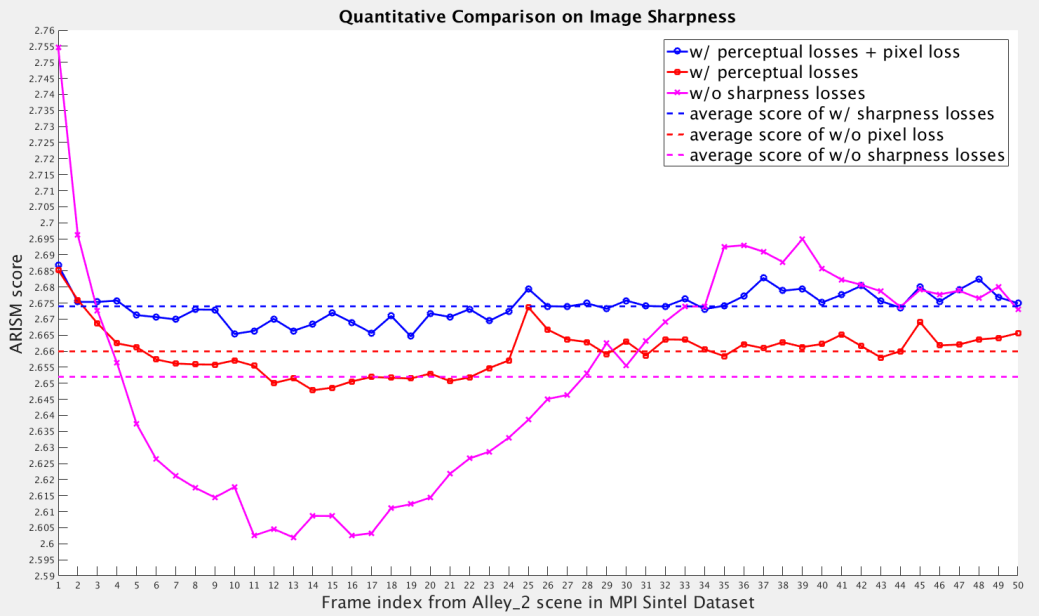

Figure 17. Ablation study on Sharpness Losses of Alley_2 scene from MPI Sintel dataset [66]. *The higher ARISM score is better*. The outputs with our Sharpness Losses achieve higher ARISM scores than those without pixel loss and Sharpness Losses, which indicates that Perceptual Losses and Pixel Loss in the proposed Sharpness Losses both contribute to reducing blurriness artefacts.

Coherent Losses. We carried out the quantitative ablation study on Coherent Losses in five testing scenes to compare the stability errors of two groups of Temporal Losses: multi-frame RGB-level only and both levels. Table 1 shows the detailed stability errors in the baseline method [10] and two groups. It is noticeable that multi-frame RGB-level only Coherent Loss contributes to a 58.8% improvement compared to the baseline method [10], while multi-frame Feature-level Coherent Loss contributes a further approximately 2.2% improvement which finally leads to 61.0% improvement in total.

Table 1. Ablation study on Coherent Losses ($\times 10^{-2}$) of five testing videos in MPI Sintel dataset. Style: Woman with A Hat. Scene1: alley_2, Scene2: bamboo_2, Scene3: bandage_1, Scene4: cave_4, Scene5: market_2. The best result is highlighted with underline. As baseline method doesn't have improvement of itself, thus we add * in below.

Method	MPI Sintel Dataset + Style						
	Scene1	Scene2	Scene3	Scene4	Scene5	ave	Improvement
Baseline [10]	10.55	7.21	6.34	12.06	6.34	8.50	*
RGB-Level only	2.71	3.17	2.79	5.93	2.90	3.50	58.8%
Both Levels	2.43	3.00	2.59	5.74	2.75	3.31	61.0%

