LEARNING LONG TERM STYLE PRESERVING BLIND VIDEO TEMPORAL CONSISTENCY

Hugo Thimonier*, Julien Despois*, Robin Kips*†, Matthieu Perrot*

* L’Oréal Research and Innovation, France
† LTCI, Télécom Paris, Institut Polytechnique de Paris, France
{firstname.lastname}@rd.loreal.com

ABSTRACT
When trying to independently apply image-trained algorithms to successive frames in videos, noxious flickering tends to appear. State-of-the-art post-processing techniques that aim at fostering temporal consistency, generate other temporal artifacts and visually alter the style of videos. We propose a post-processing model, agnostic to the transformation applied to videos (e.g. style transfer, image manipulation using GANs, etc.), in the form of a recurrent neural network. Our model is trained using a Ping Pong procedure and its corresponding loss, recently introduced for GAN video generation, as well as a novel style preserving perceptual loss. The former improves long-term temporal consistency learning, while the latter fosters style preservation. We evaluate our model on the DA VIS and videvo.net datasets and show that our approach offers state-of-the-art results concerning flicker removal, and better keeps the overall style of the videos than previous approaches.

Index Terms— Time consistency, deep learning, video post-processing, style preserving perceptual loss

1. INTRODUCTION
Since the first introduction of deep convolutional neural networks (CNN) and the rise of GANs, more and more powerful image processing applications have been proposed. A still popular topic in the literature discusses how these image processing techniques can be efficiently extended to videos.

A natural way to transfer image processing algorithms to videos is to perform such transformation frame by frame. However, the latter scheme necessarily causes unpleasant visual artifacts (e.g. flickering) originating from temporal inconsistency. For instance, successive frames are not necessarily transformed coherently since image trained algorithms fail to capture the temporal dimension of videos. The literature offers several types of approaches to circumvent such problems: video-to-video translation algorithms and post-processing techniques of per-frame transformed videos. The former consists in conditional video generation algorithms aiming at generating temporally coherent outputs, while the latter post-processes videos transformed using image-trained algorithms to make them temporally coherent. On the one hand, video-to-video translation algorithms exhibit interesting performances even for relatively high-resolution videos, however, they require video datasets which are very scarce especially for supervised algorithms. On the other hand, post-processing techniques offer satisfactory results and do not necessitate any retraining of image processing algorithms, but often fail to reduce localized flickering without generating new visual artifacts and tend to modify the overall color of videos.

We propose a blind post-processing algorithm in the form of a recurrent neural network. The main contributions involved in the model are:

• Following up on recent work [2] in the image style transfer literature, we introduce a novel style preserving perceptual loss which reduces deviation from the overall style of the original processed video.

• We propose to enforce long term consistency by adapting the Ping Pong training procedure [3] to this task and its corresponding Ping Pong Loss.
• We empirically highlight some weaknesses of the perceptual distance as a measure of style preservation.

• Our model offers state-of-the-art results in flicker removal for any transformation and does not deteriorate the visual quality of videos when enforcing time consistency, nor generate new temporal artifacts.

2. BACKGROUND AND RELATED WORK

When independently applying image-trained algorithms (e.g. image-trained style transfer algorithms) to successive frames in a video, several temporal artifacts tend to appear. In that context, the recent literature has focused on two approaches to tackle the problem of video processing and temporal inconsistency.

The first approach, namely video-to-video translation, consists in adapting image processing algorithms to take into account the temporal dimension of the input. Most approaches involve computing optical flows during training and consider networks taking several inputs including the current frames (raw and stylized), the previous output frame, and the previous stylized frame warped using the estimated optical flow [4]. Optical flow estimation is computationally costly, thus, later approaches also considered similar methods but circumvent the problem of computing optical flows during the inference stage such as [5]. Recent approaches offer satisfying results but tend to require complex training procedures such as a two-stage progressive training strategy and need to be retrained for each new application considered.

