Deep Video Color Propagation

Simone Meyer[1,2]
simone.meyer@inf.ethz.ch
Victor Cornillère[2]
covictor@student.ethz.ch
Abdelaziz Djelouah[2]
aziz.djelouah@disneyresearch.com
Christopher Schroers[2]
christopher.schroers@disneyresearch.com
Markus Gross[1,2]
grossm@inf.ethz.ch

[1] Department of Computer Science
ETH Zurich
[2] Disney Research

Abstract

Traditional approaches for color propagation in videos rely on some form of matching between consecutive video frames. Using appearance descriptors, colors are then propagated both spatially and temporally. These methods, however, are computationally expensive and do not take advantage of semantic information of the scene. In this work we propose a deep learning framework for color propagation that combines a local strategy, to propagate colors frame-by-frame ensuring temporal stability, and a global strategy, using semantics for color propagation within a longer range. Our evaluation shows the superiority of our strategy over existing video and image color propagation methods as well as neural photo-realistic style transfer approaches.

1 Introduction

Color propagation is an important problem in video processing and has a wide range of applications. For example in movie making work-flow, where color modification for artistic purposes [2] plays an important role. It is also used in the restoration and colorization of heritage footage [1] for more engaging experiences. Finally, the ability to faithfully propagate colors in videos can have a direct impact on video compression.

Traditional approaches for color propagation rely on optical flow computation to propagate colors in videos either from scribbles or fully colored frames. Estimating these correspondence maps is computationally expensive and error prone. Inaccuracies in optical flow can lead to color artifacts which accumulate over time. Recently, deep learning methods have been proposed to take advantage of semantics for color propagation in images [5, 6] and videos [21]. Still, these approaches have some limitations and do not yet achieve satisfactory results on video content.

In this work we propose a framework for color propagation in videos that combines local and global strategies. Given the first frame of a sequence in color, the local strategy warps these colors frame by frame based on the motion. However this local warping becomes less

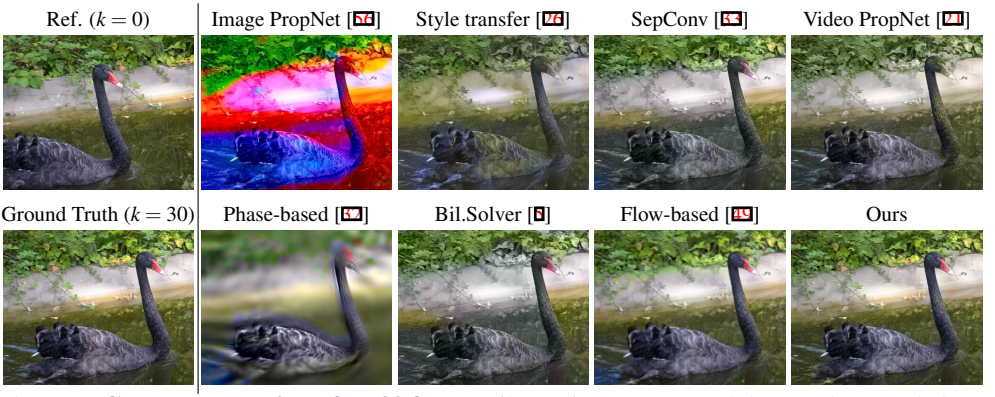

Figure 1: **Color propagation after 30 frames** ($k = 30$)**.** Our approach is superior to existing strategies for video color propagation. (Image source: [38])

reliable with increasing distance from the reference frame. To account for that we propose a global strategy to transfer colors of the first frame based on semantics, through deep feature matching. These approaches are combined through a fusion and refinement network to synthesize the final image. The network is trained on video sequences and our evaluation shows the superiority of the proposed method over image and video propagation methods as well as neural style transfer approaches, see Figure 1.

Our main contribution is a deep learning architecture, that combines local and global strategies for color propagation in videos. We use a two-stage training procedure necessary to fully take advantage of both strategies. Our approach achieves state-of-the-art results as it is able to maintain better colorization results over a longer time interval compared to a wide range of methods.

2 Related work

2.1 Image and Video Colorization

A traditional approach to image colorization is to propagate colors or transformation parameters from user scribbles to unknown regions. Seminal works in this direction considered low level affinities based on spatial and intensity distance [23]. To reduce user interaction, many directions have been considered such as designing better similarities [29]. Other approaches to improve edit propagation include embedding learning [9], iterative feature discrimination [50] or dictionary learning [10]. Achieving convincing results for automatic image colorization [11, 20], deep convolutional networks have also been considered for edit propagation [12] and interactive image colorization [56]. To extend edit propagation to videos, computational efficiency is critical and various strategies have been investigated [3, 53].

