LEARNING TASK AGNOSTIC TEMPORAL CONSISTENCY CORRECTION

Anonymous authors
Paper under double-blind review

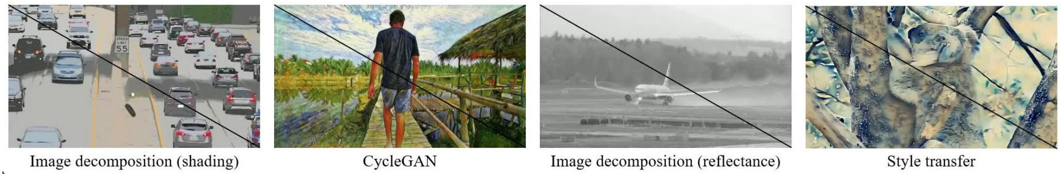


Image decomposition (shading)
CycleGAN
Image decomposition (reflectance)
Style transfer

Figure 1: **Applications of the proposed method.** The proposed algorithm takes only the per-frame processed videos with severe temporal flicker (bottom-left) and produces temporally consistent results (top-right). The proposed method is agnostic to the image processing operations used to process videos, and it does not require the availability of raw/unprocessed videos. The figure presented above contains animated content and is best viewed on a computer screen with Adobe PDF reader.

ABSTRACT

In many video restoration/translation tasks, image processing operations are naively extended to the video domain by processing each frame independently. This disregard for the temporal connection in video processing often leads to severe temporal inconsistencies. State-of-the-art techniques that address these inconsistencies rely on the availability of unprocessed videos to siphon consistent video dynamics to restore the temporal consistency of frame-wise processed videos. We propose a novel general framework for this task that learns to infer consistent motion dynamics from inconsistent videos to mitigate the temporal flicker while preserving the perceptual quality for both the temporally neighboring and relatively distant frames. The proposed framework produces state-of-the-art results on two benchmark datasets, DAVIS and videvo.net, processed by numerous image processing applications in a frame-wise processing manner. The code and the trained models will be released upon acceptance.

1 INTRODUCTION

Video sharing social media platforms like Snapchat and TikTok have introduced the common populace to a plethora of computer vision applications such as Style Transfer (Gatys et al., 2015), Colorization (Zhang et al., 2016b), Denoising (Zhang et al., 2017), and Dehazing (Yan et al., 2016). With the wide-scale integration of classic computer vision applications in such platforms, various image processing operations are naively extended to videos due to scarcity of annotated video dataset and their computational complexity. This naive extension of static image processing methodologies to videos disregards the temporal connection of the video progression and introduces severe temporal flickering in the videos. This temporal flicker can appear for various reasons; for instance, these image processing methods can produce drastically different results for temporally neighboring frames due to slight changes in their global or local content statistics, or it could also happen due to the multimodality of the application as there could exist a number of valid solutions for images with similar content as highlighted in (Bonneel et al., 2015; Zhang et al., 2019). Therefore, the extension of these image-to-image translation tasks to the video domain is an active area of research in computer vision, and these extensions generally require redesigning the whole algorithm with particular attention to the video dynamics as presented in (Chen et al., 2017; Lei & Chen, 2019; Liu & Freeman, 2010; Liu et al., 2018). Re-formulating these video processing methods is a challenging task as there are only a handful of available datasets for video translation tasks. While adequate for one task, these methodologies perform poorly in their applicability to other tasks. Therefore, a method that can help extend a plethora of image processing applications with little to no knowledge of the operations used to process videos is quite useful. In this work, we propose a novel task agnostic framework capable of correcting the temporal consistency of the jagged videos produced with image processing operations without requiring a raw/unprocessed version of the video at the test stage.

There are several task-dependent temporal consistency correction approaches available such as (Chen et al., 2017; Liu et al., 2021; Thimonier et al., 2021), but only a handful of approaches have been proposed to tackle the problem of blind temporal consistency correction due to the complex nature of this task. (Bonneel et al., 2015) defined the problem of temporal consistency correction with gradient-domain minimization of per-frame processed video with the unprocessed video to minimize the warping error between the frames. (Lai et al., 2018) extended the deterministic formulation of (Bonneel et al., 2015) with the help of recurrent Convolutional Neural Networks (CNN) and introduced a perceptual penalty in their formulation to restrict the deviation of perceptual content of the restored video from the frame-wise processed video. Deep Video Prior (DVP) (Lei et al., 2020) extended Deep Image Prior (DIP) (Ulyanov et al., 2018) to the temporal dimension and proposed to formulate enforcing temporal consistency by training a CNN to generate processed video from the unprocessed video without utilizing optical flow. All of these approaches rely on the availability of unprocessed videos. To tackle the problem of blind video temporal consistency correction, we first focus on the shortcomings of the naive image transformation extensions and then look at the currently available approaches to see if the shortcomings have been addressed appropriately.

All the previously proposed approaches for the task of temporal consistency correction define it with the help of unprocessed videos. This kind of definition helps the model in developing a sense of consistent motion dynamics. This implicit definition, with the help of raw videos, limits the applicability of the previously proposed approaches to only the videos for which their raw/unprocessed counterparts are available. In order to overcome this limitation, we propose to learn a motion representation capable of generating consistent motion dynamics solely from inconsistent videos. This is achieved by extending the conventional bi-frame motion estimation to a tri-frame strategy. This strategy to evaluate consistent motion representation from inconsistent videos eliminates the requirement of the unprocessed counterpart of the video, making it the first-ever truly task agnostic video temporal consistency correction method. For the task of learning this motion representation, we fine-tune conventional optical flow estimations networks like (Ilg et al., 2017; Sun et al., 2018) along with the generative part of the network. Our model pipeline is illustrated in figure 2. The proposed model achieves state-of-the-art qualitative and quantitative results. The proposed formulation also can deal with the resolution mismatch problem in processed and raw videos and makes it possible to extend the Single Image Super-Resolution (SISR) (Ledig et al., 2017) methods to Video Super-Resolution (VSR) methods without any modification. The detailed description of our formulation is presented in 3. We summarize our contributions as follows:

- A novel framework for task agnostic temporal consistency correction that overcomes the need for siphoning motion dynamics from the raw videos.
- Propose a novel tri-frame extension for motion representation that generates a robust representation for both temporally consistent and inconsistent videos to mitigate the flicker introduced by naive extension of image processing operations to videos.
- Identify and propose tailored solutions for various challenges of naive extensions of image processing applications to videos and produce state-of-the-art results for this task.

2 Related Work

The literature is divided into two main streams to generate visually appealing videos through image transformation models. The first stream tackles the task of mitigating temporal inconsistencies with the reformulation of the task at hand with temporal information. The second stream focuses on developing post-processing models that refine the frame-wise processed videos by penalizing the temporal deviation of the processed video from its unprocessed counterpart. The second stream is further divided into two sub-streams: task-specific and task-agnostic. The following subsections describe the details of both of the above-mentioned streams, respectively.

Reformulation stream (video-to-video translation): These approaches are generally termed video-to-video translation of image-to-image transform applications. Generally, these approaches either consider multiple frames as input and produce multiple output frames or generate a single frame from multiple input frames in the form of frame recurrence such as (Chu et al., 2018; Liu et al., 2021). There are also cases like video style transfer where content information is propagated to the next time timestep with the help of optical flow to initiate the optimization of the next frame. This frame-recurrent methodology has also been proven effective in applications like video super resolution (Sajjadi et al., 2018). Designing these approaches for each task and training them from scratch is a hectic task and data scarcity can make these approaches unfeasible. Therefore, such models often do not adapt well to different tasks. Therefore, methods that can restore the temporal consistency of multiple frame-wise processed videos are highly valuable and are being actively investigated.

