

BiSTNet: Semantic Image Prior Guided Bidirectional Temporal Feature Fusion for Deep Exemplar-based Video Colorization

Yixin Yang¹ Zhongzheng Peng¹ Xiaoyu Du¹ Zhulin Tao² Jinhui Tang¹ Jinshan Pan¹

¹Nanjing University of Science and Technology

² Communication University of China

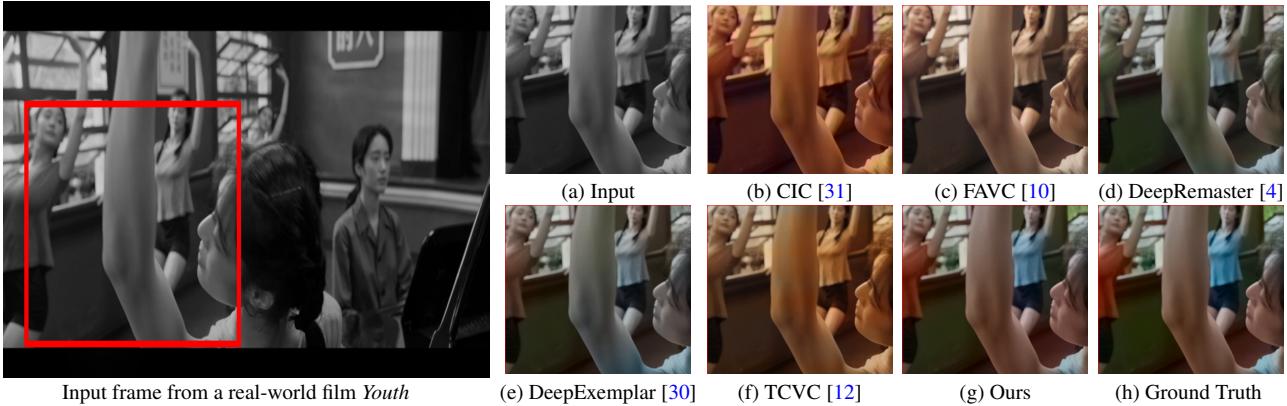


Figure 1. Colorization results in a real-world challenging film *Youth*. Our method explores well-colorized pixels from reference exemplars by a semantic correspondence and adaptively propagates the estimated colors by a bidirectional temporal feature fusion under the guidance of the semantic image prior to achieve better video colorization. As our analysis shows, reducing the influence of the inaccurately matched colors with the proposed bidirectional temporal feature fusion module and utilizing semantic information to guide the colorization process can obtain better-colorized results than both the exemplar-based methods [4, 30] and the automatic colorization methods [10, 12, 31] in terms of accuracy and temporal consistency.

Abstract

How to effectively explore the colors of reference exemplars and propagate them to colorize each frame is vital for exemplar-based video colorization. In this paper, we present an effective BiSTNet to explore colors of reference exemplars and utilize them to help video colorization by using a bidirectional temporal feature fusion with the guidance of semantic image prior. We first establish the semantic correspondence between each frame and the reference exemplars in deep feature space to explore color information from reference exemplars. Then, to better propagate the colors of reference exemplars into each frame and avoid the inaccurate matches colors from exemplars we develop a simple yet effective bidirectional temporal feature fusion module to better colorize each frame. We note that there usually exist color-bleeding artifacts around the boundaries of the important objects in videos. To overcome this problem, we further develop a mixed expert block to extract semantic information for modeling the object boundaries of frames so that the semantic image prior can better guide

the colorization process for better performance. In addition, we develop a multi-scale recurrent block to progressively colorize frames in a coarse-to-fine manner. Extensive experimental results demonstrate that the proposed BiST-Net performs favorably against state-of-the-art methods on the benchmark datasets. Our code will be made available at <https://yyang181.github.io/BiSTNet/>.

1. Introduction

Video colorization is a problem of generating fully colorized videos from their grayscale (monochrome) version. There are amounts of legacy black-and-white movies captured in the past decades due to the low quality of the film technology at the time, this problem has attracted lots of attention recently. Video colorization is an ill-posed problem as only monochrome videos are given. Compared to the single image colorization problem, the video colorization problem is more challenging as it not only needs to restore high-quality frames but also requires colorized videos with better temporal consistency.

The simplest way to solve video colorization is to process each frame independently using an image-based colorization model. Recently, amounts of single-image colorization methods are proposed and have achieved remarkable progress [2, 5, 9, 11, 22, 31, 33]. However, these methods often lead to temporal flickering artifacts and discontinuities. Thus, the demand for high-quality colorized videos, that look natural and are free of flickering artifacts, brings a huge challenge for existing video colorization methods.