One of the first method considering gray scale video colorization was proposed by Welsh *et al.* [48] as a frame-to-frame color propagation. Later, image patch comparisons [43] were used to handle large displacements and rotations. However this method targets cartoon content and is not directly adaptable to natural videos. Yatzi *et al.* [54] consider geodesic distance in the 3d spatio-temporal volume to color pixels in videos and Sheng *et al.* [40] replace spatial distance by a distance based on Gabor features. The notion of reliability and priority [19] for coloring pixels allow better color propagation. These notions are extended to

entire frames [49], considering several of them as sources for coloring next gray images. For increased robustness, Pierre *et al.* [37] use a variational model that rely on temporal correspondence maps estimated through patch matching and optical flow estimation.

Instead of using pixel correspondences, some recent methods have proposed alternative approaches to the video colorization problem. Meyer *et al.* [32] transfer image edits as modifications of the phase-based representation of the pixels. The main advantage is that expensive global optimization is avoided, however propagation is limited to only a few frames. Paul *et al.* [35] uses instead of motion vectors the dominant orientations of a 3D steerable pyramid decomposition as guidance for the color propagation of user scribbles. Jampani *et al.* [21], on the other hand, use a temporal bilateral network for dense and video adaptive filtering, followed by a spatial network to refine features.

2.2 Style Transfer

Video colorization can be seen as transferring the *color* or *style* of the first frame to the rest of the images in the sequence. We only outline the main directions of *color* transfer as an extensive review of these methods is available in [13]. Many methods rely on histogram matching [39] which can achieve surprisingly good results given their relative simplicity but colors could be transferred between incoherent regions. Taking segmentation into account can help to improve this aspect [44]. Color transfer between videos is also possible [6] by segmenting the images using luminance and transferring chrominance. Recently Arbelot *et al.* [4] proposed an edge-aware texture descriptor to guide the colorization. Other works focus on more complex transformations such as changing the time of the day in photographs [41], artistic edits [42] or season change [34].

Since the seminal work of Gatys *et al.* [15], various methods based on neural networks have been proposed [24]. While most of them focus on painterly results, several recent works have targeted photo-realistic style transfer [18, 26, 28, 31]. Mechrez *et al.* [31] rely on Screened Poisson Equation to maintain the fidelity with the style image while constraining the results to have gradients similar to the content image. In [28] photo-realism is maintained by constraining the image transformation to be locally affine in color space. This is achieved by adding a corresponding loss to the original neural style transfer formulation [14]. To avoid the resulting slow optimization process, patch matching on VGG [18] features can be used to obtain a guidance image. Finally, Li *et al.* [26] proposed a two stage architecture where an initial stylized image, estimated through whitening and coloring transform (WCT) [25], is refined with a smoothing step.

3 Overview

The goal of our method is to colorize a gray scale image sequence by propagating the given color of the first frame to the following frames. Our proposed approach takes into account two complementary aspects: short range and long range color propagation, see Figure 2.

The objective of the short range propagation network is to propagate colors on a frame by frame basis. It takes as input two consecutive gray scale frames and estimates a warping function. This warping function is used to transfer the colors of the previous frame to the next one. Following recent trends [22, 33, 50], warping is expressed as a convolution process. In our case we choose to use spatially adaptive kernels that account for motion and re-sampling simultaneously [33], but other approaches based on optical flow could be considered as well.

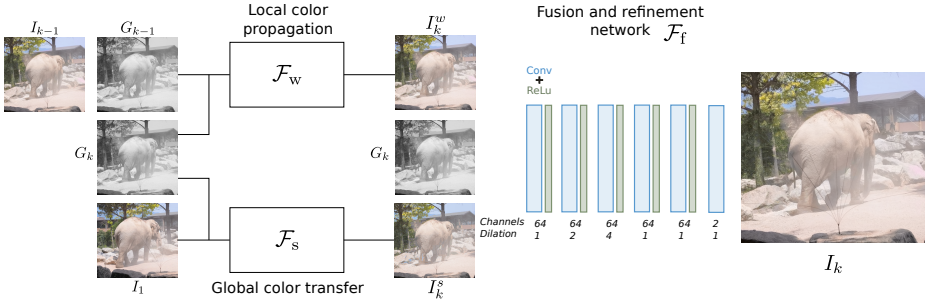

Figure 2: **Overview.** To propagate colors in a video we use both short range and long range color propagation. First, the local color propagation network $\mathcal{F}_w$ uses consecutive gray scale frames G_{k-1} and G_k to predict spatially adaptive kernels that account for motion and re-sampling from I_{k-1}. To globally transfer the colors from the reference frame I_1 to the entire video a matching based on deep image features is used. The results of these two steps, I_k^s and I_k^w, are together with G_k the input to the fusion and refinement network which estimates the final current color frame I_k. (Image source: [38])

