DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement

SATOSHI IIZUKA, University of Tsukuba
EDGAR SIMO-SERRA, Waseda University / JST PRESTO

Fig. 1. **Vintage film remastering results.** Our approach is able to remaster 700 frames of video using only 6 reference color images in a single processing step. The first row shows various frames from the input video, the second row shows the restored black and white frames, the third row shows the variation between the input and restored black and white frames, and the fourth row shows the final colorized output. We show the reference color images used on the right. Using source-reference attention, our model automatically matches similar regions to the reference color images, and using self-attention with temporal convolutions it is able to enforce temporal consistency. Our approach is able to restore the noisy and blurring input, and, afterwards, with the few manually colored reference images, we are able to obtain a temporally-consistent natural looking color video. Images are taken from "A-Bomb Blast Effects" (1952) and licensed under the public domain. Figure best viewed in color.

The remastering of vintage film comprises of a diversity of sub-tasks including super-resolution, noise removal, and contrast enhancement which aim to restore the deteriorated film medium to its original state. Additionally, due to the technical limitations of the time, most vintage film is either recorded in black and white, or has low quality colors, for which colorization becomes necessary. In this work, we propose a single framework to tackle the entire remastering task semi-interactively. Our work is based on temporal convolutional neural networks with attention mechanisms trained on videos with data-driven deterioration simulation. Our proposed source-reference attention allows the model to handle an arbitrary number of reference color images to colorize long videos without the need for segmentation while maintaining temporal consistency. Quantitative analysis shows that our framework outperforms existing approaches, and that, in contrast to existing approaches, the performance of our framework increases with longer videos and more reference color images.

CCS Concepts: • **Computing methodologies** → **Image processing**; *Neural networks*;

Additional Key Words and Phrases: remastering, restoration, colorization, convolutional network, source-reference attention

ACM Reference format:
Satoshi Iizuka and Edgar Simo-Serra. 2019. DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement. *ACM Trans. Graph.* 38, 6, Article 176 (November 2019), 13 pages.
DOI: 10.1145/3355089.3356570

1 INTRODUCTION

Since the invention of motion pictures in the late 19th century, an incredible amount of hours of film have been recorded and released.

However, in addition to visual artefacts and the low quality of the film technology at the time, many of the earlier works of significant historical value have suffered from degradation or been lost. Restoration of such important films, given their analogue nature, is complicated, with the initial efforts beginning on restoring the film at a physical level. Afterwards, the content is transferred to the digital medium, where it is remastered by removing noise and artefacts in addition to adding color to the film frames. However, such remastering processes require a significant amount of both time and money, and is currently done manually by experts with a single film costing in the order of hundreds of thousands to millions dollars. Under these circumstances, huge industries such as publishers, TV, and the print industry, which own an enormous quantity of archived deteriorated old videos, show a great demand for efficient remastering techniques. In this work, we propose a semi-automatic approach for remastering old black and white films that have been converted to digital data.

Remastering an old film is not as simple as using a noise removal algorithm followed by colorization approach in a pipeline fashion: the noise and colorization processes are intertwined and affect each other. Furthermore, most old films suffer from blurring and low resolution, for which increasing the sharpness also becomes important. We propose a full pipeline for remastering black and white motion pictures, made of several trainable components which we train in a single end-to-end framework. By using a careful data creation and augmentation scheme, we are able to train the model to remaster videos by not only removing noise and adding color, but also increasing the resolution and sharpness, and improving the contrast with temporal consistency.

Our approach is based on fully convolutional networks. In contrast to many recent works that use recursive models for processing videos [Liu et al. 2018; Vondrick et al. 2018], we use temporal convolutions that allow for processing video frames by taking account information from multiple frames of the input video at once. In addition, we propose using an attention mechanism, which we denote as *source-reference* attention, that allows using multiple reference frames in an interactive manner. In particular, we use this *source-reference* attention to provide the model with an arbitrary number of color mages to be used as references when adding color. The model is able to not only dynamically choose what reference frames to use when coloring each output frame, but also choose what regions of the reference frames to use for each output region in a computationally efficient manner. We show how this approach can be used to remaster long sequences composing of multiple different scenes (close-up, panorama, etc.), using an assortment of reference frames as shown in Fig. 1. The number of reference frames used is not fixed and it is even possible to remaster in a fully automatic way by not providing reference frames. Additionally, by manually creating and/or colorizing reference frames, it is possible for the user to control the colorization results when remastering, which is necessary for practical applications.

We perform an in-depth evaluation of our approach both quantitatively and qualitatively, and find the results of our framework to be favorable in comparison with existing approaches. Furthermore, the performance of our approach increases on longer sequences with more reference color images, which proves to be a challenge for existing approaches. Our experiments show that using *source-reference* attention it is possible to remaster thousands of frames with a small set of reference images in a efficiently with stable and consistent colors.

To summarize, our contributions are as follows: (1) the first single framework for remastering vintage film, (2) source-reference attention that can handle an arbitrary number of reference images, (3) an example-based film degradation simulation approach for generating training data for film restoration, and (4) an in-depth evaluation with favorable results with respect to existing approaches and strong baselines. Models, code, and additional results are made available at http://iizuka.cs.tsukuba.ac.jp/projects/remastering/.

2 RELATED WORK

2.1 Denoising and Restoration

One of the more classical approaches to denoising and restoration is the family of Block-Matching and 3D filtering (BM3D) algorithms [Dabov et al. 2007; Maggioni et al. 2012, 2014], which are based on collaborative filtering in the in the transform domain. Although fairly limited in the types of noise patterns they can eliminate, these approaches have wide applicability to both images and video. Besides noise removal, other restoration related applications such as image super-resolution and deblurring [Danielyan et al. 2012] have also been explored with the BM3D algorithm.

More recently, Convolutional Neural Networks have been used for denoising-type applications, and, in particular, for single images [Lefkimmiatis 2018; Zhang et al. 2018b]. However, these generally assume simple additive Gaussian noise [Lefkimmiatis 2018], blurring [Fan et al. 2018; Shi et al. 2016; Yu et al. 2018], or JPEG-deblocking [Zhang et al. 2017b], or are applied to specialized tasks such as Monte Carlo rendering denoising [Bako et al. 2017; Chaitanya et al. 2017; Vogels et al. 2018] for which it is easy to create supervised training data. Extensions for video based on optical flow and transformer networks have also been proposed [Kim et al. 2018]. However, restoration of old film requires more than being able to remove Gaussian noise or blur: it requires being able to remove film artefacts that can be both local, affecting a small region of the image, or global, affecting the contrast and brightness of the entire frame, as shown in Fig. 2. For this it is necessary to create higher quality and realistic film noise as we propose in our approach.

2.2 Colorization

Colorization of black and white images is an ill-posed problem in which there is no single solution. Most approaches have relied on user inputs, either in the form of scribbles [Huang et al. 2005; Levin et al. 2004], reference images similar to the image being colorized [Irony et al. 2005; Pitié et al. 2007; Reinhard et al. 2001; Tai et al. 2005; Welsh et al. 2002; Wu et al. 2013], or internet queries [Chia et al. 2011; Liu et al. 2008]. While most traditional approaches have focused on solving an optimization problem using both the input greyscale image and the user provided hints or references images [An and Pellacini 2008; Levin et al. 2004; Xu et al. 2013], recent approaches have opted to leverage large datasets and employ learning-based models such as Convolutional Neural Networks (CNN) to colorize images automatically [Iizuka et al. 2016; Larsson

(a) Synthetic noisy images.

(b) Vintage black and white movies from the early 20th century.