Post-processing methods for temporal consistency correction: (Bonneel et al., 2015) initiated the investigation of task agnostic temporal consistency correction using a gradient-based optimization strategy in which temporally consistent (unprocessed) videos were used as restoration guides for correcting the temporal inconsistency of frame-wise processed videos. Their formulation motivated various task-dependent and task-agnostic, and temporal consistency correction approaches with slight variations from their original formulation. Despite the efficacy of their formulation, there only exist a handful of approaches that address task-agnostic temporal consistency correction due to the difficult nature of the task. (Yao et al., 2017) proposed a keyframe strategy that accounted for the motion of different objects in those keyframes to handle occlusion as well as temporal consistency. (Lai et al., 2018) proposed the first deep learning approach for this task by employing ConvLSTM (Shi et al., 2015) and a perceptual loss (Johnson et al., 2016). (Lei et al., 2020) proposed an extension of Deep Image Prior (Ulyanov et al., 2018) and demonstrated its capability to mitigate the temporal flicker of per-frame processed videos. On the other hand, various task-specific approaches have been proposed, such as (Chu et al., 2018; Liu et al., 2021; Thimonier et al., 2021). Defining these task-specific approaches is relatively straightforward. These approaches define a temporal extension around a backbone method and penalize the deviation of generated content with the help of optical flow.

All of the previously proposed approaches rely on the availability of unprocessed videos for the successful restoration of the temporal dynamics. In this work, we propose a novel formulation that alleviates this problem by learning a consistent motion representation from the temporally inconsistent videos. This learned motion representation is then processed with the content of unprocessed videos in a branched network that ensures the natural dynamics in the generated videos. Our proposed network also contains a recurrent module (Shi et al., 2015) that helps consistent propagation of content throughout the video sequence and produces state-of-the-art qualitative and quantitative results.

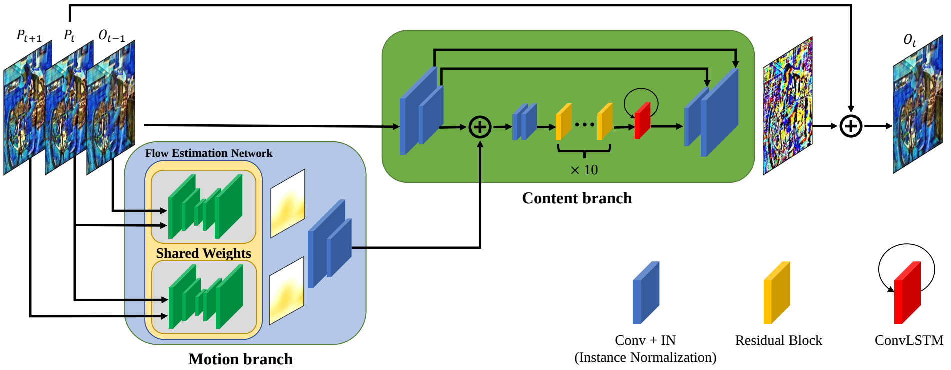


Figure 2: **Model Architecture.** An illustration of the proposed model architecture. The proposed model consists of a two-branch architecture that separately processes motion and content. The combined motion and content features are then passed through a succession of residual blocks and a decoder to generate the restored frame. The optical flow estimation networks used in the motion branch have shared weights.

3 Proposed Method

In this section, we describe the details of the proposed method. Consider a raw (temporally consistent) video $\{I_0, I_1, I_2, ..., I_n\}$ with n frames and its frame-wise processed temporally inconsistent video $\{P_0, P_1, P_2, ..., P_n\}$ acquired by an image processing function h as $\{h(I_0), h(I_1), h(I_2), ..., h(I_n)\}$. The goal of this task is to restore the temporal consistency of the temporally inconsistent video and to produce a temporally consistent version $\{O_0, O_1, O_2, ..., O_n\}$. These notations are chosen to be in line with the notations used in related literature (Bonneel et al., 2015; Lai et al., 2018; Lei et al., 2020).

The proposed formulation draws inspiration from the disentanglement of a video into its base components: content and motion representations to manipulate them separately to generate temporally consistent frames as presented in (Lin et al., 2017) for short length video generation. Therefore, we propose to learn and utilize a tangible motion representation that can be evaluated directly from a collection of frames. Generally, the motion in videos is defined over a pair of similar frames with the help of optical flow. In optical flow conventions, these individual frames are considered as content, and the estimated optical flow is considered as the inter-frame motion. By following this convention, we use I_t and $of_{t \Rightarrow t-1}$ for content ant motion components, respectively. The estimated optical flow can also be utilized to define the approximate content of future frames with the help of past frames as follows:

$$I_{t+1} \approx w\left(I_t, of_{t+1 \Rightarrow t}\right). \tag{1}$$

In Eq. 1, I_t and $of_{t+1 \Rightarrow t}$ denote content and motion components (optical flow) respectively, and w denotes the warping operator. The warping function can approximate the content of a frame with the help of estimated optical flow, provided that there exists a one-to-one correspondence for all pixels in frame t, i.e., no occlusion and de-occlusion happens in between the frames I_t and I_{t+1}. This equation also highlights the interdependence of both content and motion components of a video and implies that in a frame recurrent setting, the change in either of the components can lead to significantly different results in the subsequent frames.

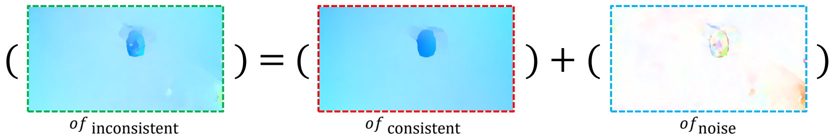


Figure 3: **Noise in optical flow.** An illustration of the decomposition of optical flow generated from temporally inconsistent frames to consistent optical flow and inconsistent noise.

A shortcoming of employing conventional optical flow is the deterministic nature of the optical flow and the accompanying warping operator. If the estimated optical flow cannot evaluate temporally consistent correspondences from the input frames, the evaluated optical flow becomes inconsistent. We observe that the evaluated optical flow from the per-frame processed video contains the consistent optical flow along with some additive noise (presented in figure 3) as follows:

$$of_{\text{inconsistent}} = of_{\text{consistent}} + of_{\text{noise}}. \tag{2}$$

Here, ”of” denotes the optical flow. This additive noise is introduced due to the temporal inconsistencies. Separating the consistent component from the estimated flow is challenging as optical flow is defined over only two consecutive frames. Thus, the content of both the participating frames is given equal importance, and the decision to separate consistent content from the inconsistent content with conventional optical flow becomes impractical. Another observation that can be made from figure 3 is that the flicker in frame-wise processed frames generally occurs near the motion boundaries. This impracticality of the conventional optical flow has hindered a truly task-agnostic solution to this task because the methods that deal with this task rely on the availability of the unprocessed counterparts of the per-frame processed videos to siphon temporally coherent motion dynamics for inter-frame flicker removal. To solve this problem, we propose deviating from the conventional bi-frame motion estimation methodologies by extending it to a tri-frame method to overcome the equal contribution of frames in motion estimation.

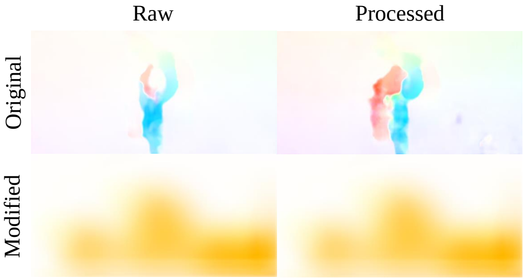


Figure 4: **Conventional and Modified Optical flow.** Comparison of original and Modified flow evaluated from temporally consistent (left) and inconsistent (right) frames. The modified flow evaluated from both the temporally consistent and inconsistent frames is quite similar.

This tri-frame motion estimation strategy also aids the network in developing a sense of rigidity to distinguish consistent content from a set of temporally inconsistent frames. To learn the consistent motion representation, an optical flow network has been integrated into the network as illustrated in figure 2. A similar integration of a small-scale optical flow estimation network (SpyNet (Ranjan & Black, 2017)) is presented in (Xue et al., 2019) for the task video denoising. We utilize a relatively larger network (PWC-Net (Sun et al., 2018)) in our method and provide an ablation study on different sized optical flow models in the accompanying supplementary text. The optical flow network in the proposed methodology is finetuned in an end-to-end manner without any special supervision. The objective functions used for the optimization are described in detail in the next section.