For longer range propagation, simply smoothing warped colors according to the gray scale guide image is not sufficient. Semantic understanding of the scene is needed to transfer color from the first colored frame of the video to the rest of the video sequence. In our case, we find correspondences between pixels of the first frame and the rest of the video. Instead of matching pixel colors directly we incorporate semantical information by matching deep features extracted from the frames. These correspondences are then used in order to sample colors from the first frame. Besides the advantage for long range color propagation, this approach also helps to recover missing colors due to occlusion/dis-occlusion.

To combine the intermediate images of these two parallel stages, we use a convolutional neural network. This corresponds to the fusion and refinement stage. As a result, the final colored image is estimated by taking advantage of information that is present in both intermediate images, i.e. local and global color information.

4 Approach

Let's consider a grayscale video sequence $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$ of n frames, where the colored image I_1 (corresponding to G_1) is available. Our objective is to use the frame I_1 to colorize the set of grayscale frames $\mathcal{G}$. Using a local (frame-by-frame) strategy, colors of I_1 can be sequentially propagated to the entire video using temporal consistency. With a global strategy, colors present in the first frame I_1 can be simultaneously transfered to all the frames of the video using a style transfer like approach. In this work we propose a unified solution for video colorization combining local and global strategies.

4.1 Local Color Propagation

Relying on temporal consistency, our objective is to propagate colors frame by frame. Using the adaptive convolution approach developed for frame interpolation [33], one can similarly write color propagation as convolution operation on the color image: given two consecutive

grayscale frames G_{k-1} and G_k, and the color frame I_{k-1}, an estimate of the colored frame I_k can be expressed as

$$I_k^{\mathrm{w}}(x,y) = P_{k-1}(x,y) * K_k(x,y) \,, \tag{1}$$

where $P_{k-1}(x,y)$ is the image patch around pixel $I_{k-1}(x,y)$ and $K_k(x,y)$ is the estimated pixel dependent convolution kernel based on G_k and G_{k-1}. This kernel is approximated with two 1D-kernels as

$$K_k(x,y) = K_k^v(x,y) * K_k^h(x,y) \,. \tag{2}$$

The convolutional neural network architecture used to predict these kernels is similar to the one originally proposed for frame interpolation [33], with the difference that 2 kernels are predicted (instead of 4 in the interpolation case). Furthermore, we use a softmax layer for kernel prediction which helps to speedup training [46]. If we note $\mathcal{F}_w$ the prediction function, the local color propagation can be written as

$$I_k^w = \mathcal{F}_{\mathrm{w}}(G_k, G_{k-1}, I_{k-1}; \Lambda_{\mathrm{w}}) \,, \tag{3}$$

with Λ_{w} being the set of trainable parameters.

4.2 Global Color Transfer

The local propagation strategy becomes less reliable as the frame to colorize is further away from the first frame. This can be due to occlusions/dis-occlusions, new elements appearing in the scene or even complete change of background (due to camera panning for example). In this case, a global strategy with semantic understanding of the scene is necessary. It allows to transfer color within a longer range both temporally and spatially. To achieve this, we leverage deep feature extracted with convolutional neural networks trained for classification and image segmentation. Similar ideas have been developed for style transfer and image inpainting [24, 52].