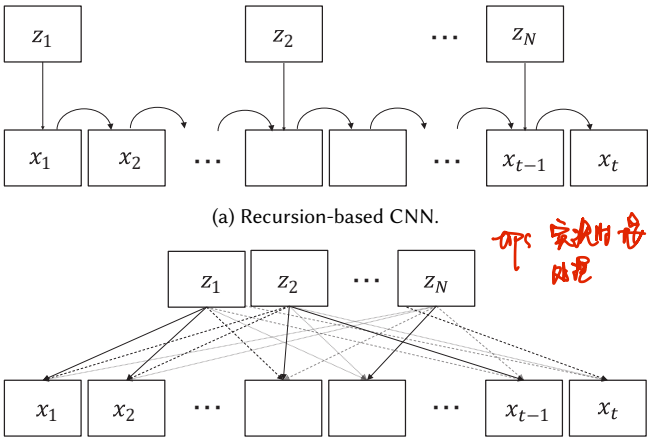
Fig. 2. **Comparison between denoising and restoration tasks.** (a) Example of generated synthetic images for denoising tasks [Martin et al. 2001]. The top row shows the original images and the bottom row shows them with added Gaussian noise. (b) Example of vintage film which requires restoration. The old movies suffer from a plethora of deterioration issues such as film grain noise, scratches, dampness, vignetting, and contrast bleed, which make them challenging to restore to their original quality. (a) Images are taken from [Martin et al. 2001], and (b) videos licensed in the Public Domain.

et al. 2016; Zhang et al. 2016]. Analogous to the optimization-based approaches, CNN-based approaches have been extended to handle user inputs as both scribbles [Sangkloy et al. 2017; Zhang et al. 2017a], and a single reference images [He et al. 2018; Meyer et al. 2018]. Our approach, while related to existing CNN-based methods, extends the colorization to video and an arbitrary number of reference images, in addition to performing restoration of the video.

Related to the current work are Recursive Neural Network (RNN) approaches for colorizing videos [Liu et al. 2018; Vondrick et al. 2018]. They process the video frame-by-frame by propagating the color from an initial colored key frame to rest of the scene. While this is a simple way to colorize videos, it can fail to propagate the color when there are abrupt changes in scene. In particular, RNN-based methods have the following limitations:

(1) They require the first frame to be colored and cannot use related frames.
(2) They are unable to propagate between scene changes, and thus require precise scene segmentation. This doesn't allow handling scenes that alternate back and forth, as commonly done in movies, which end up requiring many additional colorized references.
(3) Once they make an error they continue amplifying it. This severely limits the number of frames that can be propagated.

In contrast to RNN-based approaches, our approach is able to handle multiple scenes or entire videos seamlessly as shown in Fig. 3. Instead of using a RNN, we use a CNN with temporal convolutions and attention, which allows incorporating non-local information from multiple input frames to colorize a single output frame.



(a) Recursion-based CNN.

(b) Our Source-Reference Attention-based CNN.

Fig. 3. **Comparison between recursion-based and attention-based Convolutional Neural Networks (CNN) when processing an input video x with reference images z.** Recursion-based networks simply propagate the information frame-by-frame, and because of this can not be processed in parallel and are unable to form long-term dependencies. Each time a new reference image is used, the propagation is restarted, and temporal coherency is lost. Source-reference attention-based networks, such as our approach, are able to use all the reference information when processing any the frames.

2.3 Attention

Attention mechanisms for neural networks were original developed for Natural Language Translation (NLT) [Bahdanau et al. 2015]. Similar to human attention, attention for neural network allows the model to focus on different parts of the input. For NLT, attention allows to find a mapping between the input language words and the output language words, which can be in different orders. For natural language processing, many different variants have been proposed such as global and local attention [Luong et al. 2015], self-attention [Cheng et al. 2016; Parikh et al. 2016] with large-scale studies being performed [Britz et al. 2017; Vaswani et al. 2017].

Computer vision has also seen applications of attention for caption generation of images [Xu et al. 2015], where the generation of

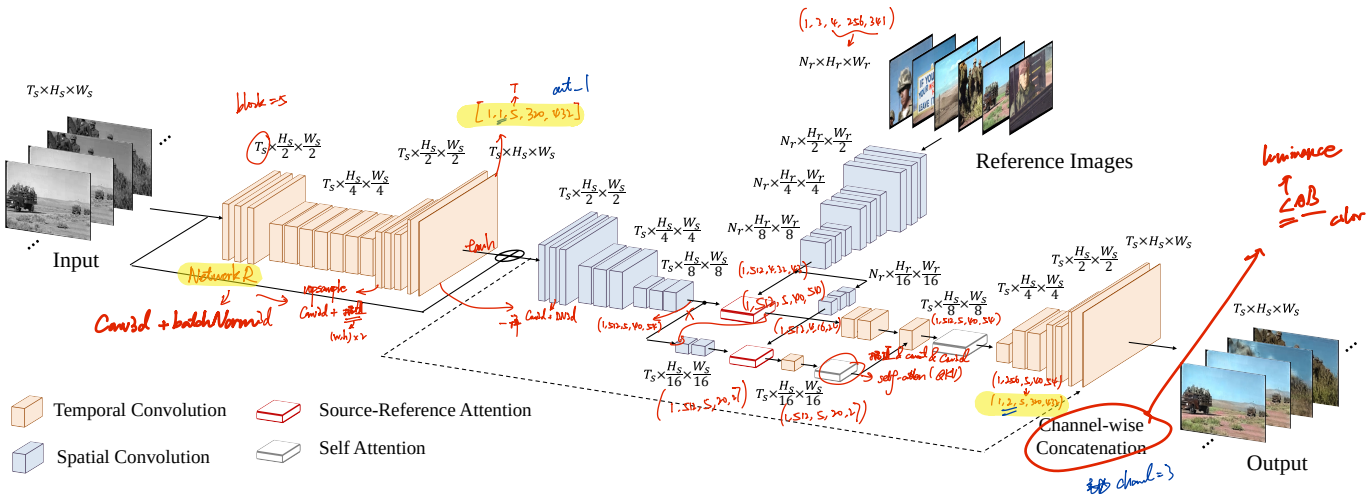


Fig. 4. **Overview of the proposed approach.** The model is input a sequence of black and white images which are restored using a pre-processing network and used as the luminance channel of the final output video. Afterwards, a source-reference network uses an arbitrary number of reference color images in conjunction with the output ofthe pre-processing network to produce the final chrominance channels of the video. Source-reference attention is used to allow the model to employ the color of similar regions in the reference color images when colorizing the video. The output of the model is a remastered video of the input.

each word in the caption can focus on different parts of the image using attention. Parmar *et al.* [2018] proposed using self-attention for image generation where pixels locations are explicitly encoded. This was later simplified in [Zhang et al. 2018a] to not need to explicitly encode the pixel locations. More related to our approach is the extension of self-attention to videos classification [Wang et al. 2018], in which the similarity of objects in the different video frames is computed with a self-attention mechanism. This is shown to improve the classification results of videos. Our approach is based on the same concept, however, we extend it to compute the similarity between the input video frames and an arbitrary number of reference images.

3 APPROACH

Our approach is based on fully convolutional networks, which are a variant of convolutional neural networks in which only convolutional layers are employed. This allows processing images and videos of any resolution. We employ a mix of temporal and spatial convolution layers, in addition to attention-based mechanisms that allow us to use an arbitrary number of reference color images during the remastering. An overview of the proposed approach can be seen in Fig. 4.

3.1 Source-Reference Attention

We employ source-reference attention to be able to supply an arbitrary number of reference color images that the model can be use as hints for the remastering of videos. In particular, source-reference attention layers take as an input two different variable length volumetric feature maps, one corresponding to the source data and the other to the reference data, and allow the model to exploit non-local similarities between the source data and the reference data. The model can thus use the color from the reference data to colorize similar areas of the source data.