The proposed model takes in three consecutive frames P_{t-1}, P_t and P_{t+1} of the temporally inconsistent video as input. These three frames are propagated through both the content and motion branches. The proposed model consists of a UNet (Ronneberger et al., 2015) like structure with multiple encoder streams and a single decoder stream. The decoder contains skip connections from the content stream to encourage better reconstruction. The bottleneck part of the model contains a recurrent ConvLSTM (Shi et al., 2015) to transmit information of the generated frames to temporally distant frames. The motion branch comprises of two passes through a PWCNet (Sun et al., 2018) with shared weights followed by a conventional encoder-like architecture. The detailed architecture of the proposed model is presented in figure 2. This two-pass strategy in the motion branch is along with the attached encoder, and it extends the conventional bi-frame motion estimation strategy to a tri-frame motion estimation strategy.

Figure 4 highlights the learned motion representation generated by the finetuned PWC-Net (Sun et al., 2018). It is evident from figure 4 that the network learns to regress similar motion representations from both temporally consistent and inconsistent frames. This disentangled representation allows the users to further improve the level of restored temporal consistency with the help of an iterative arrangement; please refer to the supplemental for the results generated with an iterative arrangement.

4 Objective Functions

This section provides the details of the learning objectives used in the training phase of the proposed network. These objectives can be classified into two categories, short-term and long-term objectives. Both of these optimization categories are presented below.

4.1 Local Neighborhood Losses

In this work, we aim to find a general solution for correcting the temporal consistency of frame-wise processed videos. Due to the diversity of the applications addressed, a simplistic spatial content matching reconstruction loss cannot justify the generation of temporally consistent frames, as the content can vary greatly in applications like style transfer (Gatys et al., 2015) and image adjustment (Yan et al., 2016). This discrepancy in the image space of processed and raw videos hinders a straightforward definition of a feasible loss function for this task. Due to the challenging nature of this task, we define a flow gradient loss that provides a supervision signal for the reconstruction with the help of optical flow acquired through the raw videos. We further extend this loss to only compare the gradients of the optical flow acquired through the synthesized and raw frames. This comparison encapsulates the spatio-temporal information necessary to synthesize temporally consistent videos. This spatio-temporal loss is defined as follows:

$$\mathcal{L}_{fg} = \sum_{t=2}^{T} \left\| \nabla \left(of \left(O_t, O_{t-1} \right) \right), \nabla \left(of \left(I_t, I_{t-1} \right) \right) \right\|_1 . \quad (3)$$

Here, $\mathcal{L}_{fg}$ and of represents flow gradient loss and the optical flow estimation network (Ilg et al., 2017) respectively. The ∇ operator denotes the spatial gradient of the estimated optical flow. This optical flow based loss abstracts the objective of flow denoising as introduced in Eq. 2 by providing a supervision signal for the motion boundaries where most of the spatio-temporal noise is introduced in per-frame processing. A similar loss in optical flow domain has been proposed in (Li et al., 2022) where the raw optical flow of the generated video was compared with that of the ground truth video. Although effective for the task of video inpainting, the raw optical flow matching approach fails in applications like style transfer in which the entire content of the generated video and the raw video differ greatly, i.e., contains excessive spatio-temporal noise. In our experiments, the models trained with raw optical flow performed poorly, whereas the model trained with flow gradients converged in a relatively shorter time and reduced the warp error significantly.

This spatio-temporal loss by itself is not sufficient for faithfully correcting the temporal consistency as there could exist multiple solutions for optical flow equation, e.g., it does not take into account the perceptual information of the synthesized and raw frames. In order to address this issue and to guide the network for the generation of perceptually credible frames, a non-local optical flow based loss in the image space is employed as presented below:

$$\mathcal{L}_{reconstruction} = \sum_{t=2}^{T} M_{t \Rightarrow t-1} \left\| O_t - w(O_{t-1}, of(I_t, I_{t-1})) \right\|_1 , \quad (4)$$

Here T represents the total number of frames in a sequence. The working of this loss function can be summarized as the propagation of content from succeeding frames to the subsequent frames in image domain to ensure that the generated frames contain similar content to the previous frames and the inter-frame motion dynamics of the synthesized frames are similar to that of the temporally consistent counterpart. By unrolling Eq. 4 it can be seen that this loss provides supervision that is directly proportional to the optical flow of the consistent video hence further supporting the hypothesis presented in Eq. 2. This non-local optical flow based loss function takes into account the occlusion problem of the Eq. 1 by masking the occluded and de-occluded content as defined below:

$$M_{t \Rightarrow t-1} = \exp\left(-\alpha \left\| I_t - w(I_{t-1}, of(I_t, I_{t-1})) \right\|_2^2 \right) . \quad (5)$$

The value for $\alpha = 50$ is chosen according to the previous work where a similar strategy is used for evaluating occlusion (Ruder et al., 2016).

An additional short-term perceptual similarity loss (Johnson et al., 2016) was also introduced in the training phase to minimize the deviation of the synthesized frames from the original processed frames. This loss is defined as follows:

$$\mathcal{L}_p = \sum_{t=2}^{T} \sum_{l} \left\| \phi_l \left(O_t \right) - \phi_l \left(P_t \right) \right\|_1 , \quad (6)$$

whHere $\phi(.)$ represents $relu\_4\_3$ layer of a pretrained VGG-16 network, trained on the ImageNet dataset (Deng et al., 2009).

4.2 Temporal Constancy Loss

The loss defined in Eq. 4 adequately defines the content propagation from the previous frame to the current frame but lacks the ability to enforce it in temporally distant sequences. For instance, consider the task of colorization where a car appears in a portion of the frames and is assigned some random color, and it disappears and re-appears in a later instance. The frame-wise colorization method can assign it a different color in each sequence, and the loss function defined in Eq. 4 will be sufficient to enforce the color constancy of the car in each interval where the car is visible. This, however, does not ensure the constancy of the color assigned to the car in both intervals, i.e., the car could be assigned a solid blue color in the first interval and a solid green color in the next interval. To address this problem of temporally distant yet similar instances, a recurrent module is introduced in the proposed model. To ensure that the recurrent module works as intended, a long term extension of the short term loss is introduced in the training phase. This temporally distant extension of the short term loss is termed as constancy loss and is presented below:

$$\mathcal{L}_{constancy} = \sum_{p=1}^{T-2} \sum_{t=p+2}^{T} M_{t \Rightarrow p} \left\| O_t - w(O_p, of(I_{t-1}, I_p)) \right\|_1 . \tag{7}$$

Here subscript p highlights that the preceding distant images are compared with all the subsequent frames. The final loss for each training sequence is given by,

$$\mathcal{L}_{total} = \lambda_a \mathcal{L}_{fg} + \lambda_b \mathcal{L}_{reconstruction} + \lambda_c \mathcal{L}_p + \lambda_d \mathcal{L}_{constancy}. \tag{8}$$

The λ(s) in the equation above define the contribution of each loss in the optimization phase. The details of these hyperparameters are provided in the supplementary text.

5 Experiments and Results

During our experimentation phase, we tested various optical flow estimation networks such as (Ilg et al., 2017; Ranjan & Black, 2017; Sun et al., 2018). Flownet2.0 (Ilg et al., 2017) has approximately 162M trainable parameters, and it produced relatively better results in lesser training iterations as compared to the other two variants. Having a vast number of parameters compromises both the training and testing time. Therefore, we opted for a medium-sized optical flow estimation network (PWC-Net (Sun et al., 2018)) and a higher number of training iterations. The results produced by both variants are discussed in the supplementary text. We also experimented with an iterative arrangement of the trained models where restored videos are again subjected to temporal consistency correction models and evaluated that the proposed model architecture consistently reduces warp error with every iteration (as presented in figure 5). This reduced warp error comes at the cost of quality. Generally, the frames subjected to a high number of consistency correction iterations lose the fine details. The generated results by various restoration iterations are provided in the supplemental. Please note that the comparative results provided in this paper are generated through a single restoration iteration.