The second method, task-independent post-processing, aims at considering time consistency in a more agnostic manner not focusing on any type of image processing. Blind post-processing consists in (i) first transforming a video using any algorithm and (ii) using an algorithm, independent of the transformation previously performed, to make the video temporally consistent. This approach requires to consider a twofold problem: on the one hand, videos must be made temporally coherent, on the other hand, the style of videos must not be altered. Recent papers have obtained satisfying results using general approaches totally blind to the filter applied to videos. For instance, [6] and [7] display very satisfactory results, however, the computational complexity of their proposed algorithm makes it difficult to process long video sequences. More recently, in [11] the authors propose a neural network to enforce time consistency in videos. The model takes as an input the raw unprocessed video frames, the processed video frames, and the previous output of the recurrent network. The model also makes use of optical flow estimation for temporal consistency loss computation and of the VGG19 [8] network for perceptual loss computation. Their model offers interesting results but fails to coherently learn long-term temporal consistency and generates new temporal artifacts when the motion of an object changes direction. Moreover, their model tends to alter the overall style of the video once post-processed, thus reducing the quality of the video transformation. Recent work in the style transfer literature has offered possible solutions to foster style preservation when learning temporal coherence.

Neural style transfer consists in transferring the style of an image (e.g. a painting) to another one while preserving the content of the latter image. Style transfer in the form of deep neural networks was first introduced in [9]. Their seminal model makes use of the VGG19 [8] network to perform style transfer through both a content and style loss. Their approach opened a wide range of applications making use of deep features (e.g. VGG19 [8]) as measures of similarity between images. Recent approaches such as [2] have offered a better balance between the structure and the style of the output by using moments of different orders (mean and standard deviation) of the pre-trained features.

3. PROPOSED METHOD

3.1. Task

We argue that post-processing techniques can manage to efficiently remove any temporal artifacts on videos originating from per-frame processing. This approach is all the more appealing since it needs not any retraining for each task; this is a key aspect since it can be challenging to find a satisfying equilibrium in the case of video-to-video translation algorithms. We thus propose a blind post-processing model in the form of a recurrent neural network which improves current state-of-the-art results. Our approach does not require optical flow estimation during inference but necessitates to dispose of the original non-transformed video.

3.2. Architecture

The deep network is comprised of a classical encoder-decoder architecture. The encoder part is composed of two downsampling strided convolutional layers, each followed by Instance Normalization. The encoder is then followed by 5 residual blocks and a ConvLSTM module [10]. The decoder placed after the ConvLSTM module is composed of two transposed convolutional layers also followed by Instance Normalization. We also include skip-connections via concatenation from the encoder to the decoder.

ConvLSTM modules are well-suited to our problem since they can compress all the previous information into a hidden state which can be used to estimate the current state. This hidden state can capture the spatial information of the whole input sequence and allows the model to learn coherently time consistency. In tandem with temporal losses, ConvLSTM modules were shown to offer very satisfactory results in fostering temporal consistency on video style transfer networks.

Let us denote the original unprocessed frames \{I_t\}_{t=1,...,T}, the per-frame processed videos \{P_t\}_{t=1,...,T},
and \( \{O_t\}_{t=1,...,T} \) the corresponding outputs of the model. Since we consider residual blocks in our network, \( O_t \) can be seen as a residual transformation of \( P_t \) to make it temporally consistent. To output \( O_t \), the network takes as an input \( I_t, I_{t-1}, P_t \) and \( O_{t-1} \). Formally,

\[
O_t = P_t + F(I_t, I_{t-1}, P_t, O_{t-1})
\]

where \( F(.) \) denotes our network. The recurrent form of the network thus originates from the fact that our network makes use of the previous output to generate the current output. Let us note here that for of each sequence \( P_1 = O_1 \).