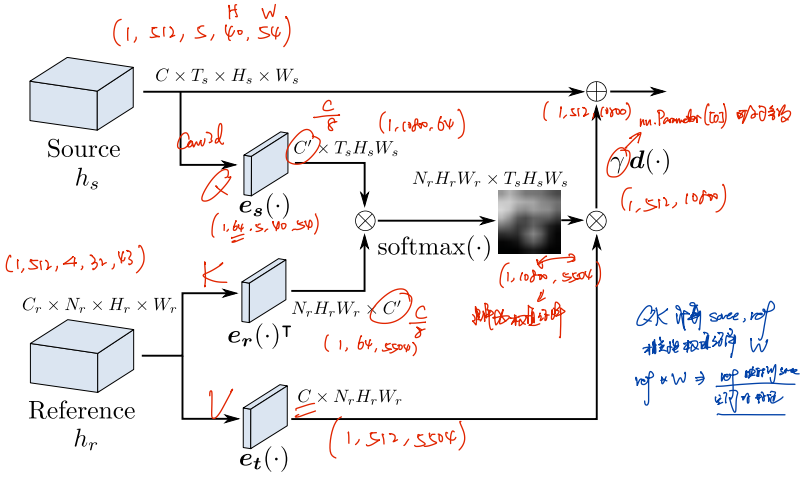


Fig. 5. **Overview of the source-reference attention layer.** This layer takes both a set of reference feature maps h_r and a set of source feature maps h_s as an input, and outputs a new set of feature maps of the same dimension as the source feature maps. This attention allows using non-local features from the reference features to perform a transformation of the source reference features. This transformation is done while conserving the local information similar to a purely convolutional layer. We denote matrix multiplication as "⊗" and matrix addition as "⊕". The input and output dimensions of the different components are shown for reference.

More formally, let the source data feature representation be $h_s \in \mathbb{R}^{C \times T_s \times H_s \times W_s}$ with C channels, T_s frames of height H_s and width W_s, and let the reference data features be $h_r \in \mathbb{R}^{C_r \times N_r \times H_r \times W_r}$ with C_r channels, N_r maps of height H_r and width W_r. The source-reference attention layer $A_{sr}(\cdot,\cdot)$ can be defined as

$$A_{sr}(h_s, h_r) = h_s + \gamma\, \boldsymbol{d}\Big(\boldsymbol{e_t}(h_r)\ \mathrm{softmax}\big(\boldsymbol{e_r}(h_r)^\intercal\ \boldsymbol{e_s}(h_s)\big)\Big) \quad , \quad (1)$$

where $\gamma \in \mathbb{R}$ is a learnt parameter and

$$\begin{aligned} \boldsymbol{e_s} &: \mathbb{R}^{C\times T_s\times H_s\times W_s} \rightarrow \mathbb{R}^{C'\times T_s H_s W_s} \\ \boldsymbol{e_r} &: \mathbb{R}^{C_r\times N_r\times H_r\times W_r} \rightarrow \mathbb{R}^{C'\times N_r H_r W_r} \\ \boldsymbol{e_t} &: \mathbb{R}^{C_r\times N_r\times H_r\times W_r} \rightarrow \mathbb{R}^{C\times N_r H_r W_r} \end{aligned} \tag{2}$$

are encoding functions that map the input source and reference feature tensors to matrices with a reduced number of channels, and $\boldsymbol{d}: \mathbb{R}^{C\times T_s H_s W_s} \rightarrow \mathbb{R}^{C\times T_s\times H_s\times W_s}$ is a decoding function that simply reshapes the tensor without modifying any values. For the encoding functions, we use temporal convolution operators with 1×1×1-pixel kernels followed by reshaping to the appropriate output dimensions. A visual overview of the source-reference attention layer is shown in Fig. 5.

Note that if reference data features are not provided, the output of the source-reference attention layer becomes simply the source data features. Furthermore, in the case the same features are used for both the source and reference features, the source-reference attention layer becomes a self-attention layer similar to the implementation of [Zhang et al. 2018a], except using temporal convolutions instead of spatial convolutions for the encoders, and the dot-product implementation of [Wang et al. 2018], where a single multiplicative parameter γ is used in place of a convolution operator in the decoder. To reduce the computational burden of the attention, we set $C' = C/8$ unless mentioned otherwise.

3.2 Model

The model is composed fundamentally of two trainable parts: a pre-processing network, and a source-reference network. Both are fully differentiable and trained together in an end-to-end fashion. We follow the best practices of fully convolutional networks by having each convolution layer consist of a convolution operator, followed by a Batch Normalization (BN) layer [Ioffe and Szegedy 2015], and a Exponential Linear Unit (ELU) activation function [Clevert et al. 2015], unless specified otherwise. Unless specified otherwise, all convolutions operate in the temporal domain, with spatial convolution operators using kernel of size $1 \times 3 \times 3$, and temporal convolution operators using kernel of size of $3 \times 3 \times 3$. All layers use padding so that the output is the same size as the input .The resolution is decreased with layers that used strides of $1 \times 2 \times 2$-pixels, and increased with trilinear up-sampling before the convolution layers when necessary. A full overview of the model can be seen in Fig. 4.

3.2.1 Pre-Processing Network. The pre-processing network is formed exclusively by temporal convolution layers, and uses a skip connection between the input and output. The main objective of the pre-processing network is to remove artefacts and noise from the input greyscale video. The network uses an encoder-decoder architecture in which the resolution is halved twice and restored to the full size at the end with trilinear upsampling. A full overview of the pre-processing model architecture is shown in Table 1. Most of the processing is done at the low resolution to decrease the computational burden, and the output of this network is used as the luminance channel of the final output image.

3.2.2 Source-Reference Network. The source-reference network forms the core of the model and takes as an input the output of the pre-processing network along with an arbitrary number of user-provided reference color images. Two forms of attention are employed to allow non-local information to be used when computing the output chrominance maps: source-reference attention allows information from reference color images to be used, giving the user indirect control of the colorization; and self-attention allows non-local temporal information to be used, increasing the temporal consistency of the colorization. For self-attention, we use the source-reference attention layer implementation and use the same features for both the source and reference feature maps. An overview of the source-reference model architecture is shown in Table 2.

As with the pre-process network, the model is based on a encoder-decoder architecture, where the resolution is reduced to allow for more efficient computation and lower memory usage, and restored for the final output. While temporal convolutions allow for better temporal consistency, they also complicate the learning and increase the computational burden. Unlike the pre-processing network, the source-reference network uses a mix of temporal and spatial convolutions. In particular, the decoder and 1/8 middle branch use temporal convolutions while the encoders of both the input video and reference images use spatial convolutions, and the 1/16 middle branch uses a mixture of both, which we found decreases memory usage and simplifies the training, while not sacrificing any remastering accuracy. Furthermore, in the case of the reference color images, there is no temporal coherency to be exploited by using temporal convolutions as the images are not necessarily related.

First, the input video and reference images are separately reduced to 1/8 of the original width and height in three stages by separate encoders. The encoded input video and reference video features are then split into two branches: one processes the video at 1/8 width and height, and one decreases the resolution another stage to 1/16 of the original width and height to further process the video. Both branches employ source-reference attention layers, additional temporal convolution layers, and self-attention layers. In particular,

Table 1. **Overview of the pre-processing model architecture.** We abbreviate Temporal Convolution with "TConv.". Layer irregularities are specified in the notes column. When the same layer is repeated several times consecutively, we indicate this with the number of times in parenthesis.