The dataset for training the proposed model contains videos from the DAVIS dataset for video segmentation (Perazzi et al., 2016) and videos gathered from Videovo.net by (Lai et al., 2018). The dataset contained videos from a diverse range of applications such as, Artistic Style Transfer (Gatys et al., 2015; Johnson et al., 2016), Colorization (Waechter et al., 2014; Zhang et al., 2016a), Image Enhancement (Gharbi et al., 2017), Intrinsic Image Decomposition (Bell et al., 2014), and Image-to-Image Translation tasks (Li et al., 2017; Zhang et al., 2016a; Zhu et al., 2017). The qualitative and quantitaive results are described below.

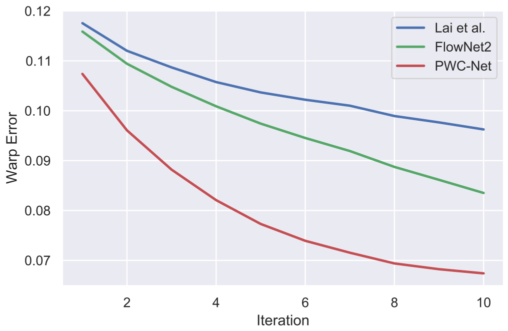


Figure 5: **Temporal warp error vs. iterations.** Temporal warp error consistently decreases with a higher number of restoration iterations.

5.1 Quantitative Results

The quantitative results are evaluated on the basis of temporal warp error. Tab. 1 shows the quantitative result produced by the methods proposed by (Bonneel et al., 2015), (Lai et al., 2018) and our method. Due to the tedious and computationally expensive nature of DVP (Lei et al., 2020) we compare some of their results in the accompanying supplemental for qualitative and quantitative comparison. The lower temporal warp error represents better temporal consistency. The lowest

Task	Trained	DAVIS			VIDEVO		
		Bonneel	Lai	Ours	Bonneel	Lai	Ours
WCT (Li et al., 2017)/antimono	✓	0.0029	0.0031	0.0026	0.0015	0.0021	0.0015
WCT (Li et al., 2017)/asheville		0.0047	0.0059	0.0047	0.0032	0.0043	0.0039
WCT (Li et al., 2017)/candy	✓	0.0034	0.0047	0.0035	0.0020	0.0032	0.0025
WCT (Li et al., 2017)/feathers		0.0040	0.0040	0.0027	0.0027	0.0030	0.0021
WCT (Li et al., 2017)/sketch	✓	0.0036	0.0029	0.0021	0.0025	0.0022	0.0017
WCT (Li et al., 2017)/wave		0.0033	0.0035	0.0027	0.0023	0.0026	0.0021
Fast-neural-style (Johnson et al., 2016)/princess		0.0043	0.0063	0.0042	0.0035	0.0053	0.0042
Fast-neural-style (Johnson et al., 2016)/udnie		0.0021	0.0023	0.0022	0.0012	0.0014	0.0014
DBL (Gharbi et al., 2017)/expertA	✓	0.0017	0.0011	0.0010	0.0014	0.0007	0.0006
DBL (Gharbi et al., 2017)/expertB		0.0016	0.0010	0.0008	0.0011	0.0006	0.0004
Intrinsic (Bell et al., 2014)/reflectance		0.0015	0.0008	0.0007	0.0011	0.0006	0.0006
Intrinsic (Bell et al., 2014)/shading	✓	0.0014	0.0008	0.0008	0.0008	0.0004	0.0004
CycleGAN (Zhu et al., 2017)/photo2ukiyoe		0.0024	0.0019	0.0015	0.0017	0.0013	0.0010
CycleGAN (Zhu et al., 2017)/photo2vangogh		0.0026	0.0026	0.0019	0.0020	0.0020	0.0015
Colorization (Zhang et al., 2016b)	✓	0.0016	0.0011	0.0008	0.0010	0.0006	0.0004
Colorization (Waechter et al., 2014)		0.0015	0.0009	0.0007	0.0010	0.0005	0.0003
Average		0.0027	0.0027	0.0021	0.0018	0.0019	0.0015

Table 1: **Quantitative comparison of Temporal Warping Error.**

temporal warp error is highlighted in red, and the second best is highlighted in blue. It is noteworthy that the temporal warp error does not take into account the perceptual quality of the produced frames and assigns a lower value for videos that resemble raw/unprocessed videos (as produced by (Bonneel et al., 2015)). It is evident from Tab. 1 that the proposed model produces state-of-the-art consistency results. Please note that unlike the proposed method, both the compared methods require the availability of raw/unprocessed videos for siphoning temporal dynamics.

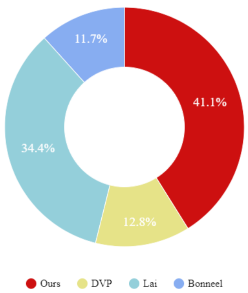


Figure 6: **User Study.** This donut chart highlights the user preferences collected through the user study. A majority of the participants preferred the results produced by the proposed method.

It is worth mentioning that some attempts to judge the perceptual quality through metrics like mean PSNR (performance degradation (Lei et al., 2020)) have been proposed for this task. We intentionally disregarded this metric and argue that this metric might be helpful to evaluate the performance of successive iterations of their prior-based methods where each iteration leads to progressively better results, but it can be a bit misleading in judging the overall perceptual quality of restored videos. In order to understand our arguments, let us assume a toy example of three consecutive frames, the first and the third frame are purely white, and the middle frame is black. All of the proposed methods for this task will treat the middle frame as an inconsistent frame and will try to correct it and produce a blended/blurred gray frame. If the models do not restore the middle frame, this metric will indicate that the models have faithfully restored the middle frame, whereas, in the case where the middle frame is corrected, the same metric will indicate that the produced frame is of lower quality, despite better temporal coherence. A similar phenomenon can be observed with other local perceptual metric such as LPIPS (Zhang et al., 2018). Therefore, to properly evaluate the models on their perceptual quality, thorough user studies are necessary. The findings of our conducted user studies are presented below and in the appended supplemental text.

5.1.1 User Study

We conducted 2 user studies; a comprehensive and a factorized user study to properly evaluate the performance of the proposed model. The first user study consisted of 36 participants with 130 different scenes processed through each of the proposed model. Each participant was asked to judge 5 videos processed through each of the proposed method for this task. The users were shown a randomly sampled scene from the video pool processed through (Bonneel et al., 2015), (Lai et al., 2018), (Lei et al., 2020) and this work, simultaneously. The participants were instructed to judge the videos on the basis of naturalness, quality, consistency, content and style preservation. Figure

6 presents the findings of the first user study. On average, 41% of the users preferred the videos restored by the proposed method. More details about the environment of the user study and the factorized study are presented in the appended supplemental.

5.2 Qualitative Results

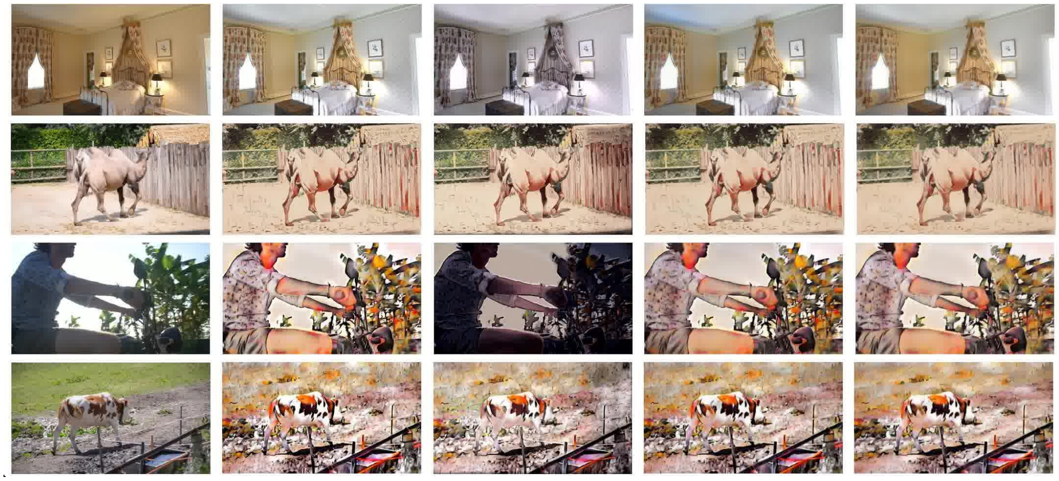


Figure 7: **Qualitative Results comparison.** From Left to Right: Raw Videos, Processed videos, videos generate by (Bonneel et al., 2015), (Lai et al., 2018) and the proposed method. The comparing videos in the above presented figure can be played with Adobe PDF Reader on a computer screen.