5.2.2. Ablation Study on Initialization

As mentioned in Section 3.1, we propose a set of new mask techniques to address the texture discontinuity problem. We now give a detailed analysis of image quality of the proposed techniques including the multi-scale scheme, incremental mask and multi-frame mask fusion. To fairly compare these techniques, we follow Section 5.1.1 and divide them into four groups, and we compare the general Image Quality Assessment (IQA) scores of the four groups using the following term proposed in [60]:

$$Q = \frac{1}{N_p} \sum_{i}^{N_p} y_i \tag{14}$$

where N_p denotes the number of patches which are chosen from the given image, and y_i denotes the estimated visual qualities of patch i. We chose this IQA score (refer to DIQaM-NR method in [60]) as our general image quality metric because it is capable of assessing image quality by coping with several distortion types such as luminance and contrast changes, compression, Gaussian noise and the Rayleigh fading channel. Figure 18 shows the detailed ablation study carried out on the proposed mask techniques, and **a lower score indicates better visual image quality**. Note that the multi-scale scheme only (magenta line) reduces untraceable flow errors (see Figure 8) but it causes higher scores than the naive method (blue line). The incremental mask technique (red line) helps to achieve lower average scores than the naive method (blue line) and w/ multi-scale scheme method (magenta line), and multi-frame mask fusion (green line) further improves the image quality by decreasing the average score from 30.69 (dashed blue line) to 29.71 (dashed green line). This observation practically follows the qualitative evaluation in Figure 11.

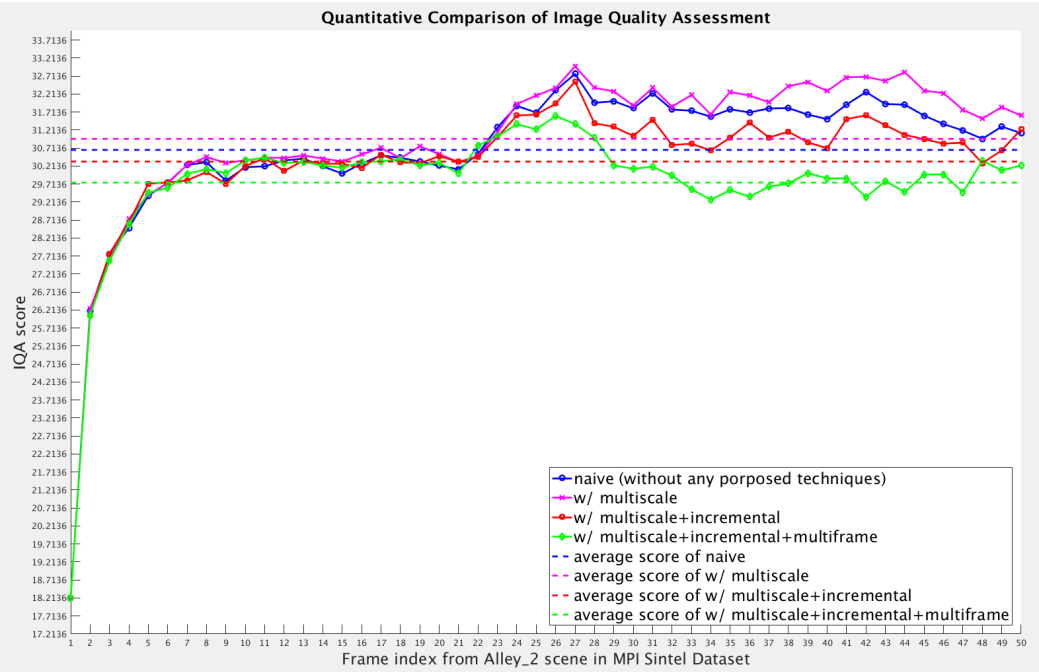

Figure 18. Ablation study on image quality assessment of Alley_2 scene from MPI Sintel dataset [66]. *A lower score indicates better visual image quality*. Note that adding multi-scale scheme (magenta line) causes image quality loss (higher score) compared to naive method (blue line), while adding incremental mask and multi-frame fusion (red line and green line) contributes to achieve lower scores than naive method (blue line).