Layer Type	Output Resolution	Notes
Input	$1 \times T_s \times W_s \times H_s$	Input greyscale image
TConv.	$64 \times T_s \times W_s/2 \times H_s/2$	Replication padding, spatial stride of 2
TConv. (×2)	$128 \times T_s \times W_s/2 \times H_s/2$	
TConv.	$256 \times T_s \times W_s/4 \times H_s/4$	Spatial stride of 2
TConv. (×4)	$256 \times T_s \times W_s/4 \times H_s/4$	
TConv.	$128 \times T_s \times W_s/2 \times H_s/2$	Trilinear upsampling
TConv. (×2)	$64 \times T_s \times W_s/2 \times H_s/2$	
TConv.	$16 \times T_s \times W_s \times H_s$	Trilinear upsampling
TConv.	$1 \times T_s \times W_s \times H_s$	TanH output, input is added, and finally clamped to $[0, 1]$ range

Table 2. **Overview of the source-reference model architecture.** This model takes as an input both the output of the pre-processing model and a set of reference images. Both these inputs are processed by separate encoders (a), then processed in two different middle branches corresponding to 1/16 width and height (b), and 1/8 width and height (c), before being decoded to the chrominance channels of the output video with a decoder (d). We abbreviate Spatial Convolutions with "SConv.", Temporal Convolutions with "TConv.", and Source-Reference Attention with "SR Attn.". For the source and reference encoders, we refer to the temporal dimension generically as T, where $T = T_r$ for the reference encoder and $T = T_s$ for the source encoder. We specify layer irregularities in the notes column. When the same layer is repeated several times consecutively, we indicate this with the number of times in parenthesis.

(a) Source and Reference Encoders.

Layer Type	Output Resolution	Notes
Input	$(1 \text{ or } 3) \times T \times W \times H$	3 channels (RGB) for reference, 1 channel (greyscale) for source
SConv.	$64 \times T \times W/2 \times H/2$	Spatial stride of 2
SConv. (×2)	$128 \times T \times W/2 \times H/2$	
SConv.	$256 \times T \times W/4 \times H/4$	Spatial stride of 2
SConv. (×2)	$256 \times T \times W/4 \times H/4$	
SConv.	$512 \times T \times W/8 \times H/8$	Spatial stride of 2
SConv. (×2)	$512 \times T \times W/8 \times H/8$	

(b) Middle 1/16 branch.

Layer Type	Output Resolution	Notes
SConv.	$512 \times T_s \times W_s/16 \times H_s/16$	Input is source encoder output, spatial stride of 2
SConv.	$512 \times T_s \times W_s/16 \times H_s/16$	Outputs 1/16 source
SConv.	$512 \times N_r \times W_r/16 \times H_r/16$	Input is reference encoder output, spatial stride of 2
SConv. (×2)	$512 \times N_r \times W_r/16 \times H_r/16$	Outputs 1/16 reference
SR Attn.	$512 \times T_s \times W_s/16 \times H_s/16$	Uses 1/16 source and reference as inputs
TConv.	$512 \times T_s \times W_s/16 \times H_s/16$	
Self Attn.	$512 \times T_s \times W_s/16 \times H_s/16$	

(c) Middle 1/8 branch.

Layer Type	Output Resolution	Notes
SR Attn.	$512 \times T_s \times W_s/8 \times H_s/8$	Input is source and reference encoder output
TConv. (×2)	$512 \times T_s \times W_s/4 \times H_s/4$	
TConv. (×2)	$512 \times T_s \times W_s/4 \times H_s/4$	Output of the 1/16 branch is concatenated to the input
Self Attn.	$512 \times T_s \times W_s/4 \times H_s/4$	

(d) Decoder.

Layer Type	Output Resolution	Notes
TConv.	$256 \times T_s \times W_s/8 \times H_s/8$	
TConv.	$128 \times T_s \times W_s/4 \times H_s/4$	Trilinear upsampling
TConv.	$64 \times T_s \times W_s/4 \times H_s/4$	
TConv.	$32 \times T_s \times W_s/2 \times H_s/2$	Trilinear upsampling
TConv.	$16 \times T_s \times W_s/2 \times H_s/2$	
TConv.	$8 \times T_s \times W_s \times H_s$	Trilinear upsampling
TConv.	$2 \times T_s \times W_s \times H_s$	Sigmoid output represents chrominance

the 1/16 branch is processed with a self-attention layer before being upsampled trilinearly and concatenated to the 1/8 branch output. The resulting combined features are processed with self-attention to be more temporally uniform. Afterwards, a decoder converts the features to chrominance channels in three stages using trilinear upsampling. Finally the output of the network is used as the image chrominance with two channels corresponding to the *ab* channels of the *Lab* color-space, while the output of the pre-processing network is used as the image luminance corresponding to the L channel.

4 TRAINING

We train our model using manually curated supervised training data. In order to improve both the generalization and quality of the results, we perform large amounts of both synthetic data augmentation and example-based film deterioration.

4.1 Objective Function

We train the model in a fully supervised fashion with a linear combination of two L_1 losses. In particular, we use a supervised dataset $\mathcal{D}$ consisting of pairs of deteriorated black and white videos x and restored color videos split into luminance y_l and chrominance y_{ab} using the *Lab* color-space, and reference color images z, and optimize the following expression:

$$\arg\min_{\theta,\phi} \mathbb{E}_{(x,y_l,y_{ab},z)\in\mathcal{D}} \|P(x;\theta) - y_l\| + \beta \|S(P(x;\theta),z;\phi) - y_{ab}\|, \tag{3}$$

where P is the pre-processing model with weights θ, S is the source-reference model with weights ϕ, and $\beta \in \mathbb{R}$ is a weighting hyper-parameter.

Training is done using batches of videos with 5 sequential frames each, that are chosen randomly from the training data. For each 5-frame video, a random number of color references images z is chosen uniformly from the $[0, 6]$ range. If the number of references is not 0, one of the reference images is chosen to be from within five neighboring frames of the input frames, while the remaining reference images are randomly sampled from the whole training data set.

Table 3. **Overview of the different types of data augmentation used during training.** The target refers to which data is being augmented. Values in parenthesis indicate that the same transformation is done jointly to both variables, instead of independently. Probability indicates how likely that particular transformation is likely to occur, and range is how the transformation parameters are sampled. We note that in the case of the input video x and target video y, the same transformation is done to all the frames in the video, while in the case of the reference images z, the transformation is done independently for each image as they are not related to each other.

Name	Target	Prob.	Range	Notes
Horiz. Flip	$(x, y), z$	50%	-	
Scaling	(x, y)	100%	$\mathcal{U}(256, 400)$	Size of the smallest edge (px), randomly crops
Rotation	(x, y)	100%	$\mathcal{U}(-5, 5)$	In degrees
Brightness	(x, y)	20%	$\mathcal{U}(0.8, 1.2)$	
Contrast	(x, y)	20%	$\mathcal{U}(0.9, 1.0)$	
JPEG	x, z	90%	$\mathcal{U}(15, 40)$	Encoding quality
Noise	x, z	10%	$\mathcal{N}(0, 0.04)$	Gaussian
Blur	x	50%	$\mathcal{U}(2, 4)$	Bicubic down-sampling
Contrast	x	33%	$\mathcal{U}(0.6, 1.0)$	
Scaling	z	100%	$\mathcal{U}(256, 320)$	Size of the smallest edge (px), randomly crops
Saturation	z	10%	$\mathcal{U}(0.3, 1.0)$	

4.2 Training Data

We base our dataset on the YouTube-8M dataset [Abu-El-Haija et al. 2016] which consists of roughly 8 million videos corresponding to about 500 thousand hours of video data. The dataset is annotated with 4,803 visual entities which we do not use. We convert the videos to black and white and corrupt them, simulating old film degradation, to create supervised training data for our model.

As YouTube-8M dataset was created by mostly automatically, a large amount of videos depict gameplay, black and white video, static scenes from fixed cameras, and unnaturally colored scenes such as clubs with live music. We randomly select videos from the full dataset and manually annotate them as suitable for training and evaluating a remastering model. In particular, we end up with 1,569 videos totalling 10,243,010 frames, of which we use 1,219 (7,993,132 frames) for training our model, 50 (321,306) for validation, and 300 (1,928,572) for testing.