Fig. 7 presents some of the results produced by the methods proposed by (Bonneel et al., 2015), (Lai et al., 2018) and the proposed model. The results produced by (Bonneel et al., 2015) fail to retain perceptual quality of the processed videos. In the first video, the method proposed by (Lai et al., 2018) produces flicker and odd colors in the top part. In contrast, the proposed model produces temporally consistent frames while retaining the perceptual quality of the processed video. Due to the space limitation, further higher quality results, along with the results on super resolution (Ledig et al., 2017), inpainting (Lee et al., 2021), multiple iterations and user studies are presented in the supplemental.

6 Limitations and Future work

The proposed model has difficulties in restoring videos with very low frame rate and in the case of style transferred videos, some finer inconsistencies like finer brush strokes are lost, which makes the resulting video seem a little dull (this is to be expected from all the proposed approaches for this task). A good direction for future work in this field can be the exploration of consistent flow estimation models with deterministic warp operators. The developed consistent flow models can be used for a number of video processing applications like denoising and dehazing.

7 Conclusion

We present a task agnostic temporal consistency correction framework that restores natural video dynamics by learning consistent motion representations from the temporally inconsistent videos. We address various aspects of temporal consistency correction of per-frame processed videos in a task-agnostic manner. Unlike the previously proposed approaches for this task, the proposed method does not rely on siphoning video dynamics from the unprocessed videos, hence mitigates the requirement of unprocessed videos at inference time, and produces state-of-the-art qualitative and quantitative results. Through extensive experimentation and user studies, we demonstrate that the proposed method compares favorably to the pre-existing methods available for this task.

REFERENCES

Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014.

Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM Transactions on Graphics (TOG)*, 34(6):1–9, 2015.

Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1105–1114, 2017.

Mengyu Chu, You Xie, Laura Leal-Taixé, and Nils Thuerey. Temporally coherent gans for video super-resolution (tecogan). *arXiv preprint arXiv:1811.09393*, 1(2):3, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36 (4):1–12, 2017.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2462–2470, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 170–185, 2018.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Eunhye Lee, Jeongmu Kim, Jisu Kim, and Tae Hyun Kim. Restore from restored: Single-image inpainting. *arXiv preprint arXiv:2102.08078*, 2021.

Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3753–3761, 2019.

Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33:1083–1093, 2020.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. *arXiv preprint arXiv:2204.02663*, 2022.

Xunyu Lin, Victor Campos, Xavier Giro-i Nieto, Jordi Torres, and Cristian Canton Ferrer. Disentangling motion, foreground and background features in videos. *arXiv preprint arXiv:1707.04092*, 2017.

Ce Liu and William T Freeman. A high-quality video denoising algorithm based on reliable motion estimation. In *European conference on computer vision*, pp. 706–719. Springer, 2010.

Jiaying Liu, Wenhan Yang, Shuai Yang, and Zongming Guo. Erase or fill? deep joint recurrent rain removal and reconstruction in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3233–3242, 2018.

Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong. Temporally consistent video colorization with deep feature propagation and self-regularization learning. *arXiv preprint arXiv:2110.04562*, 2021.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4161–4170, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Manuel Ruder, Alexey Dosovitskiy, and Thomas Brox. Artistic style transfer for videos. In *German conference on pattern recognition*, pp. 26–36. Springer, 2016.

Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6626–6634, 2018.

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. *Advances in neural information processing systems*, 28, 2015.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.

Hugo Thimonier, Julien Despois, Robin Kips, and Matthieu Perrot. Learning long term style preserving blind video temporal consistency. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *European conference on computer vision*, pp. 836–850. Springer, 2014.

Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2):1–15, 2016.

Chun-Han Yao, Chia-Yang Chang, and Shao-Yi Chien. Occlusion-aware video temporal consistency. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 777–785, 2017.

Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8052–8061, 2019.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016a.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Learning Task Agnostic Temporal Consistency Correction
– Supplementary Material –

8 Overview

Due to the space limitation in the main manuscript, additional results, experiments, comparisons, ablation, and user studies are presented in this supplemental. Specifically, we present the details of the ablation study that highlights the choice of optical flow estimation network and the effects of loss functions presented in the main paper. Then, this supplemental provides quantitative and qualitative comparisons of the proposed method alongside the results generated with Deep Video Prior (DVP) (Lei et al., 2020). Lastly, we also include the results generated with iterative experiments, Single Image Super-Resolution (SISR) (Ledig et al., 2017) and image inpainting (Lee et al., 2021).

9 Ablation Study

In order to properly evaluate the contribution of each of the loss functions and modules presented in the main manuscript, a thorough ablation study is presented in this section. We first define the contribution of each loss function used and later present the variants of the proposed model trained with the effective loss functions but different optical flow estimation modules.

9.1 Effects of various Loss Functions

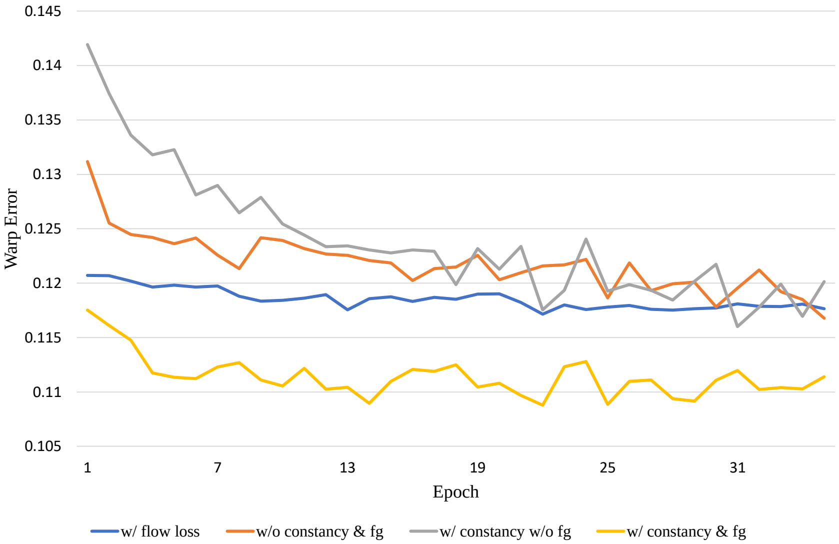


Figure 8: **Ablation of loss functions.** The graph containing temporal warp error against the number of epochs for each model trained with a different combination of losses. The warp error for the final model is highlighted with a solid yellow line.

As presented in the main paper, training this model with a simplistic reconstruction loss is infeasible; therefore, a combination of various loss functions is used to address the task at hand properly. figure 8 presents the graph containing temporal warp error against the number of epochs for each model trained with a different combination of losses. We started our experiments with a simplistic

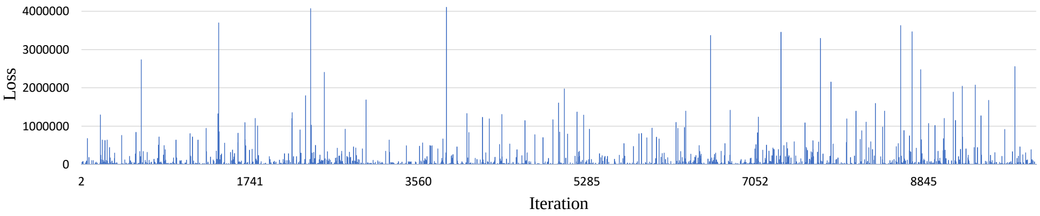


Figure 9: **The problems with raw optical flow loss.** The variety of domain transfer applications included in the dataset, optimizing the model based on raw optical flow becomes challenging, as shown in this figure.

loss function which provided the means to compare videos having mainly the same motion profiles but different content. This loss proved ineffective for the task, as the optical flow estimated from style transfer videos can vary significantly while optical flow evaluated from applications like image adjustment (Yan et al., 2016) can be very similar to the same evaluated from their unprocessed counterparts. The graph presented in figure 9 highlights this variation for simplistic optical flow loss for different batches containing various applications. This, however, can be effective for applications where the domain of the translated videos remains the same as presented by (Li et al., 2022). We also observed that for the variety of applications addressed, most of the inconsistencies occur near the motion boundaries; hence we extended this simplistic loss with the help of spatial gradients of optical flow and obtained visually sharper and quantitatively smoother videos as highlighted in figure 8. The model trained without the proposed flow gradient loss produced blurrier results near the motion boundaries as highlighted with a gray line in figure 8. We observed that the inclusion of the flow gradient loss helped overcome the blurriness near the motion boundaries. We also trained a model without the proposed constancy loss and found that the model learned to regress locally consistent frames and hence resulted in higher warp error. The introduction of the constancy loss mitigated the problem of global consistency and resulted in lower overall warp errors. It can be deduced from the figure that each of the losses presented in the main manuscript compliments each other and plays significant roles in the production of the final results. Please note that a perceptual loss (Johnson et al., 2016) was used with all the configurations presented in figure 8.

