Fully Automatic Video Colorization with Self-Regularization and Diversity

Chenyang Lei
HKUST

Qifeng Chen
HKUST

Abstract

We present a fully automatic approach to video colorization with self-regularization and diversity. Our model contains a colorization network for video frame colorization and a refinement network for spatiotemporal color refinement. Without any labeled data, both networks can be trained with self-regularized losses defined in bilateral and temporal space. The bilateral loss enforces color consistency between neighboring pixels in a bilateral space and the temporal loss imposes constraints between corresponding pixels in two nearby frames. While video colorization is a multi-modal problem, our method uses a perceptual loss with diversity to differentiate various modes in the solution space. Perceptual experiments demonstrate that our approach outperforms state-of-the-art approaches on fully automatic video colorization.

1. Introduction

There exist numerous classic films and videos in black-and-white. It is desirable for people to watch a colorful movie rather than a grayscale one. *Gone with the Wind* in 1939 is one of the first colorized films and is also the all-time highest-grossing film adjusted for inflation [1]. Image and video colorization can also assist other computer vision applications such as visual understanding [17] and object tracking [29].

Video colorization is highly challenging due to its multi-modality in the solution space and the requirement of global spatiotemporal consistency. First, it is not reasonable to recover the ground-truth color in various cases. For example, given a grayscale image of a balloon, we can not predict the correct color of the balloon because it may be yellow, blue, and so on. Instead of recovering the underlying color, we aim to generate a set of colorized results that look natural. Second, it often does not matter what color we assign to a region (i.e. a balloon), but the whole region should be spatially consistent. Third, video colorization is also inherently more challenging than single image colorization since temporal coherence should be also enforced. Image colorization methods usually do not generalize to video colorization. In Figure 1, we show some results of our approach and two state-of-the-art image colorization methods on classic film colorization.

Colorization of black-and-white images has been well studied in the literature [18, 6, 32, 16]. Colorization methods in the early days are mostly user-guided approaches that solve an objective function to propagate user input color scribbles to other regions [18, 25]. These approaches require users to provide sufficient scribbles on the grayscale image. On the other hand, researchers explore automatic image colorization with deep learning models. Some deep learning based approach for image colorization defines a classification based loss function with hundreds of discrete sampled points in chrominance space [32, 16]. However, the colorized image often exhibits evident discretization artifacts. To tackle this challenge, we suggest using a perceptual loss function combined with diversity. Our approach does not rely on sampling a discrete set of color in chrominance space and thus avoids discretization artifacts in the colorized video.

We may apply image colorization methods to colorize video frames independently, but the overall colorized video tends to be temporally inconsistent. Recently, Lai et al. [15] proposed a framework to enhance temporal coherence of a synthesized video where each frame is processed independently by an image processing algorithm such as colorization. However, this is a post-processing step and its performance is dependant on an image colorization approach that does not utilize multiple-frame information. Propagation-based video colorization methods require some colorized frames as reference to propagate the color of the given reference frames to the whole video [23, 29], but colorizing some frames also requires non-trivial human effort. Also, the quality of the colorized video frames decays quickly when the future frames are different from the reference frames. In this paper, we study the problem of automatic video colorization without both labeled data and user guidance.

We propose a self-regularized approach to automatic video colorization with diversity. We regularize our model with nearest neighbors in both bilateral and temporal spaces, and train the model with a diversity loss to dif-

Figure 1. Two colorized video frames by Zhang et al. [32], Iizuka et al. [12], and our approach on the classic film *Behind the Screen* in 1916 by Charlie Chaplin. State-of-the-art image colorization methods may not perform well on video colorization. The temporal inconsistency between the colorized video frames by Zhang et al. [32] and Iizuka et al. [12] is obvious. More results of classic film colorization are shown in the supplement.

ferentiate different modes in the solution space. The self-regularization encourages information propagation between pixels expected to have similar color. Specifically, we can build a graph with explicit pairwise connections between pixels by finding K nearest neighbors in some feature space or following the optical flow. By enforcing pairwise similarity between pixel pairs, we can preserve spatiotemporal color consistency in a video. Our model is also capable of generating multiple diverse colorized videos with a diversity loss [19]. We further suggest a simple strategy to select the most colorful video among all colorized videos.

We conduct experiments to compare our model with state-of-the-art image and video colorization approaches. The results demonstrate that our model can synthesize more natural colorized videos than other approaches do. We evaluate the performance on PSNR and LPIPS [33], and conduct perceptual comparison by a user study. Furthermore, controlled experiments show that our self-regularization and diversity are critical components in our model.

2. Related Work

In this section, we briefly review the related work in image and video colorization.