4.3 Data Augmentation

We perform large amounts of data augmentation to the input video, ground truth video, and reference images with two objectives: first, we wish to increase the generalization of the model to different types of video, and secondly, we want the model to be able to restore different artefacts which can be commonly found in the input videos, such as blur or low contrast. This data augmentation is done in parallel with example-based deterioration that further degrades the input greyscale videos.

We use batches of 5-frame videos with their associated reference images with a resolution of 256×256 pixels. As data augmentation, we perform a large amount of transformations that affect the input video x and ground truth video $y = (y_l, y_{ab})$ together, only the input video x, only the reference images z, or any combination of the previous three. An overview of the different transformations we apply is shown in Table 3, which include changes to brightness, contrast, JPEG noise, Gaussian noise, blurring, and saturation.

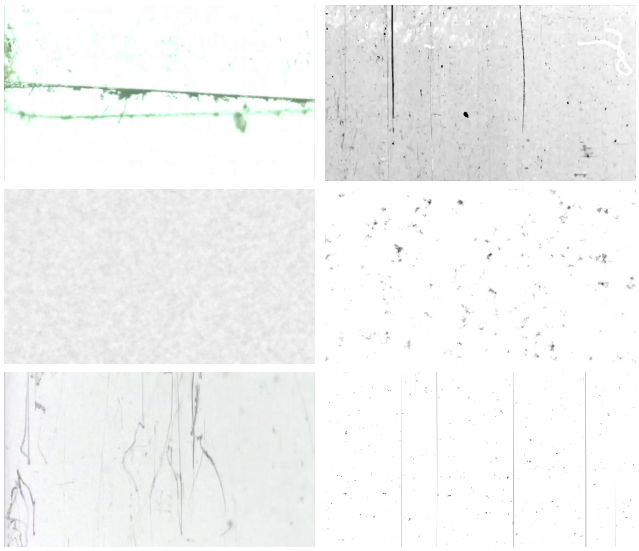


Fig. 6. **Example-based deterioration effects.** These effects are generated offline and stored as a dataset of images which can then be applied to training data inputs as additive noise.

4.4 Example-based Deterioration

In addition to all the different data augmentation techniques, we also simulate deterioration of the film medium from a dataset of 6,152 examples, such as fractal noise, grain noise, dust, and scratches. These deterioration examples are manually collected by web search using the keywords "film noise", and also generated using software such as Adobe After Effects. For generated noise, fractal noise is used to generate a base noise pattern, which can then be improved by modifying the contrast, brightness, and tone curves to obtain scratch and dust-like noise. In total, 4,340 noise images were downloaded and 1,812 were generated. Some of the deterioration examples are shown in Fig. 6. In particular, as these deterioration effects simulate the degradation of the physical medium which is supporting the film, they are implemented as additive noise: the noise data is randomly added to the input greyscale video, independently for each frame. Furthermore, they are added independently of each other and combined to create the augmented input videos.

For all the noise, we use similar data augmentation techniques as used for the input video. In particular, the noise images are scaled randomly such that the shortest edge is between [256, 720] pixels, both horizontally and vertically flipped with 50% probability, rotated randomly between $[-5, 5]$ degrees, cropped to 256×256 pixels, randomly scaled by $\mathcal{U}(0.5, 1.5)$, and randomly either subtracted or added to the original image. Some generated training examples are shown in Fig. 7.

(a) Original (b) Deteriorated Example 1

(c) Deteriorated Example 2 (d) Deteriorated Example 3

Fig. 7. **Examples of synthetic deterioration effects applied to a black and white image.** (a) For the original image, (b-d) various types of both algorithm-based and example-based deterioration effects, such as JPEG compression artifacts and film scratches, are randomly added. Video licensed in the Public Domain.

4.5 Optimization

Training is initially done of the pre-processing network and source-reference network separately for 500,000 iterations. Afterwards, they are trained together in an end-to-end fashion by optimizing Eq. (3). For the optimization method, we rely on the ADADELTA algorithm [Zeiler 2012], which is a variant of stochastic gradient descent which heuristically estimates the learning rate parameter, thus requiring no hyper-parameters to tune.

5 RESULTS

We train our model on our dataset with $\gamma = 10^{-4}$ and a batch-size of 20. We use the model with the lowest validation loss as our final model. We evaluate both quantitative and qualitative and compare with existing methods.

5.1 Comparison with Existing Approaches

We compare the results of our approach with both existing approaches and strong baselines with a quantitative evaluation. In particular, for restoration, we compare against the approach of [Zhang et al. 2017b] and [Yu et al. 2018], and for colorization we compare against the propagation-based approach of [Vondrick et al. 2018] and single-image interactive approach of [Zhang et al. 2017a]. For both remastering, *i.e.*, joint restoration and colorization, we compare against all possible combinations of restoration and colorization approaches, *e.g.*, the combination of [Zhang et al. 2017b] and [Vondrick et al. 2018] used together. The approach of [Zhang et al. 2017b] and [Yu et al. 2018] consists of a deep residual convolutional neural network for single image restoration. We note that the approach of [Yu et al. 2018] is an extension of [Fan et al. 2018] and winner of the NTIRE 2018 super resolution image challenge[1]. We modified the model of [Yu et al. 2018] by removing the up-sampling layer at the end as the target task is restoration and not super-resolution. The approach of [Vondrick et al. 2018] is a recursive convolutional neural network that can propagate color information. The approach of [Zhang et al. 2017a] is a single-image convolutional neural network approach that can use user-provided hints, which we use to provide the reference image color information. We also compare against two strong colorization baselines consisting of our proposed model with the temporal convolution layers replaced with spatial convolution layers, and of our proposed model without self-attention layers. For restoration, we compare to a baseline consisting of our pre-processing network without the skip connection. Finally, we also compare against a baseline consisting of the restoration and colorization networks of our approach trained independently, *i.e.*, without joint training. All approaches are trained using exactly the same training data and training approach for fair comparison.

We compare using our test set consisting of 300 videos from the Youtube-8M dataset. For each video we randomly sample a subset of either 90 or 300 frames, and use the subset as the ground truth. Given that these videos are not noisy nor degraded, we follow the same approach for generating training data to generate deteriorated inputs for evaluation. For the example-based deterioration effects, we use a different set of images from those of the training set to evaluate generalization. We use Peak Signal-to-Noise Ratio (PSNR) as an evaluation metric, and compute the PSNR using the luminance channel only for the restoration task, using the chrominance channels only for the colorization task, and using all the image channels for the remastering task.

For the reference color images, in the case of the 90 frame subset, we only provide the first frame as a reference image, while in the case of the 300 frame subset, we provide every 60th frame starting from the first frame as a reference image. For our approach, all the reference frames are provided at all times. In the case of the approach of [Vondrick et al. 2018], as it only propagates the color and is unable to naturally handle multiple reference images, we replace the output image with the new reference image when necessary as shown in Fig. 3. We note that the same random subset of all videos is used for all the approaches.