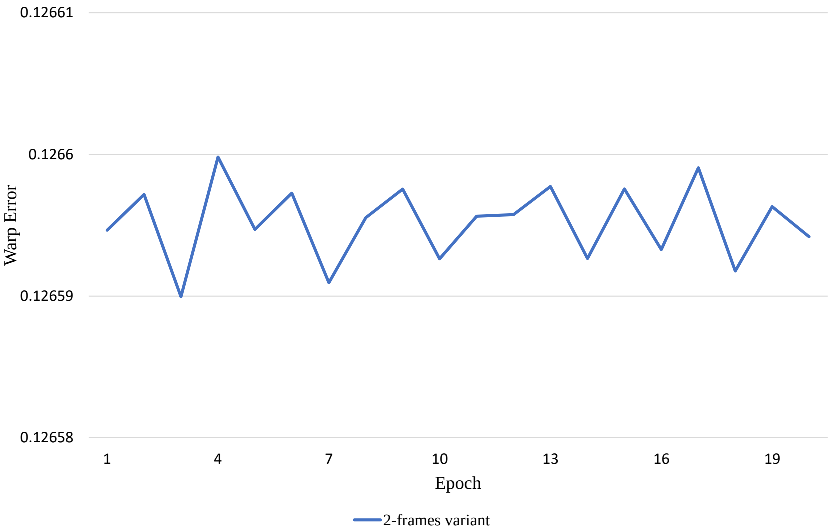


Figure 10: **Warp error for the two frame variant.** This figure presents the temporal warp error for the 2 frame variant of the proposed model.

Figure 10 presents the temporal warp error for the model that takes in two frames as input and tries to synthesize the restored version of the second frame. It is evident from figures 8 and 10 that the two frame model fails to learn the gist of temporal stability, whereas, the models trained with three frames learn the necessary reasoning to restore temporal coherence through added information that is essential to develop a sense of rigidity without the need for siphoning motion dynamics from the unprocessed videos.

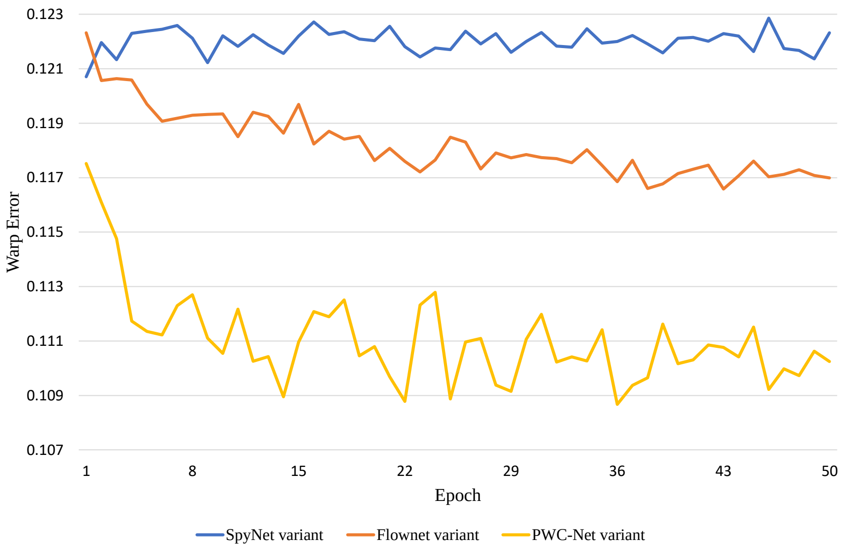


Figure 11: **Ablation of optical flow estimation models.** Temporal warp error vs. the number of epochs graph for models containing various optical flow estimation networks.

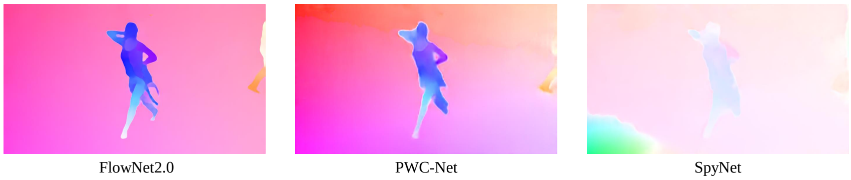


Figure 12: **Optical flow quality comparison.** This figure highlights the quality of optical flow estimated with various optical flow estimation networks. The optical flow estimated with FlowNet2.0 (Ilg et al., 2017) contains sharp boundaries with minimal noise. Whereas, the optical flow estimated through PWC-Net (Sun et al., 2018) and SpyNet (Ranjan and Black, 2017) contains a significant amount of noise.

9.2 Choice of Optical Flow Estimation Network

After finalizing the objective functions, we repeated the experiments with various optical flow estimation models such as FlowNet2.0 (Ilg et al., 2017) and SpyNet (Ranjan and Black, 2017), and finally settled for the model containing PWC-Net (Sun et al., 2018) due to its mediocre size. The temporal warp error vs. the number of epochs graph for each of these models is presented in figure 11. FlowNet2.0 consists of nearly 162M trainable parameters and requires huge computational resources for both training and inference. PWC-Net (Sun et al., 2018), on the other hand, consists of nearly 17 times lesser parameters than FlowNet2.0 (Ilg et al., 2017) and produces comparative results and is easier to train as well. SpyNet (Ranjan and Black, 2017) on the other hand, consists of less than 1M trainable parameters; hence the model saturates very quickly. The results produced by the model containing FlowNet2.0 (Ilg et al., 2017) are visually smoother as the optical flow estimated with FlowNet2.0 (Ilg et al., 2017) is quite clear and has well-defined boundaries, unlike the other compared models (as presented in figure 12). This ”noise-free” optical flow estimation helps the model converge faster as compared to the other variants but also makes it infeasible to train and test on real-world videos; therefore, PWC-Net (Sun et al., 2018) is used in the final version of the proposed model.

9.3 Hyper-Parameters

During the experimentation phase, it was observed that the hyper-parameters T (number of frames in each training sequence) and the patch size played a significant role in minimizing the temporal warp error. The models trained with a larger T and patch size achieved better warp error results as compared to the models trained with lower T and patch size. All the models discussed above were trained with a patch size of 256×256 and a sequence length of 15 frames. As for the the loss functions defined in the main paper, the weights for each loss term were selected in line with the study proposed in (Lai et al., 2018) for short term and long term losses. For the final version of our trained model, we chose $\lambda_a = 10$, $\lambda_b = 100$, $\lambda_c = 10$, $\lambda_d = 100$. The learning rate for all of the proposed experiments was set to 1×10^{-4} with a decay factor of 0.5 after every 20000 optimization iterations.

10 User Study

As mentioned in the main manuscript, the evaluation metric (Temporal Warp Error) does not take into account the various aspects of a video and can also suggests blurry videos as temporally consistent videos. Due to this shortcoming of the metric, user studies are necessary to properly evaluate any proposed method for this task. For both of the user studies, an application was developed which randomly selected a scene from the pool of videos from the DAVIS (Perazzi et al., 2016) and videvo.net dataset, and presented the results produced by the above-mentioned methods side-by-side in random order.