Formally, we note $\Phi_{I,l}$ the feature map extracted from the image I at layer l of a discriminatively trained deep convolutional neural network. We can estimate a pixel-wise matching between the reference frame G_1 and the current frame to colorize G_k using their respective features maps $\Phi_{G_1,l}$ and $\Phi_{G_k,l}$. Similarity for two positions $\mathbf{x}, \mathbf{x}'$ is measured as:

$$\mathcal{S}_{G_k,G_1}(\mathbf{x},\mathbf{x}') = ||\Phi_{G_k,l}(\mathbf{x}) \; - \; \Phi_{G_1,l}(\mathbf{x}')||_2^2 \,. \tag{4}$$

Transferring the colors using pixel descriptor matching can be written as:

$$I_k^s(\mathbf{x}) = I_1(\arg\min_{\mathbf{x}'} \mathcal{S}_{G_k,G_1}(\mathbf{x},\mathbf{x}')) \,. \tag{5}$$

To maintain good quality for the matching, while being computationally efficient, we adopt a two stage coarse-to-fine matching. Matching is first estimated for features from a deep layer $l = l_{\mathrm{coarse}}$. This first matching, at lower resolution, defines a region of interest for each pixel in the second matching step of features at level $l = l_{\mathrm{fine}}$. The different levels l of the feature maps correspond to different abstraction level. The coarse level matching allows to consider regions that have similar semantics, whereas the fine matching step considers texture-like statistics that are more effective once a region of interest has been defined. We note $\mathcal{F}_s$ the global color transfer function

$$I_k^s = \mathcal{F}_{\mathrm{s}}(G_k, I_1, G_1; \Lambda_{\mathrm{s}}) \,, \tag{6}$$

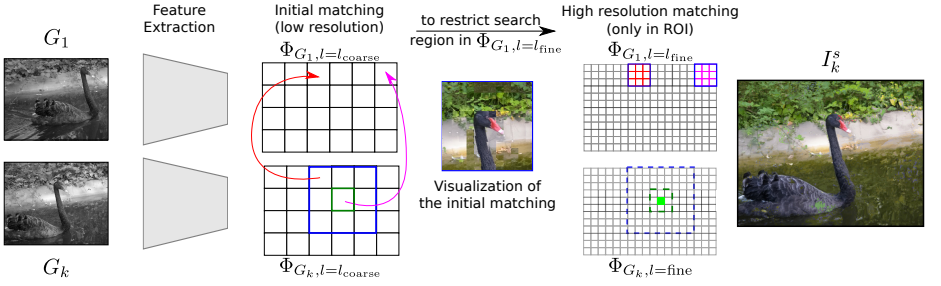

Figure 3: **Global Color Transfer.** To transfer the colors of the first frame I_1, feature maps ϕ_{G_1} and ϕ_{G_k} are extracted from both inputs G_1 and G_k. First, a matching is estimated at low resolution. This matching performed on features from a deep layer (l_{coarse}) allows to consider more abstract information. It is however too coarse to directly copy corresponding image patches. Instead, we use this initial matching to restrict the search region when matching pixels using low level image statistics (from level l_{fine} feature map). Here we show the region of interest (in blue) used to match the pixel in light green. All the pixels sharing the same coarse positions (in dark green rectangle) share the same Region Of Interest (ROI). Using the final matching, I_1 colors are transfered to the current gray scale image G_k. (Image source: [38])

with Λ_s being the set of trainable parameters. Figure 3 illustrates all the steps from feature extraction to color transfer. Any neural network trained for image segmentation could be used to compute the features maps. In our case we use ResNet-101 [17] architecture fine tuned for semantic image segmentation [8]. For l_{coarse} we use the output of the last layer of the *conv3*-block, while for l_{fine} we use the output of the first *conv1*-block (but with stride 1).

4.3 Fusion and Refinement Network

The results we obtain from the local and global stages are complementary. The local color propagation result is sharp with most of the fine details preserved. Colors are mostly well estimated except at occlusion/dis-occlusion boundaries where some color bleeding can be noticed. The result obtained from the global approach is very coarse but colors can be propagated to a much larger range both temporally and spatially. Fusing these two results is learned with a fully convolutional neural network.

For any given gray scale frame G_k, the local and global steps result in two estimates of the color image I_k: I_k^w and I_k^s. These intermediate results are leveraged by the proposed convolutional network (Figure 2) to predict the final output:

$$I_k = \mathcal{F}_{\text{f}}(G_k, I_k^w, I_k^s; \Lambda_{\text{f}}) \,, \tag{7}$$

where $\mathcal{F}_{\text{f}}$ notes the prediction function and Λ_{f} the set of trainable parameters.

Architecture details. The proposed fusion and refinement network consists of 5 convolutional layers with 64 output channels each followed by a relu-activation function. To keep the full resolution we use strides of 1 and increase the receptive field by using dilations of $1, 2, 4, 1$ and 1, respectively. To project the output to the final colors we us another convolutional layer without any activation function. To improve training and the prediction we use instance normalization [45] to jointly normalize the input frames. The computed statistics are then also used to renormalize the final output.