5.1.1 Remastering Results. As there is not a single approach that can handle the remastering of videos, we compare against a pipeline approach of first processing the video with the method of either [Zhang et al. 2017b] or [Yu et al. 2018], and then propagating the reference color on the output with the approach of either [Vondrick et al. 2018] or [Zhang et al. 2017a]. We also provide results of a baseline consisting of our full approach without the joint training, *i.e.*, the restoration and colorization networks are trained independently. Results are shown in Table 4. Of the pipeline-based approaches, we find that, while they have similar performance, the combination of [Zhang et al. 2017b] and [Zhang et al. 2017a] gives the highest performance. However, our approach outperforms the existing pipeline based approaches and the strong baseline that doesn't use joint training. This shows that even though the restoration and colorization models are first trained independently before being further trained jointly, the joint training plays an important role in improving the quality of the final results. It is also interesting to point out that while the performance of existing approaches degrades with longer

[1] http://www.vision.ee.ethz.ch/ntire18/

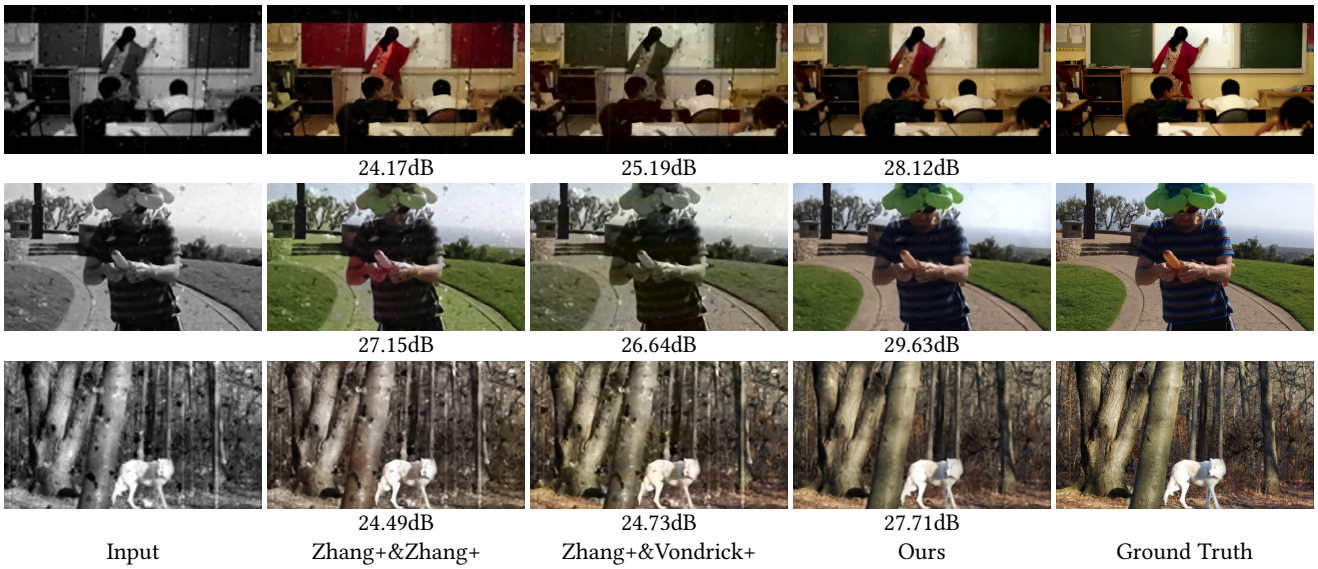


Fig. 8. **Randomly sampled examples from the Youtube-8M test dataset with degradation noise.** We show one frame from several examples and compare our approach with the combined approach of [Zhang et al. 2017b] andk [Zhang et al. 2017a], and [Zhang et al. 2017b] and [Vondrick et al. 2018]. First column shows the input frame which has been deteriorated with noise, the next two columns correspond to the remastering results with both approaches, and the last column shows the ground truth video. The PSNR of each approach is shown below each image. Videos courtesy of Naa Creation (top), Balloon Sage (middle), and Mayda Tapanes (bottom) and licensed under CC-by.

videos and more reference color images, our approach improves in performance. This is likely due to all the reference color images being used to remaster each frame. Several randomly chosen examples are shown in Fig. 8, where we can see that existing approaches fail to both remove the noise and propagate the color, while our approach performs well in both cases.

5.1.2 Restoration Results. We compare our approach with that of [Zhang et al. 2017b], [Yu et al. 2018], and a baseline for video restoration. The baseline consists of our pre-processing model without the skip connection that adds the input to the output. As color is not added, no reference color images are provided and the evaluation is done using only the 300 frame subset. Results are shown in Table 5. We can see that the baseline, the approach of [Zhang et al. 2017b], and the approach of [Yu et al. 2018] perform similarly, while our full pre-processing model, with a skip connection, outperforms both. Example results are shown in Fig. 9.

5.1.3 Colorization Results. We compare against the approach of [Zhang et al. 2017a] using global hints, the approach of [Vondrick et al. 2018] and two baselines: one consisting of our source-reference network without temporal convolutions and one without self-attention for colorization. Results are shown in Table 6, and we can see that our approach outperforms existing approaches and the baselines. Similar to the remastering case, our approach performs significantly better on longer videos with additional references images, which is indicative of the capabilities of the source-reference attention: not only is it possible to colorize long sequences with many reference images, it is beneficial for performance. An interesting result is that self-attention plays a critical role in our model. We believe this is due to the fact it allows each output pixel to be computed using information from the entire image, which would require many more convolutional layers if self-attention was not employed. Example results are shown in Fig. 10.

Table 4. **Quantitative remastering results.** We compare the results of our model with that of restoring each frame with the approach of [Zhang et al. 2017b], and propagating reference color with the approach of [Vondrick et al. 2018] on synthetically deteriorated videos from the Youtube-8M dataset, and with a baseline that consists of our model without using joint training. We perform two types of experiments: one using a random 90-frame subset from each video with 1 reference frame, and one using a random 300-frame subset with 5 reference frames.

Approach	Frames	# Ref.	PSNR
Zhang+[2017b]&Zhang+[2017a]	90	1	27.13
	300	5	27.31
Yu+[2018]&Zhang+[2017a]	90	1	26.43
	300	5	26.59
Zhang+[2017b]&Vondrick+[2018]	90	1	26.43
	300	5	26.60
Yu+[2018]&Vondrick+[2018]	90	1	26.85
	300	5	26.89
Ours w/o joint training	90	1	29.07
	300	5	29.23
Ours	90	1	**30.83**
	300	5	**31.14**

Fig. 9. **Restoration results on the Youtube-8M test dataset with degradation noise.** We show one frame from several examples and compare our approach with the approaches of [Zhang et al. 2017b] and [Yu et al. 2018]. The first column shows the input frame which has been deteriorated with noise, the next three columns correspond to the black and white restoration of each approach, and the last column corresponds to the ground truth video. Videos courtesy of Naa Creation (top), and Mayda Tapanes (bottom) and licensed under CC-by.

Fig. 10. **Colorization results on the Youtube-8M test dataset**. We show one frame from several examples and compare our approach with the colorization approach of [Zhang et al. 2017a] without using the reference image and the RNN-based approach [Vondrick et al. 2018] which uses the reference image. The first column shows the input frame, the next three columns correspond to the colorization of each approach, and the last corresponds to the reference image. Note that the input frame is not the same frame as the reference image. Videos courtesy of Naa Creation (top), and Mayda Tapanes (bottom) and licensed under CC-by.

Table 5. **Quantitative restoration results.** We compare the results of our pre-processing network with the approach of [Zhang et al. 2017b], [Yu et al. 2018], and a baseline of our approach without the skip connection for restoring synthetically deteriorated videos from the Youtube-8M dataset.