3.3. Losses

The aim of the proposed model is to reduce the temporal inconsistency while maintaining the perceptual similarity in both content and style between the output frame of the model and the processed frame. Therefore, two types of loss must be considered: perceptual losses and temporal losses.

3.3.1. Perceptual losses

The pre-trained VGG19 classification network is vastly used in a large number of application as it has been shown to be a good measure of human perception. To preserve the content of the video, as in [1], we set the content perceptual loss to be

\[
\mathcal{L}_p = \sum_{t=2}^{T} \sum_{i=1}^{N} \left\| \phi_l(O_i(t)) - \phi_l(P_i(t)) \right\|_2
\]

where \( O_i(t) \in \mathbb{R}^3 \) represent the RGB vector of pixel \( i \) in the output frame \( O_t \) composed of \( N \) pixels, while \( \phi_l(.) \) represents the feature activation of the \( l \)-th block of the VGG19 network \( \phi \) (we consider a sum of all feature activation until the 4th block, i.e. relu1-2, relu2-2, relu3-3 and relu4-3).

To ensure that our model does not alter the style of the video, we also introduce a novel loss which makes use of the VGG network discussed above. For instance, we consider

\[
\mathcal{L}_{sp} = \sum_{t=2}^{T} \sum_{i=1}^{N} \left\{ \| \mu(\phi_i(O_t)) - \mu(\phi_i(P_t)) \|^2_2 + \| \sigma(\phi_i(O_t)) - \sigma(\phi_i(P_t)) \|^2_2 \right\}
\]

where \( \mu(.) \) denotes the average in the channel dimension and \( \sigma(.) \) the standard deviation. Such approach is well suited to our case since pixel-wise comparisons between \( P_t \) and \( O_t \) can be too conservative, while considering only moments of the first and second order tend to capture well global style of a frame. We also consider its temporal equivalent to avoid excessive deviations from the global style of the raw processed videos. For instance, since images are generated in an online manner, minor deviation from the style of the video tend to accumulate as frames are being processed.

\[
\mathcal{L}_{sp}^{temp} = \sum_{t=2}^{T} \sum_{i=1}^{N} \left\{ \| \mu(\phi_i(O_t)) - \mu(\phi_i(O_{t-1})) \|^2_2 + \| \sigma(\phi_i(O_t)) - \sigma(\phi_i(O_{t-1})) \|^2_2 \right\}
\]

In both losses, the feature activation considered are relu1-2 and relu2-2. We thus consider the following style preserving perceptual loss

\[
\mathcal{L}_{SP} = \mathcal{L}_{sp} + \mathcal{L}_{sp}^{temp}
\]

3.3.2. Temporal losses

During training, we randomly fetch a sequence of \( k \) successive frames in a video which we transform to construct a temporal cycle centered around the last frame of the selected sequence. For instance, for each randomly selected sequence in the training dataset \( \{P_t, \ldots, P_{t+k-1}, P_{t+k}\} \) the following ping pong sequence is constructed \( \{P_t, \ldots, P_{t+k-1}, P_{t+k}, P_{t+k-1}, \ldots, P_t\} \). In the recurrent framework detailed in the previous section, one obtains a similar sequence of output frames

\[
\{O_t, \ldots, O_{t+k-1}, O_{t+k}, O'_{t+k-1}, \ldots, O'_{t}\}
\]

where frames with \( \sim \) denote frames constructed in a backward direction. In other words, \( O'_t \) is estimated by our network taking as inputs \( I_t, I_{t+1}, O'_{t+1} \) and \( P_t \). Considering such procedure, [3] propose the following Ping Pong loss

\[
\mathcal{L}_{PP} = \sum_{t=1}^{k-1} \| O_t - O'_t \|_2
\]

Such loss which combines short term consistency for frames close to \( n \), e.g. \( \| O_{n-1} - O'_{n-1} \|_2 \) and long term consistency for factors far from \( n \), e.g. \( \| O_t - O'_t \|_2 \), should reduce the accumulation of unwanted features throughout the video.