5.2.3. Quantitative Evaluation in Literatures

The stylized perceptual strokes/patterns in the results of the different methods are clearly distinct even for one particular style, which may make the IQA comparison unfair. Thus, we here give a detailed comparisons of the stability errors which are invariant to diverse strokes/patterns. Table 2 lists the stability errors created by the three different approaches in the MPI Sintel dataset [66] and Davis 2017 dataset [68]. Our approach takes per-frame stylized results produced by per-frame processing methods as inputs and obtains the stable outputs; thus, we give two comparisons in Table 2. There are five different scenes chosen from each dataset and combined with two representative styles. It is noticed that, for all testing videos, our method significantly reduces the stability errors compared to per-frame processing methods [10,13].

Table 3 lists the stability errors and SIFID metrics of state-of-the-art (SOTA) methods and our approach. We verified our method and six other representative coherent video style transfer approaches on 20 different scenes from the MPI Sintel dataset [66]. All

the testing videos used the 1024 × 436 resolution, and the groundtruth optical flow with corresponding masks are all provided from MPI Sintel dataset [66]. The stability errors are re-calculated. As the dataset only provides the forward direction of optical flow, the image w^t in equation (12) warps the per-frame stylized result f_s^{t+1} at time $t+1$ back to t. Our method may achieve higher average stability errors than Wang et al. [36], but we achieved the lowest SIFID distance, which indicates that our transferred results are closer to the style reference image in terms of faithful texture patterns and colour preservation. Our method also achieves both lower average stability errors and SIFID scores than more recent SOTA methods such as Li et al. [35], Deng et al. [62], Wu et al. [37] and Gu et al. [38], which is consistent with the qualitative comparisons shown in Figures 14 and 15.

Table 2. Stability errors ($\times 10^{-2}$) of per-frame processing methods and our approach on five testing videos in each dataset. Style1: Candy, Style2: Mondrian. In MPI Sintel Dataset, there are five scenes used in this table, and they are S1: alley_1, S2: bamboo_2, S3: cave_4, S4: market_5, S5: temple_3. In Davis 2017 Dataset, there are five scenes used and they are S1: dance-flare, S2: car-turn, S3: parkour, S4: soapbox, S5: stroller. The best results are shown underlined.

Method	MPI Sintel Dataset + Style1					Davis 2017 Dataset + Style2				
	S1	**S2**	**S3**	**S4**	**S5**	**S1**	**S2**	**S3**	**S4**	**S5**
Johnson et al. [10]	8.15	7.91	11.96	13.06	14.50	12.73	12.63	13.01	13.62	14.11
Ours	3.13	3.56	6.82	7.88	8.25	4.08	4.42	4.85	4.92	5.13
Huang et al. [13]	9.16	10.11	14.51	14.32	14.44	15.50	14.52	15.43	15.77	16.28
Ours	3.46	4.39	7.66	8.28	8.38	5.00	4.99	5.86	5.90	6.14

Table 3. Quantitative comparison of stability errors ($\times 10^{-2}$) and SIFID metrics of six state-of-the-art methods and ours on 20 videos in MPI Sintel dataset. The groundtruth flow and occlusion masks are provided by MPI Sintel dataset. All the image resolutions are 1024 × 436. The baseline model of Ruder et al. [30] and ours is Johnson et al. [10]. Style1: Candy, Style2: The_Scream, Style3: Woman With A Hat. The best two results are highlighted in one underline and two underlines.

Method	Stability Errors (↓)				SIFID (↓)			
	Style1	**Style2**	**Style3**	**Mean**	**Style1**	**Style2**	**Style3**	**Mean**
Ruder et al. [30]	8.56	5.06	6.84	6.82	1.47	0.82	1.04	1.1117
Li et al. [35]	8.19	5.06	6.13	6.46	1.70	0.84	1.30	1.2772
Deng et al. [62]	9.72	5.26	6.89	7.29	2.09	1.00	1.62	1.5695
Wu et al. [37]	9.52	5.74	7.49	7.58	1.95	1.17	1.57	1.5632
Gu et al. [38]	8.17	5.34	5.47	6.33	1.97	1.04	1.63	1.5481
Ours	7.76	4.95	6.15	6.29	1.52	0.67	0.99	1.0615
Wang et al. [36]	6.68	3.57	5.75	5.34	1.94	1.76	1.40	1.7002

5.2.4. User Study

Furthermore, we carried out a user study experiment for comparison. To be specific, we used five videos and three style images to present fifteen video clips from seven methods to 50 participants. The demographic consisted of 27 males and 23 females, aged from 22 to 40. Then, we asked for their preferences over three aspects for each video: temporal consistency, style transformation effect (including the preservation of colour and texture patterns) and overall preference. In total, we received 150 votes for each video and 2250 votes from the 50 subjects. Figure 19 shows that our method obtained the majority of votes for style transformation effect and overall preference, but took second place in terms of temporal consistency compared to Wang et al. [36]. These results are consistent with Table 3.