4.4 Training

Since all the layers we use are differentiable, the proposed framework is end-to-end trainable, and can be seen as predicting the colored frame $\hat{I}_k$ from all the available inputs

$$I_k = \mathcal{F}(G_k, G_{k-1}, I_{k-1}, I_1; \Lambda_\mathrm{w}, \Lambda_\mathrm{s}, \Lambda_\mathrm{f}). \tag{8}$$

The network is trained to minimize the total objective function $\mathcal{L}$ over the dataset $\mathcal{D}$ consisting of sequences of colored and gray scale images.

$$\Lambda_f^*, \Lambda_w^* = \arg\min_{\Lambda_f, \Lambda_w} \mathbb{E}_{I_1, I_2, G_1, G_2 \sim \mathcal{D}} \left[\mathcal{L}\right]. \tag{9}$$

Image loss. We use the ℓ_1-norm of pixel differences which has been shown to lead to sharper results than ℓ_2 [27, 30, 33]. This loss is computed on the final image estimate:

$$\mathcal{L}_1 = ||I_k - \hat{I}_k||_1 . \tag{10}$$

Warp loss. The local propagation part of the network has to predict the kernels used to warp the color image I_{i-1}. This is enforced through the warp loss. It is also computed as the ℓ_1-norm of pixel differences between the ground truth image I_i and I_i^w:

$$\mathcal{L}_w = ||I_k - I_k^w||_1 . \tag{11}$$

Since I_k^w is an intermediate result, using more sophisticated loss functions such as feature loss [14] or adversarial loss [16] is not necessary. All the sharp details will be recovered by the fusion network.

Training procedure. To train the network we used pairs of frames from video sequences obtained from the DAVIS [36, 38] dataset and Youtube. We randomly extract patches of 256×256 from a total of 30k frames. We trained the fusion net with a batch size of 16 over 12 epochs.

To efficiently train the fusion network we first apply $\mathcal{F}_\mathrm{w}$ and $\mathcal{F}_\mathrm{s}$ separately to all training video sequences. The resulting images I_k^w and I_k^s show the limitations of their respective generators $\mathcal{F}_\mathrm{w}$ and $\mathcal{F}_\mathrm{s}$. The fusion network can then be trained to synthesize the best color image from these two intermediate results. As input we provide G_k and the intermediate images I_k^w and I_k^s converted to Yuv-color space. Using the luminance channel helps the prediction process as it can be seen as an indicator on the accuracy of the intermediate results. The final image consists of the chrominance values estimated by the fusion network and G_k as luminance channel.

Running time. At test time, the matching step is the most computationally involved. Still, our naive implementation with TensorFlow computes high resolution (1280×720) edit propagation within $5s$ per frame on a Titan X (Pascal).

5 Results

For our evaluation we used various types of videos. This includes videos from DAVIS [36, 38], using the same test set as in [21], as well as from [7]. We also test our approach on HD videos from the video compression dataset [47].

Figure 4: **Ablation study.** Using local color propagation based on [33] only preserve details but is sensitive to occlusion/dis-occlusion. Using only global color transfer does not preserve details and is not temporally stable. Best result is obtained when combining both strategies. See Figure 9 for quantitative evaluation. (Image source: [7, 47])

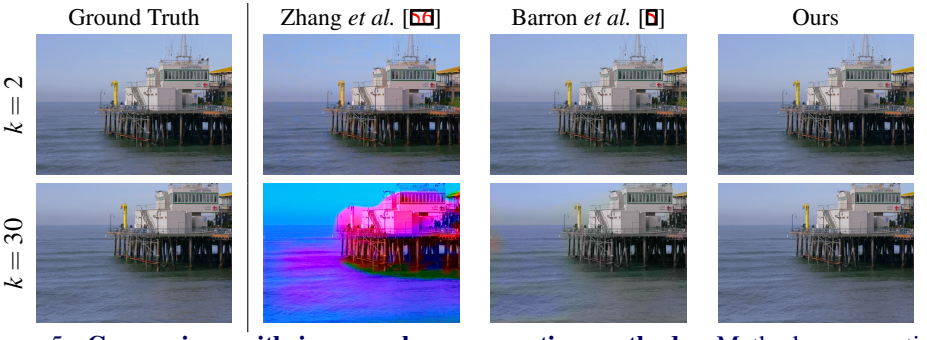

Figure 5: **Comparison with image color propagation methods.** Methods propagating colors in a single image achieve good results on the first frame. The quality of the results degrades as the frame to colorize is further away from the reference image. (Image source: [47])

Ablation Study. To show the importance of both the local and global strategy, we evaluate both configuration. The local strategy is more effective for temporal stability and details preservation but is sensitive to occlusion/dis-occlusion. Figure 4 shows an example where color propagation is not possible due to an occluding object, and a global strategy is necessary. Using a global strategy only is not sufficient, as some details are lost during the matching step and temporal stability is not maintained (see video in supplemental material).