The short term temporal error is based on the warping error of successive output frames. Notice that the Ping Pong training procedure modifies classical short term temporal loss by involving backward warping error

\[
\mathcal{L}_{st} = \sum_{t=2}^{T} \sum_{i=1}^{N} \left\{ \| O_i(t) - warp(O_{t-1}, F_{t,t-1})(i) \|_1 + M_i(t-1) \| O'_i(t) - warp(O'_{t+1}, F_{t,t+1})(i) \|_1 \right\}
\]

where \( warp(O_{t-1}, F_{t,t-1}) \) is the \( O_{t-1} \) frame warped to time \( t \) using the backward optical flow \( F_{t,t-1} \) estimated between \( I_t \) and \( I_{t-1} \). \( M_i(t-1) = \exp(-\alpha||I_i - warp(I_{t-1}, F_{t,t-1})||_2^2) \) is the visibility mask calculated from the warping error between the input frame \( I_t \) and the warped input frame \( warp(I_{t-1}, F_{t,t-1}) \) where \( \alpha \) is set to be 50 [1]. Optical flows are estimated using FlowNet2 [11].
The previously described loss is used to enforce consisten-
ty between adjacent frames, however there is no guaran-
tee regarding long-term consistency. The proposed long-term
temporal loss consists in the warping error between the first
frame of the sequence and all of the output frames of the
forward sequence,
\[ L_{it} = \sum_{i=2}^{N} \sum_{t=1}^{T} M_{i}^{(i)}(O_{t}^{i} - warp(O_{1}, F_{t}, t)^{(i)}) \|_{1} \] (6)

Following up on [12] we formulate the hypothesis on sta-
ble stylized/processed videos: non-occluded regions should
be low-rank representations. Consider during training (i) a
sequence of \( k \) consecutive frames, \( \{ I_{t}, \ldots, I_{t+k-1}, I_{t+k} \} \) (ii)
a reference time (which we simply set to \( k/2 \)) to which all
frames are warped (iii) and occlusion masks at each time step.
Denoting \( R_{t} \) the Hadamard product (element-wise product)
between the occlusion mask and the warped frame, consider
\( \chi = [ vec(R_{0}), \ldots, vec(R_{k}) ]^{T} \in \mathbb{R}^{k \times N} \) where \( N = H \times W \)
the number of pixels in the image and \( vec(\cdot) \) simply denotes
the flattening operator of a two-dimensional image. Based on
the formulated hypothesis, the rank of \( \chi \) constructed using
the raw input frames \( I_{k} \) and using the output frames of
the model \( O_{k} \), \( \chi_{O} \) should not be too different from each other. [12] propose the low rank loss using the convex relaxation of
the rank defined as
\[ L_{\text{rank}} = (||\chi_{I}||_{*} - ||\chi_{O}||_{*})^{2} \] (7)

where \( ||\cdot||_{*} \) is the nuclear norm defined as the sum of all sin-
gular values of a matrix.

### 3.3.3. Overall loss

The overall loss is a weighted sum of all previously defined losses
\[ \mathcal{L} = \lambda_{p} L_{p} + \lambda_{SP} L_{SP} + \lambda_{st} L_{st} + \lambda_{lt} L_{lt} + \lambda_{\text{rank}} L_{\text{rank}} + \lambda_{PP} L_{PP} \] (8)

where \( \lambda_{p}, \lambda_{SP}, \lambda_{st}, \lambda_{lt}, \lambda_{\text{rank}} \) and \( \lambda_{PP} \) are hyper-parameters
to be defined, that represent the weight given to each loss in
the overall training of the network.