User-guided Image Colorization. The most classical approaches on image colorization are based on optimization that requires user input on part of the image to propagate the provided colors on certain regions to the whole image [18, 25, 22, 5, 31]. Levin et al. [18] propose optimization based interactive image colorization by solving a quadratic cost function under the assumption that similar pixels in space-time should have similar colors. Zhang et al. [34] present a deep learning based model for interactive image colorization.

Instead of requiring user scribbles, exemplar-based colorization approaches take a reference image as additional input [30, 13, 21, 3, 7, 10]. The reference image should be semantically similar to the input grayscale image to transfer the color from the reference image to the input image. A

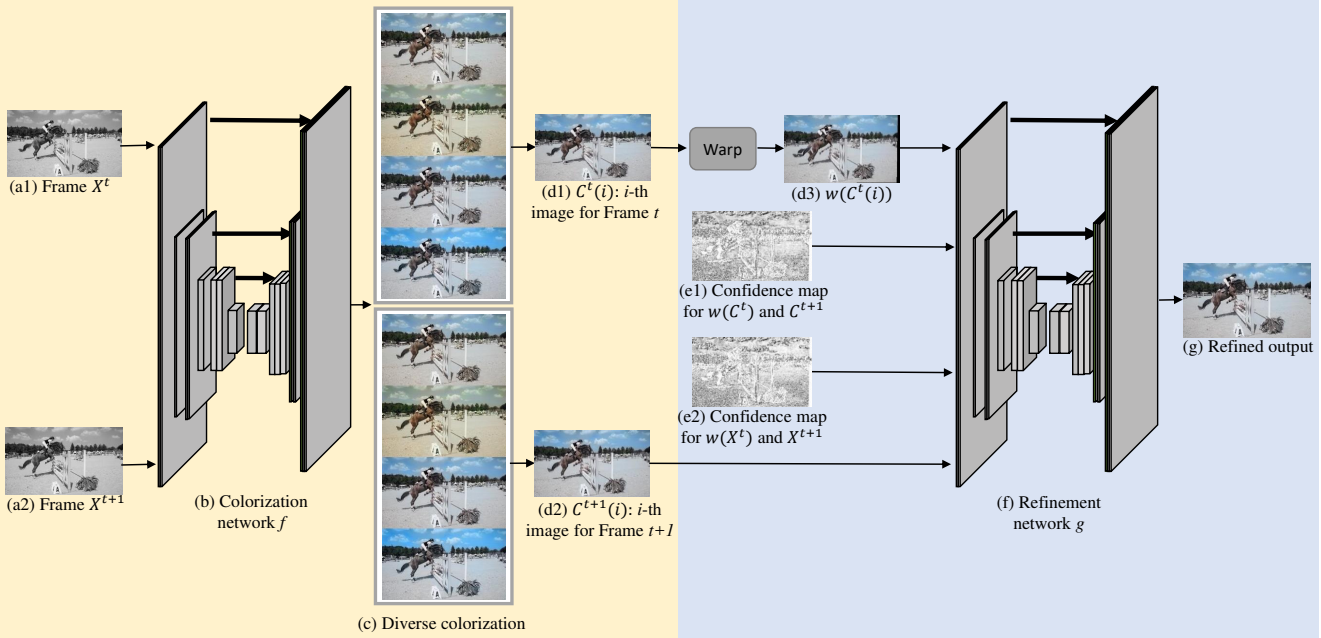

Figure 2. The overall architecture of our model. The colorization network f is designed to colorize each grayscale video frame, and produces multiple colorization candidate images. Taking i-th colorized candidate images from Frame t and Frame $t+1$ as well as two confidence maps, the refinement network g will output a refined video frame for Frame t.

recent approach by He et al. [11] combines deep learning and exemplars in image colorization and achieves the state-of-the-art performance. In this work, we are interested in fully automatic colorization approach that requires neither user input nor reference images.

Automatic Image Colorization. The most prominent work on fully automatic image colorization is deep learning based approaches that do not require any user guidance [6, 12, 32, 16, 9]. Cheng et al. [6] propose the first deep neural network model for fully automatic image colorization. Some deep learning approaches use a classification network that classifies each pixel into a set of hundreds of chrominance samples in a LAB or HSV color space to tackle to the multi-modal nature of the colorization problem [32, 16]. However, it is difficult to sample densely in the two-dimensional chrominance with hundreds of points. Thus we propose to use a perceptual loss with diversity [19] to avoid the discretization problem.