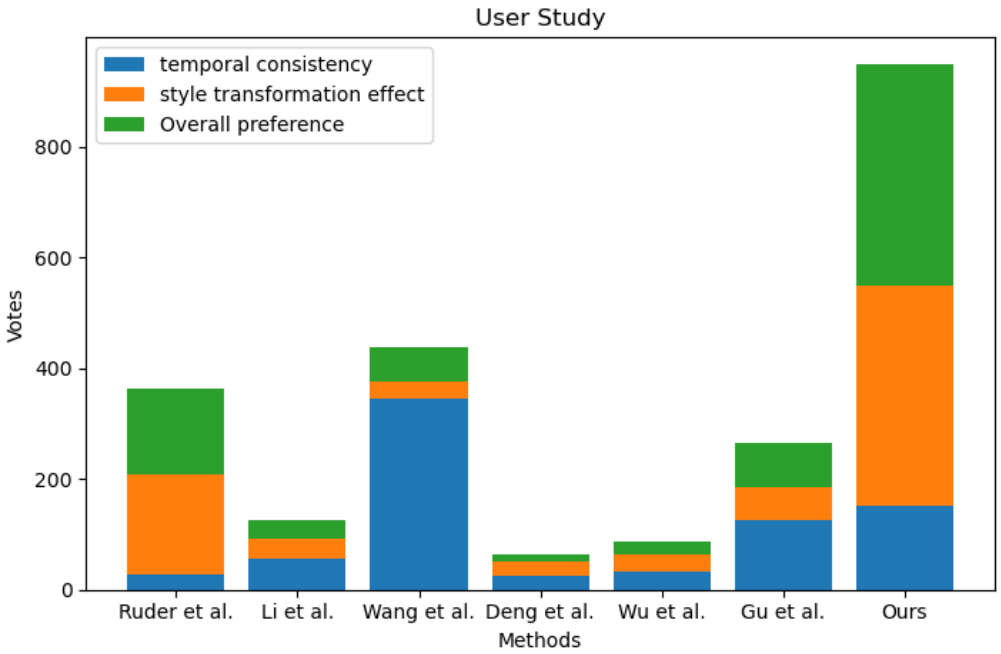

Figure 19. User Study. It is noticed that Wang et al. [36] received the most votes in terms of temporal consistency, but our proposed method obtained the most majority of votes for overall preference and style transformation effect compared to six other state-of-the-art methods. The number of votes for each method is consistent with Table 3 [30,35–38,62].

6. Conclusions

We discovered that previous state-of-the-art methods for video style transfer have difficulties in finding the balance between temporal consistency and faithful style effects in their stylized videos. To obtain the balance, we revisited the gradient-based optimization methods using optical flow, and proposed solutions to their problems, including flow errors and a low-speed run time. Our solutions consist of a new initialization strategy with proposed mask techniques that is presented to reduce flow errors and constraints with proposed losses that are performed to ensure temporal consistency and image quality over long-range frames. Specifically, we propose a multi-scale mask fusion scheme to reduce the flow untraceable errors, an incremental mask to reduce ghosting artefacts and a multi-frame mask fusion technique to reduce the flow traceable errors. The initialized images obtained via our mask techniques increase the running speed from over 3 min per frame to less than 2 s per frame for gradient-based optimization algorithms. In addition, we also introduced multi-frame Coherent Losses to enhance the temporal consistency on both RGB-level and Feature-level, and Sharpness Losses to effectively mitigate the image blurriness artefacts over times. In the end, the ablation studies validated that each component of the proposed framework serves their purposes, and experiments on public datasets demonstrate that our approach achieves a better balance between temporal consistency and faithful style effects compared to state-of-the-art methods. We notice that the proposed method still needs several seconds for each frame to perform style transformation, which may be a little bit slower in practice. We will further investigate offline feed-forward networks to quickly approximate solutions to our style transformation problem in the near future.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app14062630/s1.