Comparison with image color propagation. Given a partially colored image, propagating the colors to the entire image can be achieved using the bilateral space [5] or deep learning [56]. To extend these methods to video, we compute optical flow between consecutive frames [53] and use it to warp the current color image (details provided in supplementary material). These image based color methods achieve satisfactory color propagation on the first few frames (Figure 5) but the quality quickly degrades. In the case of the bilateral solver, there is no single set of parameters that performs satisfactorily on all the sequences. The deep learning approach [56] is not designed for videos and drifts towards extreme values.

Comparison with video color propagation. Relying on optical flow to propagate colors in a video is the most common approach. In addition to this, Xie *et al.* [49] also consider frame re-ordering and use multiple reference frames. However, this costly process is limiting as processing 30 HD frames requires several hours. Figure 1 and Figure 6 shows that

Figure 6: **Comparison with video color propagation methods.** Our approach best retains the sharpness and colors of this video sequence. Our result was obtained in less than one minute while the optical flow method [49] needed 5 hours for half the original resolution. (Image source: [47])

Figure 7: **Comparison with photo-realistic style transfer.** The reference frame is used as style image. (Image source: [38])

we achieve similar or better quality in one minute. Phase-based representation can also be used for edit propagation in videos [32]. This original approach to color propagation is however limited by the difficulty in propagating high frequencies. Recently, video propagation networks [21] were proposed to propagate information forward through a video. Color propagation is a natural application of such networks. Contrary to the fast bilateral solver [5] that only operates on the bilateral grid, video propagation networks [21] benefits from a spatial refinement module and achieve sharper and better results. Still, by relying on standard bilateral features (i.e. colors, position, time) colors can be mixed and propagated from incorrect regions, which leads to the global impression of washed out colors.

Comparison with photo-realistic style transfer. Propagating colors of a reference image is the problem solved by photo-realistic style transfer methods [26, 28]. These method replicate the global look but little emphasize is put on transferring the exact colors (see Figure 7).

Quantitative evaluation. Our test set consists of 69 videos which span a large range of scenarios with videos containing various amounts of motions, occlusions/dis-occlusion, change of background and object appearing/disappearing. Due to their prohibitive running time, some methods [28, 49] are not included in this quantitative evaluation. Figures 8 and 9 show the details of this evaluation. For a better understanding of the temporal behavior of the different methods, we plot error evolution over time averaged for all sequences. On the first frames, our results are almost indistinguishable from a local strategy (with very similar error values) but we quickly see the benefit of the global matching strategy. Our approach consistently outperforms related approaches for every frame and is able to propagate colors within a much larger time frame. Results of the video propagation networks [21] vary largely depending on the sequence, which explain the inconsistent numerical performance on our large test set compared to the selected images shown in this paper.

N	Gray	BSolver [8]	Style [26]	VideoProp [21]	SepConv [33] (local only)	Matching (global only)	Ours
10	33.65	41.00	32.94	34.96	42.72	38.90	**43.64**
20	33.66	39.57	32.81	34.65	41.01	37.97	**42.64**
30	33.66	38.59	32.70	34.45	39.90	37.43	**42.02**
40	33.67	37.86	32.61	34.26	39.08	37.02	**41.54**
50	33.68	37.40	32.54	34.13	38.56	36.75	**41.23**

Figure 8: **Quantitative evaluation:** Using PSNR in *Lab*-space we compute the average error over the first **N** frames.