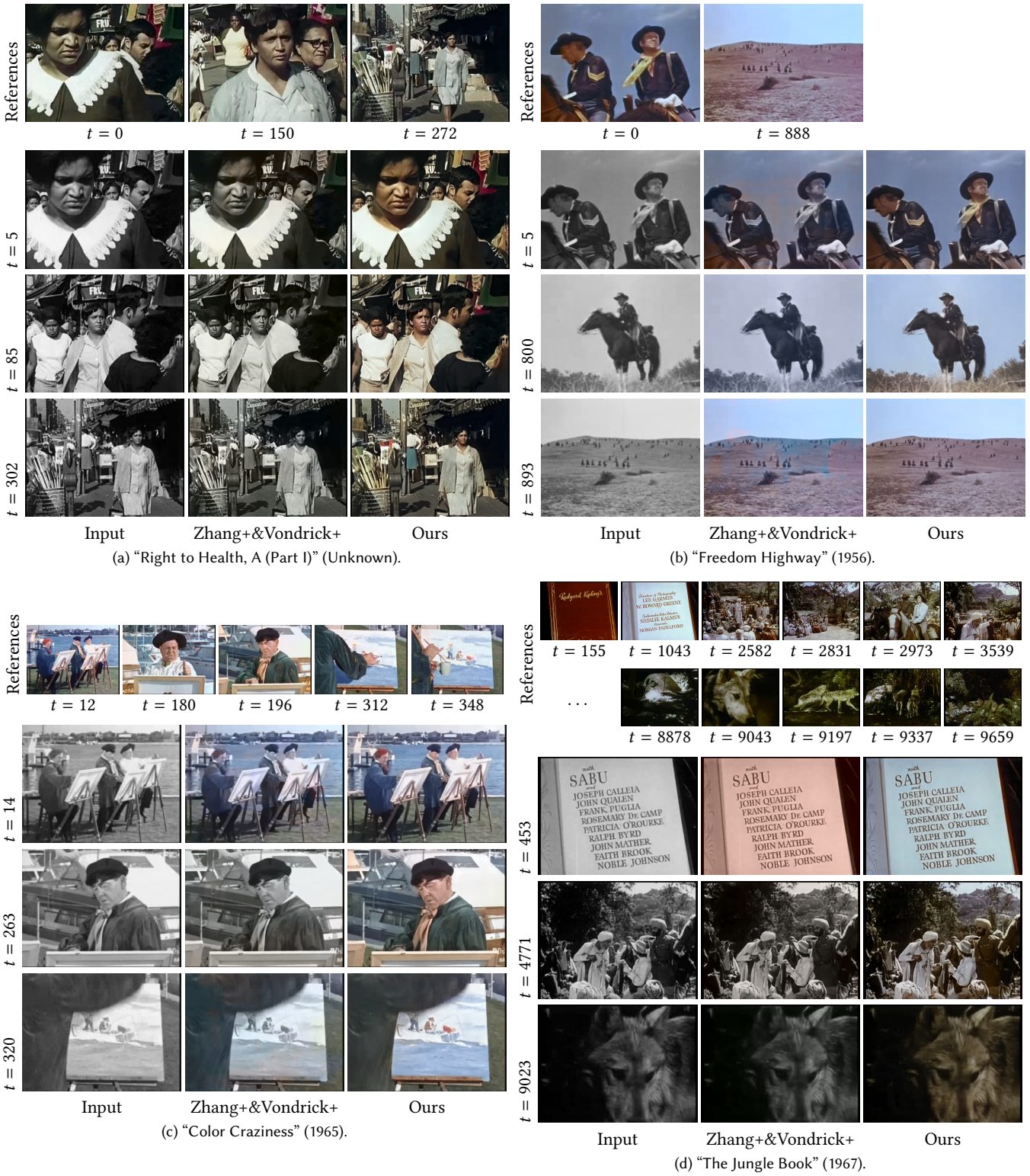
Approach	Frames	# Ref.	PSNR
[Zhang et al. 2017b]	300	-	25.08
[Yu et al. 2018]	300	-	24.49
Ours w/o skip connection	300	-	24.73
Ours	300	-	**26.13**

5.2 Qualitative Results

We show qualitative results in Fig. 11 on diverse challenging real world vintage film examples. As the videos are originally color, we use images from the original video as the reference images, and then compare both our remastering approach and a pipeline approach of denoising with the approach of [Zhang et al. 2017b] and then adding color with the method of [Vondrick et al. 2018]. We can see how our approach is able to perform a consistent remastering, while existing approaches lose track of the colorization and fail to

Table 6. **Quantitative colorization results.** We compare the colorization results of our source-reference network with the approach of [Zhang et al. 2017a] using global hints, and [Vondrick et al. 2018] for the colorization of videos from the Youtube-8M dataset. We perform two types of experiments: one using a random 90-frame subset from each video with 1 reference frame, and one using a random 300-frame subset with 5 reference frames.

Approach	Frames	# Ref.	PSNR
[Zhang et al. 2017a]	90	1	31.28
	300	5	31.16
[Vondrick et al. 2018]	90	1	31.55
	300	5	31.70
Ours w/o temporal conv.	90	1	28.46
	300	5	28.51
Ours w/o self-attention	90	1	29.00
	300	5	28.72
Ours	90	1	**34.94**
	300	5	**36.26**



(a) "Right to Health, A (Part I)" (Unknown).

(b) "Freedom Highway" (1956).

(c) "Color Craziness" (1965).

(d) "The Jungle Book" (1967).

Fig. 11. **Qualitative comparison with the combined approach of Zhang+[2017b] and Vondrick+[2018].** We show the reference color images in the first row with their timestamps. Afterwards four different frames taken from the input video and output videos are shown. Note that the example of (d) is remastered with 41 reference images of which we only show a subset. "Right to Health, A (Part I)", "Freedom Highway", "Color Craziness", and "The Jungle Book" are licensed in the public domain.

Fig. 12. **Restoration result on vintage film.** We compare with the approach of [Zhang et al. 2017b], and show the boxed area zoomed in on the bottom row. We can see that the relatively large noise is "inpainted" with our network. First two rows are frames taken from the movie "Oliver Twist" (1933) which is licensed in the public domain.

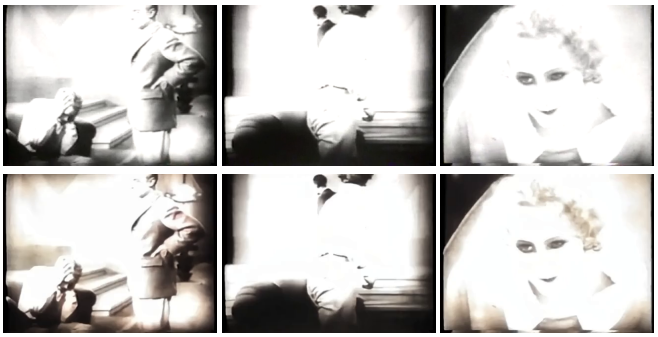


Fig. 13. **Limitation of our approach.** Example of severely deteriorated film which is not possible to remaster with the current approach. The first row shows frames from the original input video and the second row shows the output of our approach. Images taken from the movie "Metropolis" (1925) which is licensed in the public domain.

produce pleasing results, which is consistent with our quantitative evaluation.

We also perform a qualitative comparison of restoration results on vintage film in Fig. 12 with the approach of [Zhang et al. 2017b]. We can see how the approach of [Zhang et al. 2017b] can restore small noise, but fails at larger noise. Our approach is able to handle both small and large noise, while also sharpening the input image.

5.3 Computation Time

For a 528×400-px input video, our approach takes 69ms per frame with a Nvidia GTX 1080Ti GPU, with 4ms corresponding to the restoration stage, and 65ms corresponding to the colorization stage.

6 LIMITATIONS AND DISCUSSION

We have presented an approach for the remastering of vintage film based on temporal convolutional networks with source-reference attention mechanisms that allow for using an arbitrary number of reference color images. Although the source-reference attention mechanism is a powerful tool to incorporate reference images into a processing framework and is amenable to process videos of any resolution, it suffers from $O(N_r H_r W_r T_s H_s W_s)$ memory usage. Available system memory will thus limit the maximum resolution that can be processed. However,in practice, as most vintage movies are stored at low resolutions due to limits of the film technology, they do not have to be processed at resolutions that would not be possible with attention-based mechanisms.

Currently, the proposed approach relies on fully supervised learning and can not fill missing frames nor extreme degradation that leaves a large region of the image missing during many frames as shown in Fig. 13. In these cases there is too much missing information which makes it impossible to remaster, they would require image completion-based approaches to remake new plausible parts of the video, which is out of the scope of this work.