10.1 Factorized User Study

Apart from the the comprehensive user study presented in the main paper, we also conducted a secondary user study in-depth user study to know the reasoning of participants and their opinion on their selections.

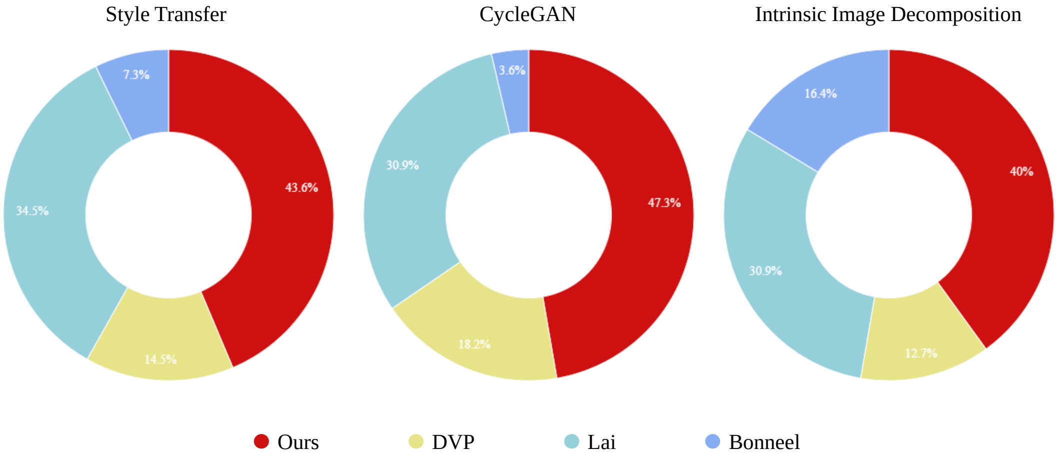


Figure 13: **Factorized user study.** The user preference scores for the tasks of style transfer, CycleGAN and intrinsic image decomposition (left to right respectively). Majority of the users preferred the videos processed through the proposed method in each category of tasks.

In this user study, we limited the video sampling to specific tasks like style transfer, CycleGAN and intrinsic image decomposition. Each user was instructed to judge 15 randomly selected videos (5 videos from each of the applications mentioned above) processed by each of the proposed models for this task. The participants were also instructed to record the reasoning for their selection. Most of the participants remarked that they chose the videos based on naturalness and lesser flickering. The results of the second user study for each of the selected tasks are provided in figure 13. On average, $\tilde{4}7\%$ of the users preferred the videos processed through the proposed model for these extremely challenging tasks. The user study application and the code will be open-sourced upon acceptance.

11 Run-time Comparison

Method	Time (s)
(Lai et al., 2018)	0.2146
DVP (Lei et al., 2020)	3.6365
Ours (FlowNet2.0 (Ilg et al., 2017))	0.2538
Ours (PWC-Net (Sun et al., 2018))	0.2236

Table 2: **Time comparison** Run-time comparison of methods proposed in (Lai et al., 2018), DVP (Lei et al., 2020) and the variants of the proposed model with different optical flow estimation networks in the motion branch.

Table 2 provides the average per-frame (910 × 480) generation time comparison of methods proposed by (Lai et al., 2018), (Lei et al., 2020) and the variants of the proposed model containing FlowNet2.0 (Ilg et al., 2017) and PWC-Net (Sun et al., 2018). As DVP (Lei et al., 2020) trains on each video, the run-time was calculated by processing a video scene containing 68 frames with all the above-mentioned methods. The time taken to process each video with DVP (Lei et al., 2020) increases with the number of frames but remains constant for both the methods proposed by (Lai et al., 2018) and this paper.

12 Comparison with DVP

Due to the tedious and resource-intensive nature of DVP (Lei et al., 2020), a large-scale comparative study was impractical, thus we provide some of the comparative results generated with DVP (Lei et al., 2020) in Fig. 14 – 15. Specifically, figure 14 presents the quantitative results (Temporal Warp Errors) and figure 15 presents the qualitative results. DVP (Lei et al., 2020) trains on each video and tries to generate a video that is similar to the per-frame processed video. This strategy, despite its efficacy, has certain drawbacks, such as temporal artifacts on recurring patches, blur, and texture loss. The videos generated with DVP (Lei et al., 2020) often resemble raw/unprocessed videos with a mere color change. This bias towards the raw videos helps it achieve a lower warp error score. Please note that the proposed method does not require the availability of raw videos in the inference time and still generates highly competitive results in terms of warp errors while retaining rich texture regions and better perceptual quality.

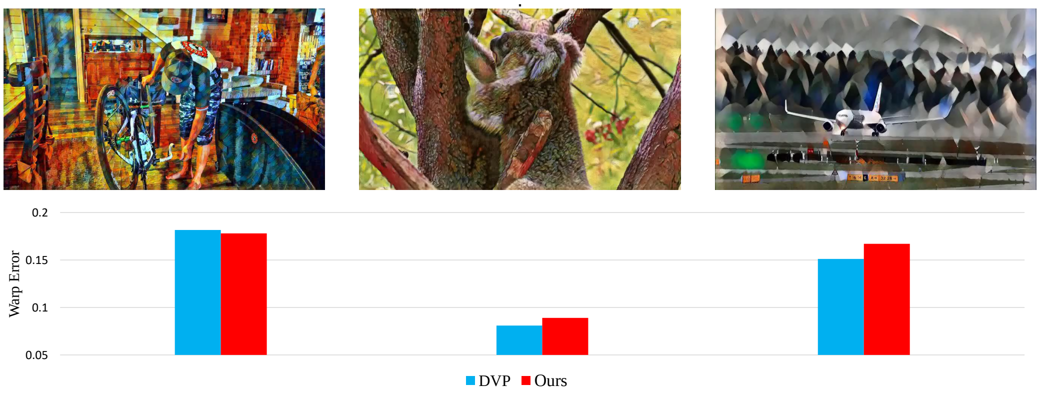
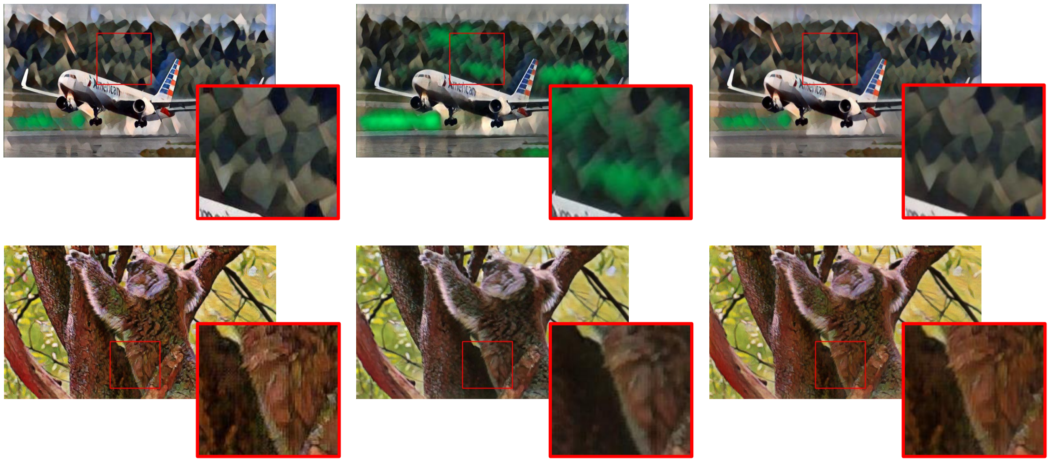


Figure 14: **Quantitative comparison with DVP (Lei et al., 2020)** The bar chart illustrates the temporal warp error on the videos generated with DVP (Lei et al., 2020) and the proposed method.