Author Contributions: Conceptualization, L.W. and X.Y.; methodology, L.W.; software, L.W.; validation, L.W. and X.Y.; formal analysis, L.W.; resources, L.W.; data curation, L.W.; writing—original draft preparation, L.W.; writing—review and editing, X.Y. and J.Z.; visualization, L.W.; supervision, X.Y. and J.Z.; project administration, X.Y.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Acknowledgments: The authors would like to thank Weidong Min, Xi Li, Yulin Zhou and Chaoyi Pang for their helpful suggestions and discussions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Bombay, India, 4–8 January 1998; pp. 839–846.
2. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410.
3. Winnemöller, H.; Olsen, S.C.; Gooch, B. Real-time video abstraction. *Acm Trans. Graph. (TOG)* **2006**, *25*, 1221–1226. [CrossRef]
4. Yang, S.; Jiang, L.; Liu, Z.; Loy, C. Vtoonify: Controllable high-resolution portrait video style transfer. *ACM Trans. Graph. (TOG)* **2022**, *41*, 1–15. [CrossRef]
5. Li, Z.; Wu, X.M.; Chang, S.F. Segmentation using superpixels: A bipartite graph partitioning approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 789–796.
6. Liu, J.; Yang, W.; Sun, X.; Zeng, W. Photo stylistic brush: Robust style transfer via superpixel-based bipartite graph. *IEEE Trans. Multimed.* **2018**, *20*, 1724–1737. [CrossRef]
7. Lee, H.-Y.; Li, Y.-H.; Lee, T.-H.; Aslam, M.S. Progressively Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. *Sensors* **2023**, *23*, 6858. [CrossRef] [PubMed]
8. Dediu, M.; Vasile, C.E.; Bîră, C. Deep Layer Aggregation Architectures for Photorealistic Universal Style Transfer. *Sensors* **2023**, *23*, 4528. [CrossRef] [PubMed]
9. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2414–2423.
10. Johnson, J.; Alahi, A.; Li, F.F. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
11. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M.H. Diversified texture synthesis with feed-forward networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3920–3928.
12. Chen, D.; Yuan, L.; Liao, J.; Yu, N.; Hua, G. Stylebank: An explicit representation for neural image style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1897–1906.
13. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1501–1510.
14. Chen, T.Q.; Schmidt, M. Fast patch-based style transfer of arbitrary style. *arXiv* **2016**, arXiv:1612.04337.
15. Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.; Xu, C. A unified arbitrary style transfer framework via adaptive contrastive learning. *ACM Trans. Graph.* **2023**, *42*, 1–16. [CrossRef]
16. Zhang, Z.; Zhang, Q.; Li, G.; Xing, W.; Zhao, L.; Sun, J.; Lan, Z.; Luan, J.; Huang, Y.; Lin, H. ArtBank: Artistic Style Transfer with Pre-trained Diffusion Model and Implicit Style Prompt Bank. *arXiv* **2023**, arXiv:2312.06135.
17. Kwon, J.; Kim, S.; Lin, Y.; Yoo, S.; Cha, J. AesFA: An Aesthetic Feature-Aware Arbitrary Neural Style Transfer. *arXiv* **2023**, arXiv:2312.05928.
18. Chu, W.T.; Wu, Y.L. Image style classification based on learnt deep correlation features. *IEEE Trans. Multimed.* **2018**, *20*, 2491–2502. [CrossRef]
19. Yang, J.; Chen, L.; Zhang, L.; Sun, X.; She, D.; Lu, S.P.; Cheng, M.M. Historical context-based style classification of painting images via label distribution learning. In Proceedings of the ACM Multimedia Conference on Multimedia Conference, Seoul, Republic of Korea, 22–26 October 2018; pp. 1154–1162.