Video Colorization. Most contemporaneous work on video colorization is designed to propagate the color information from a color reference frame or sparse user scribbles to the whole video [31, 29, 23, 20, 14]. On the other hand, Lai et al. [15] propose an approach to enforce stronger temporal consistency of a video generated frame by frame by an image processing algorithm such as colorization. To the best of our knowledge, there are no deep learning models dedicated to fully automatic video colorization. We can definitely apply an image colorization method to colorize each frame in a video, but the resulted video is usually temporally incoherent. In this paper, we present a dedicated deep learning model for automatic video colorization that encourages spatiotemporal context propagation and is capable of generating a set of different colorized videos.

3. Overview

Consider a sequence of grayscale video frames $\mathbf{X} = \{X^1, \ldots, X^n\}$. Our objective is to train a model that automatically colorizes $\mathbf{X}$ such that the colorized video is realistic. In our framework, neither user guidance neither color reference frames are needed. Before we describe our approach, we characterize two desirable properties of our fully automatic video colorization approach.

- **Spatiotemporal color consistency.** Within a video frame, multiple pixels can share a similar color. For example, all the pixels on a wall should have the same color, and all the grass should be green. Establishing nonlocal pixel neighbors (i.e. two pixels on the same wall) for color consistency can improve the global color consistency of a colorized video. Note that colorizing video frames independently can result in a temporally inconsistent video, and thus we can establish temporal neighbors between two frames to enforce temporal coherence.

- **Diverse colorization.** Most existing work on image or video colorization only generates one colorization result. It is desirable for our model to output a set of diverse set of colorized videos, as colorization is a one-to-many problem. In our model, we use a perceptual loss with diversity to differentiate different modes in the solution space.

Figure 2 illustrates the overall structure of our model. Our proposed framework contains two networks that are trained to work in synergy. The first one is the colorization network $f(X^t;\theta_f)$ that outputs a colorized video frame given a grayscale video frame X^t. The network f is self-regularized with color similarity constraints defined on K nearest neighbors in the bilateral space $(r,g,b,\lambda x,\lambda y)$ where (r,g,b) represents the pixel color, (x,y) indicates the pixel location, and λ is a weight that balances the pixel color and location. We use $K=5$ in our experiments. The second one is the refinement network $g(C^s,C^t;\theta_g)$ designed to refine the current colorized video $\mathbf{C}$ by enforcing stronger temporal consistency. The network g propagates information between two nearby frames C^s and C^t. At the test time, g can be applied multiple times to the colorized video to achieve long-term consistency.

Furthermore, our approach can produce a diverse set of colorized videos, regularized by the diversity loss introduced by Li et al. [19]. We find that our diversity loss also stabilizes the temporal consistency of the colorized video. Combining the self-regularization and the diversity loss, we obtain the overall loss function to train our model:

$$L_{self}+L_{diversity}, \tag{1}$$

where L_{self} represents the loss to regularize color similarity between pixel neighbors in a bilateral space and a temporal domain, and $L_{diversity}$ is a perceptual loss function with diversity.

4. Self-Regularization

4.1. Self-regularization for colorization network

Consider colorizing a textureless balloon. Although it is nearly impossible to infer the underlying color of the balloon from a grayscale video frame, we somehow believe that all the pixels on the balloon are similar. We can find out pixel pairs expected to be similar, and enforce color similarity on these pixel pairs when training our model.

To establish pixel pairs with similar color in a video frame, we perform the K nearest neighbor (KNN) search in a bilateral space $(r,g,b,\lambda x,\lambda y)$ on the ground-truth frame during training. We expect that two pixels with similar color and spatial locations imply that our colorized video should also have a similar color for these two pixels. A similar KNN strategy is also presented in KNN matting [4]. Suppose $\mathbf{X}=\{X^1,\dots,X^n\}$ is the input grayscale video and $\mathbf{Y}=\{Y^1,\dots,Y^n\}$ is the ground-truth color video, our bilateral loss for self-regularization is

$$L_{bilateral}(\theta_f)=\sum_{i=1}^{n}\sum_{(p,q)\in\mathcal{N}_{Y^t}}\|f_p(X^t;\theta_f)-f_q(X^t;\theta_f)\|_1, \tag{2}$$

where $\mathcal{N}_{Y^t}$ is the KNN graph build on the ground-truth color frame Y_i, and $f_p(X^t;\theta_f)$ indicates the color of pixel p on the colorized video frame $f(X^t;\theta_f)$.