Our model has a temporal resolution of 15 frames, corresponding to roughly half a second in most videos, which can lead to small temporal consistencies in the output video. For reference, existing approaches use a smaller amount such as 4 frames [Vondrick et al. 2018] or 10 frames [Lai et al. 2018]. While it should be possible to increase the temporal resolution, this leads to slower convergence and slower computation. While blind video temporal consistency techniques can alleviate this issue [Bonneel et al. 2015; Lai et al. 2018], we found that while they are able to slightly improve the temporal consistency, it comes at the cost of significantly worse results. We believe that integrating such an approach with our model and training end-to-end is a possible way to improve the temporal consistency without sacrificing the quality of the results.

We note that despite the progress in this work on remastering vintage film, due to the complexity of the task, it still is an open problem in computer graphics. Unlike most of the image and video research up until now, vintage film poses a much more difficult and realistic problem as highlighted in Fig. 2, and we hope that this work can further stimulate research in this topic.

ACKNOWLEDGMENTS

This work was partially supported by JST ACT-I (Iizuka, Grant Number: JPMJPR16U3), JST PRESTO (Simo-Serra, Grant Number: JPMJPR1756), and JST CREST (Iizuka and Simo-Serra, Grant Number: JPMJCR14D1).

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. In *arXiv:1609.08675*. https://arxiv.org/pdf/1609.08675v1.pdf

Xiaobo An and Fabio Pellacini. 2008. AppProp: All-pairs Appearance-space Edit Propagation. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 27, 3 (Aug. 2008), 40:1–40:9.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. (2015).

Steve Bako, Thijs Vogels, Brian McWilliams, Mark Meyer, Jan Novák, Alex Harvill, Pradeep Sen, Tony Derose, and Fabrice Rousselle. 2017. Kernel-predicting convolutional networks for denoising Monte Carlo renderings. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 36, 4 (2017), 97–1.

Nicolas Bonneel, James Tompkin, Kalyan Sunkavalli, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. 2015. Blind Video Temporal Consistency. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 34, 6 (2015).

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive Exploration of Neural Machine Translation Architectures. In *Conference on Empirical Methods in Natural Language Processing*.

Chakravarty R Alla Chaitanya, Anton S Kaplanyan, Christoph Schied, Marco Salvi, Aaron Lefohn, Derek Nowrouzezahrai, and Timo Aila. 2017. Interactive reconstruction of Monte Carlo image sequences using a recurrent denoising autoencoder. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 36, 4 (2017), 98.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Conference on Empirical Methods in Natural Language Processing*.

Alex Yong-Sang Chia, Shaojie Zhuo, Raj Kumar Gupta, Yu-Wing Tai, Siu-Yeung Cho, Ping Tan, and Stephen Lin. 2011. Semantic Colorization with Internet Images. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 30, 6 (2011), 156:1–156:8.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations*.

K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. 2007. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Transactions on Image Processing* 16, 8 (2007), 2080–2095.

A. Danielyan, V. Katkovnik, and K. Egiazarian. 2012. BM3D Frames and Variational Image Deblurring. *IEEE Transactions on Image Processing* 21, 4 (2012), 1715–1728.

Yuchen Fan, Jiahui Yu, and Thomas S Huang. 2018. Wide-activated Deep Residual Networks based Restoration for BPG-compressed Images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. 2018. Deep exemplar-based colorization. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 37, 4 (2018), 47.

Yi-Chin Huang, Yi-Shin Tung, Jun-Cheng Chen, Sung-Wen Wang, and Ja-Ling Wu. 2005. An Adaptive Edge Detection Based Colorization Algorithm and Its Applications. In *ACMMM*. 351–354.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. 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 (Proceedings of SIGGRAPH)* 35, 4 (2016).

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*.

Revital Irony, Daniel Cohen-Or, and Dani Lischinski. 2005. Colorization by Example. In *Eurographics Conference on Rendering Techniques*. 201–210.

T. H. Kim, M. S. M. Sajjadi, M. Hirsch, and B. Schölkopf. 2018. Spatio-temporal Transformer Network for Video Restoration. In *European Conference on Computer Vision*.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. 2018. Learning Blind Video Temporal Consistency. In *European Conference on Computer Vision*.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. 2016. Learning representations for automatic colorization. In *European Conference on Computer Vision*.

Stamatios Lefkimmiatis. 2018. Universal Denoising Networks: A Novel CNN Architecture for Image Denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Anat Levin, Dani Lischinski, and Yair Weiss. 2004. Colorization using Optimization. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 23 (2004), 689–694.

Sifei Liu, Guangyu Zhong, Shalini De Mello, Jinwei Gu, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. 2018. Switchable Temporal Propagation Network. In *European Conference on Computer Vision*.

Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. 2008. Intrinsic Colorization. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 27, 5 (December 2008), 152:1–152:9.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*.

M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian. 2012. Video Denoising, Deblocking, and Enhancement Through Separable 4-D Nonlocal Spatiotemporal Transforms. *IEEE Transactions on Image Processing* 21, 9 (2012), 3952–3966.

M. Maggioni, E. SÃąnchez-Monge, and A. Foi. 2014. Joint Removal of Random and Fixed-Pattern Noise Through Spatiotemporal Video Filtering. *IEEE Transactions on Image Processing* 23, 10 (2014), 4282–4296.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. 2001. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *International Conference on Computer Vision*.

Simone Meyer, Victor Cornillère, Abdelaziz Djelouah, Christopher Schroers, and Markus Gross. 2018. Deep Video Color Propagation. In *British Machine Vision Conference*.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Conference on Empirical Methods in Natural Language Processing*.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, and Alexander Ku. 2018. Image Transformer. In *International Conference on Machine Learning*.

François Pitié, Anil C. Kokaram, and Rozenn Dahyot. 2007. Automated Colour Grading Using Colour Distribution Transfer. *Computer Vision and Image Understanding* 107, 1-2 (July 2007), 123–137.

Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. 2001. Color Transfer between Images. *IEEE Computer Graphics and Applications* 21, 5 (sep 2001), 34–41.

Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. 2017. Scribbler: Controlling deep image synthesis with sketch and color. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. 2005. Local Color Transfer via Probabilistic Segmentation by Expectation-Maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*. 747–754.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Conference on Neural Information Processing Systems*.

Thijs Vogels, Fabrice Rousselle, Brian McWilliams, Gerhard Röthlin, Alex Harvill, David Adler, Mark Meyer, and Jan Novák. 2018. Denoising with kernel prediction and asymmetric loss functions. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 37, 4 (2018), 124.

Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. 2018. Tracking emerges by colorizing videos. In *European Conference on Computer Vision*.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. 2002. Transferring Color to Greyscale Images. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 21, 3 (July 2002), 277–280.

Fuzhang Wu, Weiming Dong, Yan Kong, Xing Mei, Jean-Claude Paul, and Xiaopeng Zhang. 2013. Content-Based Colour Transfer. 32, 1 (2013), 190–203.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*.

Li Xu, Qiong Yan, and Jiaya Jia. 2013. A Sparse Control Model for Image and Video Editing. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 32, 6 (Nov. 2013), 197:1–197:10.

Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas S. Huang. 2018. Wide Activation for Efficient and Accurate Image Super-Resolution. *CoRR* abs/1808.08718 (2018). arXiv:1808.08718

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *CoRR* abs/1212.5701 (2012).

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018a. Self-Attention Generative Adversarial Networks. *arXiv preprint arXiv:1805.08318* (2018).

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

Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2018b. FFDNet: Toward a Fast and Flexible Solution for CNN based Image Denoising. *IEEE Transactions on Image Processing* (2018).

Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European Conference on Computer Vision*.

Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. 2017a. Real-Time User-Guided Image Colorization with Learned Deep Priors. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 9, 4 (2017).