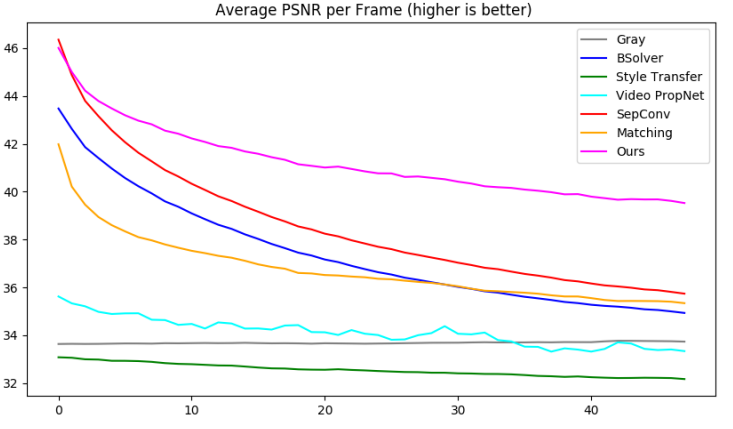

Figure 9: **Temporal evaluation:** The average PSNR error per frame shows the temporal stability of our method and its ability to maintain a higher quality over a longer period.

6 Conclusions

In this work we have presented a new approach for color propagation in videos. Thanks to the combination of a local strategy, that consists of a frame by frame image warping, and a global strategy, based on feature matching and color transfer, we have augmented the temporal extent to which colors can be propagated. Our extended comparative results show that the proposed approach outperforms recent methods in image and video color propagation as well as style transfer.

Acknowledgments. This work was supported by ETH Research Grant ETH-12 17-1.

References

[1] America In Color. https://www.smithsonianchannel.com/shows/america-in-color/1004516. Accessed: 2018-03-12.

[2] Short Documentary - Painting with Pixels: O Brother, Where Art Thou? .

[3] Xiaobo An and Fabio Pellacini. Appprop: all-pairs appearance-space edit propagation. *ACM Transactions on Graphics (TOG)*, 27(3):40:1–40:9, 2008.

[4] Benoit Arbelot, Romain Vergne, Thomas Hurtut, and Joëlle Thollot. Automatic texture guided color transfer and colorization. In *Joint Symposium on Computational Aesthet-*

ics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering, pages 21–32. Eurographics Association, 2016.

[5] Jonathan T. Barron and Ben Poole. The fast bilateral solver. In *European Conference on Computer Vision*, pages 617–632, 2016.

[6] Nicolas Bonneel, Kalyan Sunkavalli, Sylvain Paris, and Hanspeter Pfister. Example-based video color grading. *ACM Transactions on Graphics (TOG)*, 32(4):39:1–39:12, 2013.

[7] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, pages 611–625, 2012.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[9] Xiaowu Chen, Dongqing Zou, Qinping Zhao, and Ping Tan. Manifold preserving edit propagation. *ACM Transactions on Graphics (TOG)*, 31(6):132:1–132:7, 2012.

[10] Xiaowu Chen, Dongqing Zou, Jianwei Li, Xiaochun Cao, Qinping Zhao, and Hao Zhang. Sparse dictionary learning for edit propagation of high-resolution images. In *Conference on Computer Vision and Pattern Recognition*, pages 2854–2861, 2014.

[11] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *International Conference on Computer Vision*, pages 415–423, 2015.

[12] Yuki Endo, Satoshi Iizuka, Yoshihiro Kanamori, and Jun Mitani. Deepprop: Extracting deep features from a single image for edit propagation. *Computer Graphics Forum*, 35 (2):189–201, 2016.

[13] Hasan Sheikh Faridul, Tania Pouli, Christel Chamaret, Jürgen Stauder, Erik Reinhard, Dmitry Kuzovkin, and Alain Trémeau. Colour mapping: A review of recent methods, extensions and applications. *Computer Graphics Forum*, 35(1):59–88, 2016.

[14] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270, 2015.

[15] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] M. He, J. Liao, L. Yuan, and P. V. Sander. Neural color transfer between images. *arXiv preprint arXiv:1710.00756*, 2017.

[19] Junhee Heu, Dae-Young Hyun, Chang-Su Kim, and Sang-Uk Lee. Image and video colorization based on prioritized source propagation. In *International Conference on Image Processing*, pages 465–468, 2009.

[20] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4):110:1–110:11, 2016.

[21] Varun Jampani, Raghudeep Gadde, and Peter V. Gehler. Video propagation networks. In *Conference on Computer Vision and Pattern Recognition*, pages 3154–3164, 2017.

[22] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, pages 667–675, 2016.

[23] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. *ACM Transactions on Graphics (TOG).*, 23(3):689–694, 2004.

[24] Chuan Li and Michael Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Conference on Computer Vision and Pattern Recognition*, pages 2479–2486, 2016.

[25] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, pages 385–395, 2017.

[26] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. *arXiv preprint arXiv:1802.06474*, 2018.