A simple temporal regularization term $L^f_{temporal}(\theta_f)$ can be defined on f:

$$\sum_{t=1}^{n-1}\|\left(f(X^t;\theta_f)-\omega_{t+1\to t}(f(X^{t+1};\theta_f))\right)\odot M_{t+1\to t}\|_1, \tag{3}$$

where $\omega_{i+1\to i}$ is an warping operator that warps an image from Frame $t+1$ to Frame t according to the optical flow from X^{t+1} to X^t. Given the optical flow $f_{t+1->t}$ from frame $t+1$ to frame t, we use backward warping to obtain a binary mask $M_{t+1->t}$ that indicates non-occluded pixels (invisible in Frame $t+1$).

4.2. Confidence-based refinement network

In our model, a confidence-based refinement network g is used to enforce stronger temporal consistency. Temporal inconsistency appears when corresponding pixels in two frames do not share similar colors. We use confidence maps to indicate whether the color of a pixel is inconsistent or inaccurate. Given a current colorized video $\mathbf{C}=\{C^1,\dots,C^n\}$, the temporal inconsistency when warping Frame t to Frame s can be translated into a confidence map with weights in the range of $[0,1]$:

$$W_{t\to s}(C^t,C^s)=\max(\mathbf{1}-\alpha|C^s-\omega_{t\to s}(C^t)|\odot M_{t\to s},\mathbf{0}), \tag{4}$$

where α is a hyper-parameter that controls the sensitivity of temporal inconsistency and we use $\alpha=15$.

Thus, for each colorized frame C^s, the refinement network g can use another nearby frame C^t along with the computed confidence maps to refine C^s. The input to g includes C^s, $\omega_{t\to s}(C^t)$, $W_{t\to s}(C^t,C^s)$, and $W_{t\to s}(X^t,X^s)$ that is the confidence map defined on the input grayscale image pairs. g outputs a refined video frame for C^s.

Training. To train the refinement network g, we sample two neighboring frames s and t such that $|s-t|\le\lambda$ where λ specifies the window size for temporal refinement. We find $\lambda=1$ is enough in our model. Then we optimize the following temporal regularization loss for θ_g:

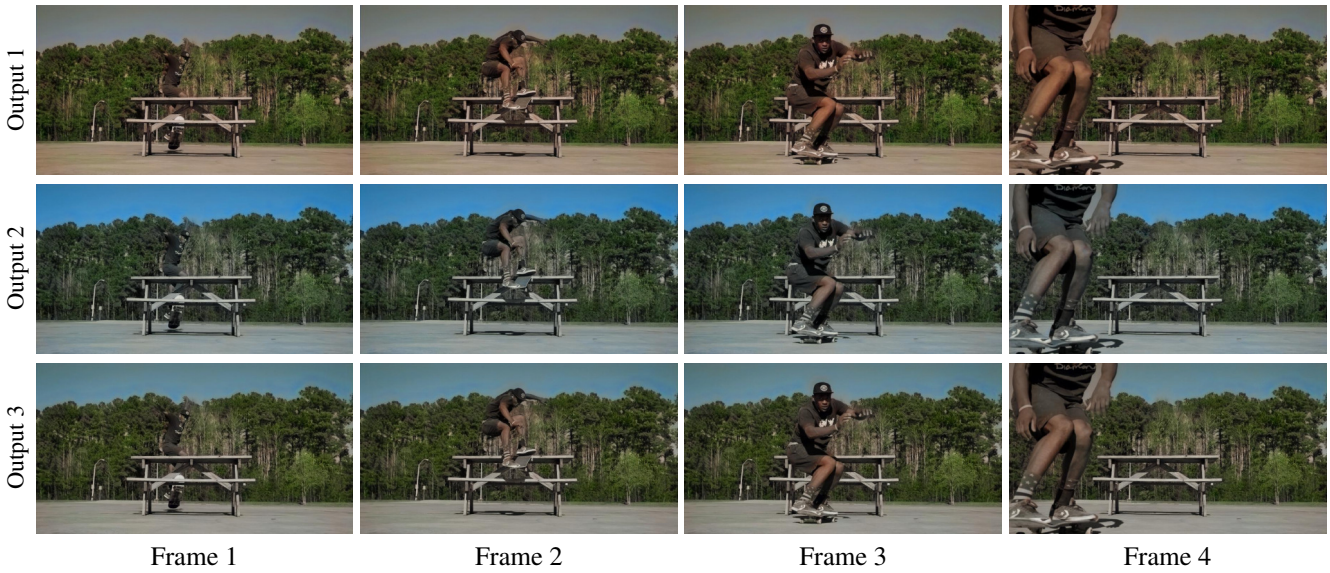

Figure 3. Four frames of three different videos colorized by our approach with diversity. Our approach is able to colorize videos in different ways. In general, different videos exhibit different global styles.

$$L^g_{temporal}(\theta_g) = \sum_{1\le|s-t|\le\lambda} \|g(f(X^s;\theta_f), f(X^t;\theta_f);\theta_g) - Y^s\|_1. \quad (5)$$

In summary, our self-regularization loss L_{self} is defined as

$$L_{bilateral}(\theta_f) + L^f_{temporal}(\theta_f) + L^g_{temporal}(\theta_g). \quad (6)$$

Inference. During the inference, we can apply g to refine each frame using the left λ frames and the right λ frames. If we perform this temporal refinement multiple times, we indirectly use the information from non-local frames to refine each frame.