### 4. EXPERIMENTAL RESULTS

#### 4.1. Evaluation metrics

Two features must be evaluated to assess the quality of the
proposed model: (i) the output frames \( O_{t} \) must be as close
as possible to unprocessed frames \( I_{t} \) in terms of time consisten-
ty and (ii) the output frames \( O_{t} \) must visually look like the
input processed frames \( P_{t} \). Two types of metrics are used in
the literature to evaluate the quality of a model: a time consis-
tency metric based on the warping error and a perceptual
similarity metric.

The literature dealing with time consistency has offered
several metrics to measure time consistency in a video. The
most vastly used is the mean warping error between succes-
sive frames of a video. The latter is defined as the difference
between a warped frame to time \( t \) and the true frame at time \( t \).
Formally, the warping error between two frames \( V_{t} \) and \( V_{t+1} \) is defined as
\[ \omega_{t,t+1}^{V} = \frac{1}{M} \sum_{i=1}^{N} M_{i}^{(i)} \| V_{t}^{(i)} - warp(V_{t+1}, F_{t}, t+1)^{(i)} \|_{2}^{2} \]
where \( M_{t} \in \{0,1\}^{N} \) is the non-occluded regions, \( N \) is the number of pixels in each frame
and \( M = \sum_{i=1}^{N} M_{i}^{(i)} \). To estimate those regions, we
resort to the method given in [4] which consists in performing
a backward-forward consistency check of the optical flows
based on the method introduced in [13]. To obtain a metric
for a whole video \( V \), one averages the warping error between
successive frames over the sequence.

Perceptual similarity is evaluated using the LPIPS metric
proposed in [14] to evaluate the perceptual similarity between
the raw processed and output videos. In [1], the authors resort
to the LPIPS metric proposed in [14] and the SqueezeNet [15]
(noted \( S \)) to evaluate the perceptual similarity between
the raw processed videos \( P \) and output \( O \) videos.

\[ D_{\text{perceptual}}(P, O) = \frac{1}{T-1} \sum_{t=2}^{T} S(O_{t}, P_{t}) \]

We discuss this metric more deeply in the next section and
argue that this metric fails to correctly capture localized style
deviations or brightness alteration.

#### 4.2. Datasets and training

For comparability and convenience, we use the dataset made
available by [1] which is constituted of videos from DAVIS-
2017 dataset [16] and videvo.net videos. The height of each
video in the training set is scaled to 480 while preserving the
aspect ratio. The training set contains in total 25,735 frames.
The applications considered in the dataset are the following:
aristic style transfer, Colorization, Image enhancement, In-
trinsic image decomposition, and Image-to-image translation.

Results presented hereafter are those obtained after train-
ing with values of lambdas set empirically: \( \lambda_{p} = 10, \lambda_{SP} =
10, \lambda_{st} = 100, \lambda_{lt} = 100, \lambda_{PP} = 100, \lambda_{\text{rank}} = 0.00001 \).
We set the number of epoch to 100, each composed of 1000
batches of size 4. The sequences considered in every batch
are composed of 5 successive frames each, and thus 9 once
the Ping Pong sequence is constructed. We consider Adam
optimizer for parameter optimization.

#### 4.3. Qualitative evaluation

We propose to compare our model to the current state-of-the-
art post-processing technique of [1]. The latter model suc-
cessfully removes flickering artifacts in most cases but suffers from two drawbacks: (i) it tends to generate new temporal artifacts when flickering is localized on videos or when the motion of an object changes direction and (ii) the overall color of the video is often modified after using their post-processing method. An example of (i) can be seen in Fig. 2a where make-up synthesis was applied which generates flickering on the lips. Once post-processed using the model of [1] a trail following the lips appears. We believe this type of artifacts originate from the form of the long-term loss considered in their model which forces the model to be coherent with the first frames of the sequence.

Our model visually outperforms the current state-of-the-art model in terms of style conservation and flicker removal. For instance Fig. 2b shows that our model successfully removes flickering and does not generate trails as [1]. Similarly Fig. 1 shows how our model manages to better remove flickering while not suffering from perceptual deterioration of the overall style of the video. For instance, in Fig. 1 one can see the videos post-processed using [1] becoming darker frame after frame, while our model coherently keeps the overall look of the video while not altering the brightness nor the style of the video. More examples showing the performance of our model are available in the supplementary material.