13 Implementation Details

The proposed model was implemented with PyTorch on a system with Two RTX 8000 GPU. A pretrained PWC-Net (Sun et al., 2018) was integrated into the network. The number of frames T



Processed Video

DVP

Ours

Figure 15: **Qualitative comparison with DVP (Lei et al., 2020)** Recurring artifacts, loss of texture, and blur produced by DVP (Lei et al., 2020). The highlighted frames are taken from per-frame processed videos, DVP (Lei et al., 2020) and the proposed method (left to right). The proposed method faithfully restores the temporal consistency of the processed video without compromising the texture and quality.

for each training sequence was set to 15, and the patch size of the image input was 256 × 256. We observed that the models trained with higher T and patch size performed better and converged relatively quicker than the models trained with smaller patch size and T. Generally, all the models converged in around 70k training iterations. Due to the space limitation, we provide a detailed ablation study regarding the choice of losses, optical flow estimation network, and various hyperparameters in the accompanied supplemental.

14 Experiments

In order to show the efficacy of the proposed model, we present some of the experiments in the following sections.

14.1 Iterative Restoration

Generally, ”blind” restorative computer vision applications lack the ability to control the amount or degree of restoration. In order to overcome this limitation (Ali et al., 2021), proposed an iterative strategy for controlling the amount of stability in video stabilization dataset. Inspired by their iterative strategy, we tested the proposed model in a similar arrangement. In our iterative arrangement, the videos produced are subjected to further restoration iterations. Through our experiments, we observed that the proposed model consistently improves the temporal consistency to an extent where the foreground starts to disintegrate in places where it occludes the background. This gives the users ability to restore the temporal consistency to their desired extent. Some of the iterative results are presented in figure 16.

14.2 Bonus Applications

figure 17 provides restored results generated with SISR (Ledig et al., 2017) and image inpainting (Lee et al., 2021). Due to the generative nature of these applications, the results produced contain severe inconsistencies, and there do not exist temporally consistent counterparts of these videos; hence restoration of these videos is impossible with the previously proposed methods. The proposed method in this paper significantly improves the temporal consistency of these videos.

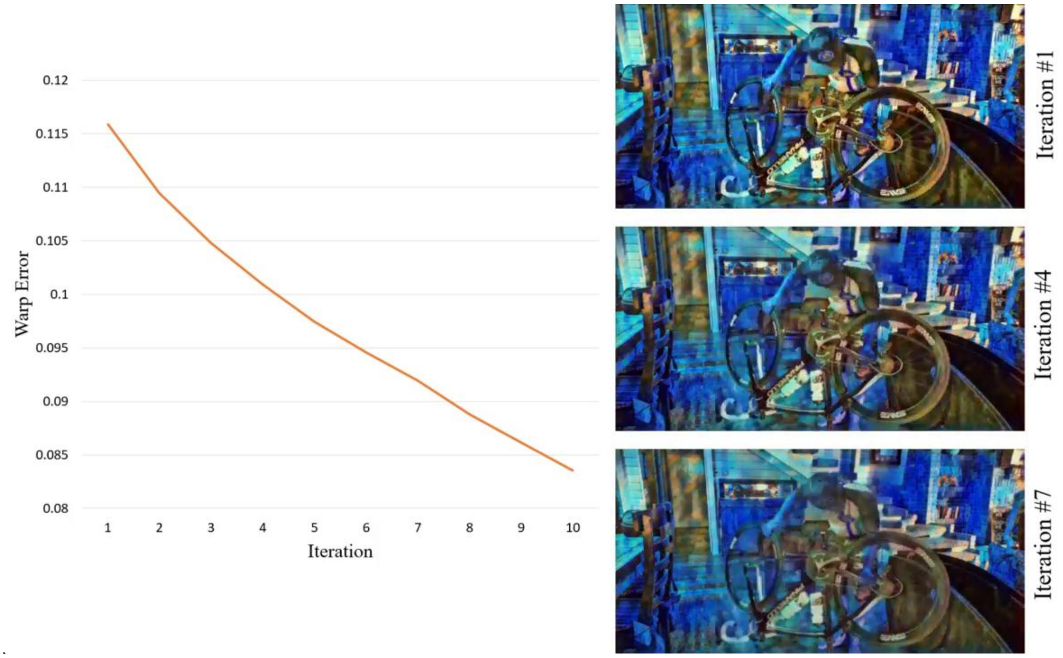


Figure 16: **Iterative results.** Warp error vs. the number of restoration iterations graph is presented on the left, and some iterative results are provided on the right side. This figure contains animation and is best viewed on a computer screen with Adobe PDF reader.

15 Additional Qualitative Results

Please refer to the accompanied supplementary videos for visualizing further results generated with the proposed model. Please note that in the accompanied supplementary video, the bottom left of each video contains the per-frame processed video, and the top right contains the videos restored by the proposed model.

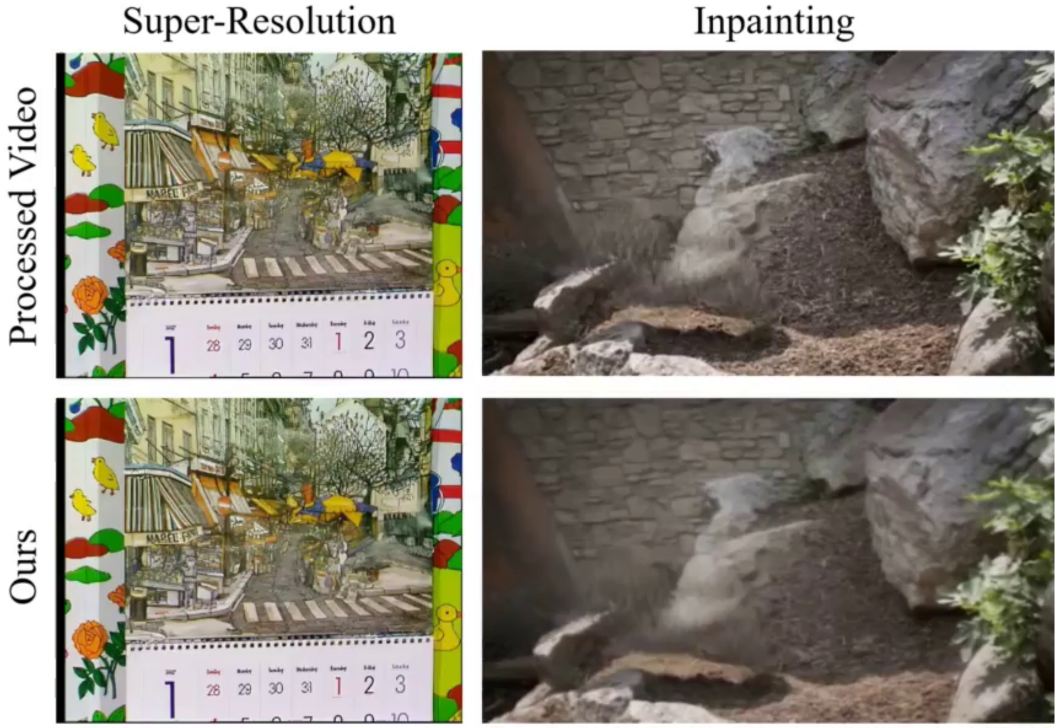


Figure 17: **SR and Inpainting results.** The right column presents results produced with an image super-resolution model, and the left side presents results generated with an image inpainting model. The restored video contains consistent letters and numerals for which the boundaries are severely affected by the super-resolution operation. The inpainted video contains high-frequency artifacts on the inpainted location; the proposed model mitigates these high-frequency temporal artifacts and produces smoother videos. This figure contains animation and is best viewed on a computer screen with Adobe PDF reader.

REFERENCES (SUPPLEMENTARY)

Muhammad Kashif Ali, Sangjoon Yu, and Tae Hyun Kim. Deep motion blind video stabilization. BMVC, 2021.

Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Blind video temporal consistency. *ACM Transactions on Graphics (TOG)*, 34(6):1–9, 2015.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

Eunhye Lee, Jeongmu Kim, Jisu Kim, and Tae Hyun Kim. Restore from restored: Single-image inpainting. *arXiv preprint arXiv:2102.08078*, 2021.

Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Advances in Neural Information Processing Systems*, 33:1083–1093, 2020.

Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. *arXiv preprint arXiv:2204.02663*, 2022.

Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2):1–15, 2016.