5. Diverse Colorization

Video colorization is essentially a one-to-many task as there are multiple feasible colorized videos given the same grayscale input. Generating a diverse set of solutions can be an effective way to tackle this multi-modality challenge. Inspired by the ranked diversity loss proposed by Li et al. [19], we propose to generate multiple colorized videos to differentiate different solution modes. Besides, the diversity loss also contributes a lot to the temporal coherence because it reduces the ambiguity of colorization by generating several modes.

Suppose we generate d different solutions in our model. The network f should be modified to generate d images as output. The diversity loss imposed on f is,

$$L_{diversity}(\theta_f) = \sum_{t=1}^{n} \min_i \{\|\phi(C^t(i)) - \phi(Y^t)\|_1\} + \sum_{t=1}^{n}\sum_{i=1}^{d} \beta_i \|\phi(C^t(i)) - \phi(Y^t)\|_1, \quad (7)$$

where $C^t(i)$ is the i-th colorized image of $f(X^t;\theta_f)$ and $\{\beta_i\}$ is a decreasing sequence. We use $d = 4$ in our experiments.

The index of the best colorized video is not always the same. In most cases, we could empirically get a good index simply by choosing the one with the highest average per-pixel saturation where the saturation of a pixel is just the S channel in the HSV color space. Our method could also be an interactive method for users to pick the results they want.

In Figure 3, we show three colorized videos by our approach given the same grayscale input. In general, each video has its only style, and all the videos are different in both global color contrast and chrominance.

6. Implementation

We augment the input to the network f by adding hypercolumn features extracted from the VGG-19 network [27]. The hypercolumn features are expected to capture both low-level and high-level information of an image. In particular, we extract 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2' and 'conv5_2' from the VGG-19 network and upsample the layers by bilinear upsampling to match the resolution of the input image. The total number of channels of the hypercolumn feature is 1472. We adopt U-Net [26] as our network

Comparison	Preference rate DAVIS	Videvo
Ours > Zhang et al.[32] + BTC [15]	80.0%	88.8%
Ours > Iizuka et al. [12]+ BTC [15]	72.8%	63.3%

Table 1. The results of perceptual user study. Both baselines are enhanced with temporal consistency by BTC [15]. Our model consistently outperforms both state-of-the-art colorization methods by Zhang et al. [32] and Iizuka et al. [12].

Comparison	Preference rate DAVIS
Ours > Ours without self-reg.	67.9%
Ours > Ours without diversity	61.5%

Table 2. The results of the ablation study of comparisons between our full model and ablated models. The evaluation is performed by perceptual user study with 15 participants. The results indicate that self-regularization and diversity are key components in our model to achieve state-of-the-art performance in fully automatic video colorization.

structure for both networks f and g, and modify the architecture to fit our purpose. We add a 1×1 convolutional layer at the beginning of each network to reduce the dimensionality of the input augmented with hypercolumn features [19]. To compute the optical flow, we use the state-of-the-art method PWC-Net [28].

For model training, we first train the network f and then train g and f jointly. During each epoch for training f, we randomly sample 5,000 images in the ImageNet dataset [8] to train with loss of $L_{bilateral} + L_{diversity}$ and sample 1,000 pairs of neighboring frames in the DAVIS training set [24] by adding the temporal regularization for f, $L^f_{temporal}$. We train f for 200 epochs in total. Then for training the refinement network g, we randomly sample 1,000 pairs of frames from the DAVIS dataset in each epoch with the loss $L^g_{temporal}$. While there are d pairs of output from f with diversity, we train g on each pair of output. We also train our model in a coarse-to-fine fashion. We train both networks on the 256p videos and images. Then we fine-tune our model on the 480p videos and images.

7. Experiments

7.1. Experimental procedure

Datasets. We conduct our experiments mainly on the DAVIS dataset [24] and the Videvo dataset [2, 15]. The test set of the DAVIS dataset consists of 30 video clips of various scenes. There are about 30 to 100 frames in each video clip. The test set of the Videvo dataset contains 20 videos and each one has about 300 video frames. In totally, we evaluate our models and baselines on 50 test videos. All the videos are resized to 480p in both datasets.