4.4. Quantitative comparisons

To evaluate the temporal coherence of videos we resort to the warping error metric previously discussed. Table 1 shows that our model outperforms the model of [1] for both DAVIS and video.net datasets for a large majority of applications when the style preserving loss is considered and for all test applications considered when not including the style preserving loss. Let us note that we obtain better results than [6] on applications where they outperformed [1]. Also note that the average warping error for the raw processed frame are 0.052 for the DAVIS dataset and 0.053 for the video.net dataset.

### Table 1: Warping Error of post-processing models

| Task            | DAVIS Average | Lai et al. | Ours | Ours + SP Loss | Lai et al. | Ours | Ours + SP Loss |
|-----------------|---------------|-----------|------|----------------|-----------|------|---------------|
| WCT/favimono    | 0.031         | 0.017     | 0.024| 0.022          | 0.013     | 0.019|
| WCT/techvillage | 0.059         | 0.029     | 0.043| 0.047          | 0.024     | 0.037|
| WCT/family      | 0.045         | 0.025     | 0.039| 0.034          | 0.020     | 0.031|
| WCT/feathers    | 0.039         | 0.022     | 0.036| 0.034          | 0.021     | 0.034|
| WCT/Ticket      | 0.030         | 0.015     | 0.024| 0.030          | 0.014     | 0.023|
| WCT/Asheville   | 0.036         | 0.020     | 0.052| 0.057          | 0.017     | 0.027|
| Fast-neutral-style/fruits | 0.060   | 0.058   | 0.060| 0.060          | 0.035     | 0.065|
| Fast-neutral-style/edible | 0.023   | 0.012   | 0.017| 0.017          | 0.009     | 0.033|
| DBL/expertA      | 0.013         | 0.008     | 0.011| 0.012          | 0.008     | 0.011|
| DBL/expertB      | 0.011         | 0.007     | 0.009| 0.009          | 0.006     | 0.010|
| Intrinsics/illumination | 0.010 | 0.006     | 0.009| 0.010          | 0.007     | 0.010|
| Intrinsics/shading | 0.009         | 0.005     | 0.007| 0.007          | 0.005     | 0.007|
| CycleGAN/photos21kmono | 0.020         | 0.013     | 0.019| 0.017          | 0.012     | 0.018|
| CycleGAN/photos21kycolor | 0.027         | 0.020     | 0.028| 0.025          | 0.017     | 0.028|
| Colorization     | 0.012         | 0.007     | 0.009| 0.009          | 0.006     | 0.008|
| Colorization     | 0.010         | 0.006     | 0.008| 0.008          | 0.006     | 0.008|
| Average          | 0.027         | 0.015     | 0.023| 0.023          | 0.014     | 0.021|

When improving temporal consistency in a video, one usually faces a trade-off between making a video more coherent and not altering the overall style of the video. The metric presented in section 4.1 fails to fully capture style conservation. For instance, one major drawback of this metric is its inability to correctly account for new temporal artifacts appearing: in Fig. 2 and the video available in the supplementary material one can see that the post-processing model of [1] generates a trail following the lips, decreasing the resemblance between the original processed video and the post-processed video. However, the perceptual distance seems to favor [1] rather than our proposed model despite such phenomenon (0.026 for the post-processing model of [1] and 0.052 for our proposed model). Similarly, [1] tends to alter the overall style of videos by either changing the contrast or the brightness of videos, while our model successfully avoids such pitfalls (see Fig. 4 where both post-processing models obtain a perceptual distance of 0.038 over the whole video) but the metric can still favor the competing model. We believe that the metric only captures blurriness and major color drift, but fails to measure true resemblance and style preservation. Our model obtains an average perceptual distance over all applications of respectively 0.058 and 0.056 with and without our style preserving perceptual loss for the DAVIS dataset, as well as 0.051 and 0.053 for the video.net dataset, while [1] obtained respectively 0.017 and 0.012.