20. Hicsonmez, S.; Samet, N.; Sener, F.; Duygulu, P. Draw: Deep networks for recognizing styles of artists who illustrate children's books. In Proceedings of the ACM on International Conference on Multimedia Retrieval, Sydney, Australia, 6–11 August 2017; pp. 338–346.
21. Zhou, X.; Liu, Z.; Gong, C.; Liu, W. Improving video saliency detection via localized estimation and spatiotemporal refinement. *IEEE Trans. Multimed.* **2018**, *20*, 2993–3007. [CrossRef]
22. Bak, C.; Kocak, A.; Erdem, E.; Erdem, A. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Trans. Multimed.* **2018**, *20*, 1688–1698. [CrossRef]
23. Anderson, A.G.; Berg, C.P.; Mossing, D.P.; Olshausen, B.A. Deepmovie: Using optical flow and deep neural networks to stylize movies. *arXiv* **2016**, arXiv:1605.08153.
24. Ruder, M.; Dosovitskiy, A.; Brox, T. Artistic style transfer for videos. In Proceedings of the German Conference on Pattern Recognition, Hannover, Germany, 12–15 September 2016; pp. 26–36.
25. Zhang, H.; Dana, K. Multi-style generative network for real-time transfer. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
26. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6924–6932.
27. Huang, H.; Wang, H.; Luo, W.; Ma, L.; Jiang, W.; Zhu, X.; Li, Z.; Liu, W. Real-time neural style transfer for videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 783–791.
28. Gupta, A.; Johnson, J.; Alahi, A.; Li, F.F. Characterizing and improving stability in neural style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4067–4076.
29. Chen, D.; Liao, J.; Yuan, L.; Yu, N.; Hua, G. Coherent online video style transfer. In Proceedings of the IEEE Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1105–1114.
30. Ruder, M.; Dosovitskiy, A.; Brox, T. Artistic style transfer for videos and spherical images. *Int. J. Comput. Vis.* **2018**, *126*, 1199–1219. [CrossRef]
31. Xu, K.; Wen, L.; Li, G.; Qi, H.; Bo, L.; Huang, Q. Learning self-supervised space-time CNN for fast video style transfer. *IEEE Trans. Image Process. (TIP)* **2021**, *30*, 2501–2512. [CrossRef] [PubMed]
32. Liu, S.; Zhu, T. Structure-guided arbitrary style transfer for artistic image and video. *IEEE Trans. Multimed.* **2021**, *24*, 1299–1312. [CrossRef]
33. Kong, X.; Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Chen, Y.; He, Z.; Xu, C. *Exploring the Temporal Consistency of Arbitrary Style Transfer: A Channelwise Perspective*; IEEE Transactions on Neural Networks and Learning Systems: Piscataway, NJ, USA, 2023; pp. 1–15.
34. Huo, J.; Kong, M.; Li, W.; Wu, J.; Lai, Y.; Gao, Y. Towards efficient image and video style transfer via distillation and learnable feature transformation. *Comput. Vis. Image Underst.* **2024**, *241*, 103947. [CrossRef]
35. Li, X.; Liu, S.; Kautz, J.; Yang, M.H. Learning linear transformations for fast image and video style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3809–3817.
36. Wang, W.J.; Yang, S.; Xu, J.Z.; Liu, J.Y. Consistent Video Style Transfer via Relaxation and Regularization. *IEEE Trans. Image Process. (TIP)* **2020**, *29*, 9125–9139. [CrossRef] [PubMed]
37. Wu, Z.; Zhu, Z.; Du, J.; Bai, X. CCPL: Contrastive coherence preserving loss for versatile style transfer. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 189–206.
38. Gu, B.H.; Fan, H.; Zhang, L.B. Two Birds, One Stone: A Unified Framework for Joint Learning of Image and Video Style Transfers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 23545–23554.
39. Gatys, L.A.; Ecker, A.S.; Bethge, M. Texture synthesis using convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canad, 7–10 December 2015; pp. 262–270.
40. Gatys, L.A.; Ecker, A.S.; Bethge, M. A neural algorithm of artistic style. *arXiv* **2015**, arXiv:1508.06576.
41. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M.H. Universal style transfer via feature transforms. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 385–395.
42. Wang, X.; Oxholm, G.; Zhang, D.; Wang, Y.F. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5239–5247.
43. Wilmot, P.; Risser, E.; Barnes, C. Stable and controllable neural texture synthesis and style transfer using histogram losses. *arXiv* **2017**, arXiv:1701.08893.
44. Shen, F.; Yan, S.; Zeng, G. Meta networks for neural style transfer. *arXiv* **2017**, arXiv:1709.04111.
45. Ulyanov, D.; Lebedev, V.; Vedaldi, A.; Lempitsky, V.S. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv* **2016**, arXiv:1603.03417.
46. Yao, Y.; Ren, J.; Xie, X.; Liu, W.; Liu, Y.J.; Wang, J. Attention-aware multi-stroke style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1467–1475.