Baselines. We compare our method with two state-of-the-art fully automatic image colorization approaches: the colorful image colorization (CIC) by Zhang et al. [32] and Iizuka et al. [12]. While these approaches are designed for image colorization, we apply their method to colorize video frame by frame. In addition, we apply the blind temporal consistency (BTC) method proposed by Lai et al. [15] improve the overall temporal consistency. Lai et al. [15] provided the results with temporal consistency for Zhang et al. [32] and Iizuka et al. [12]. We use publicly available pre-trained models and results of the baselines for evaluation. Their pre-trained models are trained on the DAVIS dataset [24] and the Videvo dataset [2, 15].

7.2. Results

Perceptual experiments. To evaluate the realism of the colorized video by each method, we conduct a perceptual experiment by user study. We compare our method with Zhang et al.[32] and Iizuka et al. [12] with enhanced temporal consistency by the blind temporal consistency (BTC) [15]. While our approach generates multiple videos, we choose the video with high saturation for evaluation.

In the user study, there are video comparisons between our approach and a baseline. In each comparison, a user is presented with a pair of colorized 480p videos side by side. The user can play both videos multiple times. We set the order of video pairs randomly and let the user choose the one that is more realistic and temporally coherent. Totally 10 users participated in this user study.

Table 1 summarizes the results of our perceptual experiment. Our method is consistently more rated preferable by most users. When our approach is compared with Zhang et al. [32], our approach is preferred in 80.0% of the comparisons on the DAVIS dataset and 88.8% of the comparisons on the Videvo dataset [2]. The perceptual user study is the key experiment to evaluate the performance of different methods.

Ablation study. Table 2 summarizes the ablation study by conducting perceptual user study on the DAVIS dataset. According to Table 2, our model without self-regularization or the diversity loss does not perform as well as our complete model. In summary, users rated our full model more realistic in 67.9% of the comparisons between our full model and the model without self-regularization and in 61.5% of the comparisons between our full model and the model without diversity.

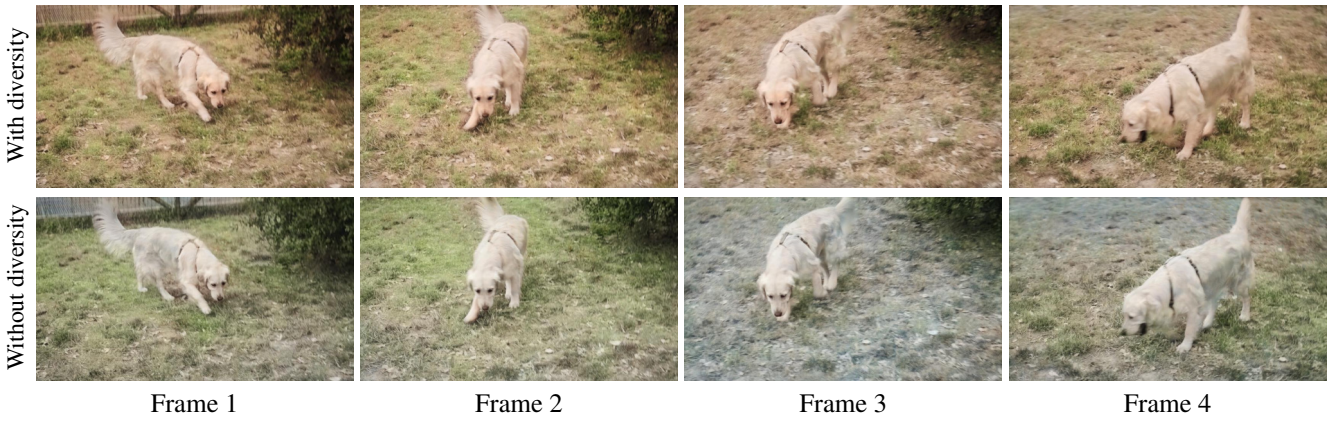

Figure 4. The visualization of the effect with and without the diversity loss. The first row shows four frames colorized by our full model, and the second shows four frames generated by our model without diversity. The diversity loss helps our model produce more temporally coherent and realistic results.

Figure 5. The visualization of the effect with and without the self-regularization. The self-regularization can help preserve global color consistency.