### 4.5. Discussion

As seen in Fig. 3 we obtain the lowest temporal warping error at the cost of a slightly larger perceptual distance than [1]. Note that when making their hyperparameters vary, [1] manage to improve their temporal coherence at the cost of a much higher perceptual distance, but they never attain as low perceptual distance as our model. The blue dots in Fig. 5 correspond to the model of [1] for several sets of hyperparameters. When including the style preserving perceptual...
Both models are compared using the temporal warping error and the perceptual distance for the application fast-neural-style/udnie. Our model significantly outperforms [1] in terms of warping error and obtains slightly higher perceptual distance.

loss, we lightly deteriorate both the perceptual distance and temporal warping error. However as seen in Fig. 4 the style preserving perceptual loss reduces deterioration of the brightness throughout videos and style deviation. This supports the idea that the perceptual distance fails to fully capture style preservation. For instance, both our model and the model of [1] can alter the brightness of videos and our proposed loss manages to reduce such issue.

Fig. 4: Without our style preserving perceptual loss the video tends to become darker when post-processed.

5. CONCLUSION

We discussed questions regarding temporal consistency linked to independently processing videos frame by frame using image trained algorithms. We put forward a new model to remove flickering artifacts, it includes recent techniques proposed in approaches orthogonal to our work in the field of video processing, as well as a novel loss to circumvent problems linked to style deterioration throughout videos. Thanks to these modifications, we obtain state-of-the-art results regarding flickering removal and also improve the visual output of the post-processed videos. We believe that future work in the field of video temporal post-processing ought to focus on algorithms that do not necessitate to have the original non-flickering video for inference.

6. REFERENCES

[1] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang, “Learning blind video temporal consistency,” 2018.
[2] Dae Young Park and Kwang Hee Lee, “Arbitrary style transfer with style-attentional networks,” 2019.
[3] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey, “Learning temporal coherence via self-supervision for gan-based video generation,” ACM Transactions on Graphics, vol. 39, no. 4, Jul 2020.
[4] Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox, “Artistic style transfer for videos,” Pattern Recognition, p. 26–36, 2016.
[5] Wei Gao, Yijun Li, Yihang Yin, and Ming Hsuan Yang, “Fast video multi-style transfer,” in Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020, United States, Mar. 2020, Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020, pp. 3211–3219, Institute of Electrical and Electronics Engineers Inc., 2020 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2020; Conference date: 01-03-2020 Through 05-03-2020.
[6] Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister, “Blind video temporal consistency,” ACM Transactions on Graphics, vol. 34, no. 6, pp. 196:1–196:9, Nov. 2015.
[7] Chun-Han Yao, Chia-Yang Chang, and Shao-Yi Chien, “Occlusion-aware video temporal consistency,” 10 2017, pp. 777–785.
[8] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014.
[9] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks.” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414–2423.
[10] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 802–810. Curran Associates, Inc., 2015.
[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 2017.
[12] Xinghao Chen, Yinan Zhang, Yunhe Wang, Han Shu, Chunjing Xu, and Chang Xu, “Optical flow distillation: Towards efficient and stable video style transfer,” 2020.
[13] N. Sundaram, T. Brox, and K. Keutzer, “Dense point trajectories by gpu-accelerated large displacement optical flow,” in *European Conference on Computer Vision (ECCV)*. Sept. 2010, Lecture Notes in Computer Science, Springer.

[14] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” 2018.

[15] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size,” *CoRR*, vol. abs/1602.07360, 2016.

[16] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, “A benchmark dataset and evaluation methodology for video object segmentation,” in *Computer Vision and Pattern Recognition*, 2016.