47. Kotovenko, D.; Sanakoyeu, A.; Ma, P.; Lang, S.; Ommer, B. A content transformation block for image style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10032–10041.
48. Kolkin, N.; Salavon, J.; Shakhnarovich, G. Style transfer by relaxed optimal transport and self-similarity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10051–10060.
49. Wang, H.; Li, Y.; Wang, Y.; Hu, H.; Yang, M.-H. Collaborative distillation for ultra-resolution universal style transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1860–1869.
50. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
51. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 12–13 December 2014; pp. 2366–2374.
52. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 2650–2658.
53. Chen, H.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D. Artistic style transfer with internal-external learning and contrastive learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 26561–26573.
54. Deng, Y.; Tang, F.; Dong, W.; Sun, W.; Huang, F.; Xu, C. Arbitrary style transfer via multi-adaptation network. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2719–2727.
55. Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Ding, E. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 6649–6658.
56. Luo, X.; Han, Z.; Yang, L.; Zhang, L. Consistent style transfer. *arXiv* **2022**, arXiv:2201.02233.
57. 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 (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5880–5888.
58. Wu, X.; Hu, Z.; Sheng, L.; Xu, D. Styleformer: Real-time arbitrary style transfer via parametric style composition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 14618–14627.
59. Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; Xu, C. Stytr2: Image style transfer with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 11326–11336.
60. Bosse, S.; Maniry, D.; Müller, K.R.; Wiegand, T.; Samek, W. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.* **2017**, *27*, 206–219. [CrossRef] [PubMed]
61. Lai, W.S.; Huang, J.B.; Wang, O.; Shechtman, E.; Yumer, E.; Yang, M.H. Learning blind video temporal consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 170–185.
62. Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Xu, C. Arbitrary video style transfer via multi-channel correlation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; pp. 1210–1217.
63. Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; Brox, T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2462–2470. Available online: http://lmb.informatik.uni-freiburg.de//Publications/2017/IMKDB17 (accessed on 20 September 2017).
64. Gao, C.; Gu, D.; Zhang, F.; Yu, Y. Reconet: Real-time coherent video style transfer network. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 637–653.
65. Yang, X.; Zhang, T.; Xu, C. Text2video: An end-to-end learning framework for expressing text with videos. *IEEE Trans. Multimed.* **2018**, *20*, 2360–2370. [CrossRef]
66. Butler, D.J.; Wulff, J.; Stanley, G.B.; Black, M.J. A naturalistic open source movie for optical flow evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 611–625.
67. Weinzaepfel, P.; Revaud, J.; Harchaoui, Z.; Schmid, C. Deepflow: Large displacement optical flow with deep matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 25–27 June 2013; pp. 1385–1392.
68. Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; Van Gool, L. The 2017 davis challenge on video object segmentation. *arXiv* **2017**, arXiv:1704.00675.
69. Gu, K.; Zhai, G.; Lin, W.; Yang, X.; Zhang, W. No-reference image sharpness assessment in autoregressive parameter space. *IEEE Trans. Image Process.* **2015**, *24*, 3218–3231.
70. Vu, P.V.; Chandler, D.M. A fast wavelet-based algorithm for global and local image sharpness estimation. *IEEE Signal Process. Lett.* **2012**, *19*, 423–426. [CrossRef]