	DAVIS		Videvo	
Method	LPIPS	PSNR	LILPS	PSNR
Input	0.227	23.80	0.228	25.30
Zhang et al. [32]	0.218	29.25	0.201	29.52
Iizuka et al. [12]	**0.189**	29.91	**0.190**	30.23
Zhang et al. + BTC [15]	0.243	29.07	0.249	29.04
Iizuka et al + BTC [15]	0.218	29.25	0.241	28.90
Ours	0.191	**30.35**	0.194	**30.50**

Table 3. The results on two image similarity metrics, PSNR and LPIPS [33]. The blind temporal consistency (BTC) does not improve the results on these metrics. Image similarity metrics can not accurately measure the realism and temporal coherence of the colorized videos.

Qualitative results. Figure 4 and Figure 5 visualize the results of our full model and the ablated models without self-regularization or diversity.

In Figure 6 and Figure 7, we show the result videos colorized by our method and prior work. Our method produces more temporally consistent and more realistic colorized videos than state-of-the-art approaches do.

Image similarity metrics. We can use the image similarity metrics as a proxy to measure the similarity between the colorized video and the ground-truth video. Table 3 summarizes the results on image similarity metrics. Note that these metrics do not directly reflect the degree of realism of colorized videos. For example, a car may be colorized as blue or red. Both colors are plausible choices, but choosing a color different from the ground-truth video can results in huge errors on these image similirity metrics.

8. Discussion

We have presented our fully automatic video colorization model with self-regularization and diversity. Our colorized videos preserve global color consistency in both bilateral space and temporal space. By utilizing a diversity loss, our model is able to generate a diverse set of colorized videos that differentiate different modes in the solution space. We also find that our diversity loss stabilizes the training and process. Our work is an attempt to improve fully automatic video colorization but the results are still far from perfect. We hope our ideas of self-regularization and diversity can inspire more future work in fully automatic video colorization and other video processing tasks.

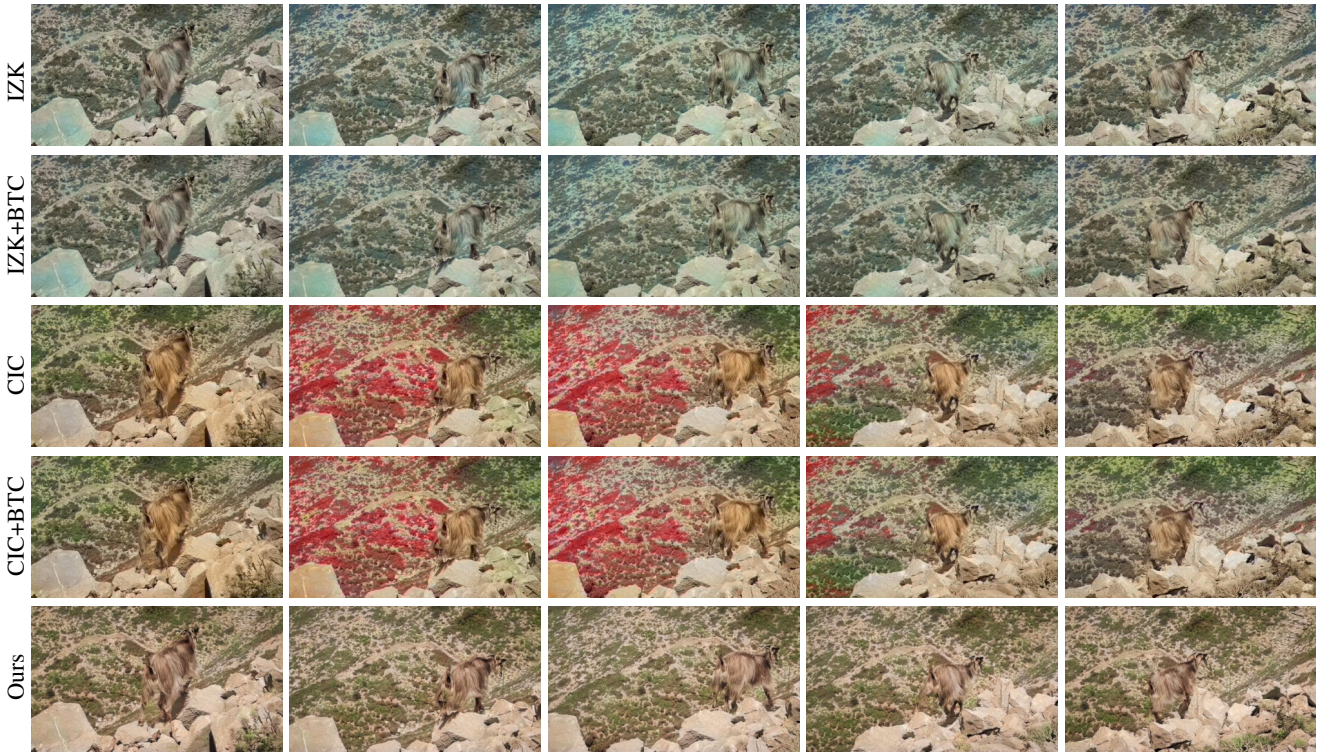

Figure 6. Qualitative results on the DAVIS dataset [24]. Here IZK refers to Iizuka et al. [12], CIC refers to the colorful image colorization method [32], and BTC refers to the blind temporal consistency method [15]. More results shown in the supplement.