[27] Gucan Long, Laurent Kneip, Jose M. Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu. Learning image matching by simply watching video. In *European Conference on Computer Vision*, pages 434–450, 2016.

[28] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Conference on Computer Vision and Pattern Recognition*, pages 6997–7005, 2017.

[29] Qing Luan, Fang Wen, Daniel Cohen-Or, Lin Liang, Ying-Qing Xu, and Heung-Yeung Shum. Natural image colorization. In *Eurographics Symposium on Rendering Techniques*, pages 309–320, 2007.

[30] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[31] Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. Photorealistic style transfer with screened poisson equation. In *British Machine Vision Conference*, 2017.

[32] Simone Meyer, Alexander Sorkine-Hornung, and Markus H. Gross. Phase-based modification transfer for video. In *European Conference on Computer Vision*, pages 633–648, 2016.

[33] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *International Conference on Computer Vision*, pages 261–270, 2017.

[34] Fumio Okura, Kenneth Vanhoey, Adrien Bousseau, Alexei A. Efros, and George Drettakis. Unifying color and texture transfer for predictive appearance manipulation. *Computer Graphics Forum*, 34(4):53–63, 2015.

[35] Somdyuti Paul, Saumik Bhattacharya, and Sumana Gupta. Spatiotemporal colorization of video using 3d steerable pyramids. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(8):1605–1619, 2017.

[36] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc J. Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[37] Fabien Pierre, Jean-François Aujol, Aurélie Bugeau, and Vinh-Thong Ta. Interactive video colorization within a variational framework. *SIAM Journal on Imaging Sciences*, 10(4):2293–2325, 2017.

[38] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[39] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[40] Bin Sheng, Hanqiu Sun, Shunbin Chen, Xuehui Liu, and Enhua Wu. Colorization using the rotation-invariant feature space. *IEEE Computer Graphics and Applications*, 31(2): 24–35, 2011.

[41] Yi-Chang Shih, Sylvain Paris, Frédo Durand, and William T. Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Transactions on Graphics (TOG)*, 32(6):200:1–200:11, 2013.

[42] Yi-Chang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. Style transfer for headshot portraits. *ACM Transactions on Graphics (TOG)*, 33 (4):148:1–148:14, 2014.

[43] Daniel Sýkora, Jan Buriánek, and Jirí Zára. Unsupervised colorization of black-and-white cartoons. In *International Symposium on Non-Photorealistic Animation and Rendering*, pages 121–127, 2004.

[44] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. Local color transfer via probabilistic segmentation by expectation-maximization. In *Conference on Computer Vision and Pattern Recognition*, pages 747–754, 2005.

[45] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[46] Thijs Vogels, Fabrice Rousselle, Brian McWilliams, Gerhard Röthlin, Alex Harvill, David Adler, Mark Meyer, and Jan Novák. Denoising with kernel prediction and asymmetric loss functions. *ACM Transactions on Graphics (TOG)*, 37(4), 2018.

[47] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeong-Hoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, Yun Zhang, Jiwu Huang, Sam Kwong, and C.-C. Jay Kuo. Videoset: A large-scale compressed video quality dataset based on JND measurement. *Journal of Visual Communication and Image Representation*, 46:292–302, 2017.

[48] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. Transferring color to greyscale images. *ACM Transactions on Graphics (TOG)*, 21(3):277–280, 2002.

[49] Sifeng Xia, Jiaying Liu, Yuming Fang, Wenhan Yang, and Zongming Guo. Robust and automatic video colorization via multiframe reordering refinement. In *International Conference on Image Processing*, pages 4017–4021, 2016.

[50] Li Xu, Qiong Yan, and Jiaya Jia. A sparse control model for image and video editing. *ACM Transactions on Graphics (TOG)*, 32(6):197:1–197:10, 2013.

[51] Tianfan Xue, Jiajun Wu, Katherine L. Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2016.

[52] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Conference on Computer Vision and Pattern Recognition*, pages 4076–4084, 2017.

[53] Tatsuya Yatagawa and Yasushi Yamaguchi. Temporally coherent video editing using an edit propagation matrix. *Computers & Graphics*, 43:1–10, 2014.

[54] Liron Yatziv and Guillermo Sapiro. Fast image and video colorization using chrominance blending. *IEEE Transactions on Image Processing*, 15(5):1120–1129, 2006.

[55] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-L^1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223, 2007.

[56] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 36(4):119:1–119:11, 2017.