Figure 7. Qualitative results on the Videvo dataset [2]. Here IZK refers to Iizuka et al. [12], CIC refers to the colorful image colorization method [32], and BTC refers to the blind temporal consistency method [15]. More results shown in the supplement.

References

[1] Highest-grossing film at the global box office (inflation-adjusted) — guinness world records. http://www.guinnessworldrecords.com/world-records/highest-box-office-film-gross-inflation-adjusted. 1

[2] Videvo. https://www.videvo.net/. 6, 8

[3] G. Charpiat, M. Hofmann, and B. Schölkopf. Automatic image colorization via multimodal predictions. In *ECCV*, 2008. 2

[4] Q. Chen, D. Li, and C. Tang. KNN matting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(9), 2013. 4

[5] X. Chen, D. Zou, Q. Zhao, and P. Tan. Manifold preserving edit propagation. *ACM Trans. Graph.*, 31(6), 2012. 2

[6] Z. Cheng, Q. Yang, and B. Sheng. Deep colorization. In *ICCV*, 2015. 1, 3

[7] A. Y. S. Chia, S. Zhuo, R. K. Gupta, Y. Tai, S. Cho, P. Tan, and S. Lin. Semantic colorization with internet images. *ACM Trans. Graph.*, 30(6), 2011. 2

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[9] A. Deshpande, J. Lu, M. Yeh, M. J. Chong, and D. A. Forsyth. Learning diverse image colorization. In *CVPR*, 2017. 3

[10] R. K. Gupta, A. Y. S. Chia, D. Rajan, E. S. Ng, and Z. Huang. Image colorization using similar images. In *Proceedings of the 20th ACM Multimedia Conference*, 2012. 2

[11] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan. Deep exemplar-based colorization. *ACM Trans. Graph.*, 37(4), 2018. 3

[12] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Trans. Graph.*, 35(4), 2016. 2, 3, 6, 7, 8

[13] R. Ironi, D. Cohen-Or, and D. Lischinski. Colorization by example. In *Proceedings of the Eurographics Symposium on Rendering Techniques*, 2005. 2

[14] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[15] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang. Learning blind video temporal consistency. In *ECCV*, 2018. 1, 3, 6, 7, 8

[16] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016. 1, 3

[17] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017. 1

[18] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *ACM Trans. Graph.*, 23(3), 2004. 1, 2

[19] Z. Li, Q. Chen, , and V. Koltun. Interactive image segmentation with latent diversity. In *CVPR*, 2018. 2, 3, 4, 5, 6

[20] S. Liu, G. Zhong, S. D. Mello, J. Gu, M. Yang, and J. Kautz. Switchable temporal propagation network. In *ECCV*, 2018. 3

[21] X. Liu, L. Wan, Y. Qu, T. Wong, S. Lin, C. Leung, and P. Heng. Intrinsic colorization. *ACM Trans. Graph.*, 27(5), 2008. 2

[22] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y. Xu, and H. Shum. Natural image colorization. In *Proceedings of the Eurographics Symposium on Rendering Techniques*, 2007. 2

[23] S. Meyer, V. Cornillère, A. Djelouah, C. Schroers, and M. H. Gross. Deep video color propagation. In *BMVC*, 2018. 1, 3

[24] F. Perazzi, J. Pont-Tuset, L. McWilliams, B.and Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 6, 8

[25] Y. Qu, T. Wong, and P. Heng. Manga colorization. *ACM Trans. Graph.*, 25(3), 2006. 1, 2

[26] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 5

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[28] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 6

[29] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy. Tracking emerges by colorizing videos. In *ECCV*, 2018. 1, 3

[30] T. Welsh, M. Ashikhmin, and K. Mueller. Transferring color to greyscale images. *ACM Trans. Graph.*, 21(3), 2002. 2

[31] L. Yatziv and G. Sapiro. Fast image and video colorization using chrominance blending. *IEEE Trans. Image Processing*, 15(5), 2006. 2, 3

[32] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *ECCV*, 2016. 1, 2, 3, 6, 7, 8

[33] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. 2018. 2, 7

[34] R. Zhang, J. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *ACM Trans. Graph.*, 36(4), 2017. 2