To constrain the temporal consistency in video colorization tasks, several automatic colorization methods are proposed, including utilizing self-regularization and temporal constraint [10], conditional generative adversarial networks (GANs) [23] and bidirectional deep feature propagation [12]. These methods mainly focus on constraining temporal consistency while the colorization performance of each frame is far from satisfactory. Moreover, as video colorization is ill-posed, only depending on the training data does not provide enough information for colorization. Therefore, another category of approaches, exemplar-based colorization methods attract more attention recently.

Exemplar-based video colorization methods [4, 24, 28, 30] usually transfer the colors from a reference exemplar image to the grayscale one. Compared with automatic methods, exemplar-based methods use the color information of the reference exemplar to colorize every frame within the entire colorization period and are flexible enough to generate any kind of style according to the provided reference exemplar images. These methods usually work well on short video clips or simple scenes, but the visual performance of colorized videos is far from satisfactory when handling long or complicated real-world scenes. Because the semantic objects within a single reference image usually cannot cover all objects shown up in the video clips, causing wrong colors to be transferred from the reference exemplar image to the input frame. Thus, to design an effective video colorization network, it is essential to enhance the temporal consistency and maintain colorization performance for individual frames at the same time.

In this paper, we present an effective exemplar-based deep video colorization method. To better explore the colors of reference exemplars and propagate them to help colorize each frame, we first establish the semantic correspondence between each frame and the reference exemplars in a deep feature space. However, we note that the inaccurately matched colors would affect colorization. In contrast to existing exemplar-based video colorization methods that directly use the match colors to colorize each frame, we develop a simple yet effective bidirectional temporal feature fusion to avoid the influence of the possible inaccurate estimated colors and better propagate the colors extracted from reference exemplars. In addition, we note that most existing ones usually suffer from the color-bleeding arti-

fact, a problematic color spreading near the boundaries of some important objects in videos. To overcome this problem, we explore the semantic information of images to localize the boundaries of objects so that they can better guide the colorization for the regions of the object boundaries. We formulate the proposed method into a multi-scale recurrent framework that can progressively colorize frames in a coarse-to-fine manner. Both quantitative and qualitative evaluations show the effectiveness of the proposed method performs favorably against state-of-the-art ones (see Figure 1).

The main contributions of this work are summarized below:

- We present an effective method that explores both spatial information and temporal clues to keep the useful color information of provided exemplars for video colorization.
- We propose a bidirectional temporal fusion block (BTFB) that is capable of combining bidirectional semantic correspondence based on temporal clues. This module is effective to avoid abrupt flicking between adjacent frames and reduce the influence of inaccurate colors from the reference exemplars for better colorization.
- To ease color bleeding or color spilling effects, we develop a mixed expert block (MEB) to explore the semantic information and edge information as prior. By utilizing such image prior, MEB effectively guides the colorization of the regions around the object boundaries.
- We formulate the proposed method into a multi-scale recurrent convolutional block (MSRB) to progressively restore colorful video frames in a coarse-to-fine manner.

We extensively evaluate the proposed method on both synthetic datasets and real-world videos and demonstrate that it generates better-colorized videos in terms of accuracy and temporal consistency.

2. Related Work

Interactive colorization. In the early stage of colorization, local user hints are the most popular [1, 11, 13, 18, 29]. These methods require input from the user either in the form of points, strikes, or scribbles and rely on the assumption that nearby pixels should have similar colors. Levin et al. [11] propose an interactive colorization technique that propagates colors from scribbles to neighboring similar pixels. Sangkloy et al. [19] use an end-to-end feed-forward deep generative adversarial architecture to colorize images. To guide structural information and color patterns, user input in the form of sketches and color strokes is employed. Zhang et al. [33] develop user interaction based on two variants, local hint and global hint networks, both of which utilize a common main branch for image colorization. The non-trivial human effort and aesthetic skills to generate colorful

images required by user-guided methods make them unsuitable for video colorizing tasks.

Exemplar-based colorization. Exemplar-based colorization controls the color styles of the output frames by using reference frames. Deep exemplar-based colorization [3, 30] aims to provide diverse colors to the same image. The system is composed of two subnetworks: both the similarity sub-network and the colorization sub-network. Inspired by stylization characteristics in feature extracting and blending, Xu et al. [28] propose a stylization-based architecture for fast deep exemplar colorization to reduce high time and resource consumption. To colorize the images with multiple objects, Su et al. [22] propose an instance-aware image colorization method by utilizing an off-the-shelf pre-trained model to detect object instances and produce cropped object images. In those works, the correspondence and the color propagation are optimized independently, therefore visual artifacts tend to arise due to correspondence errors.

Video colorization. Video colorization [10, 16, 20, 28, 30] needs to consider both colorization performance and temporal consistency. Zhang et al. [30] unify the semantic correspondence and colorization into a single network and train it end-to-end to produce temporal consistent video colorization with realistic effects. Iizuka et al. propose DeepRemaster [4], a fully 3D convolutional method, utilizing source-reference attention to colorize long videos without the need for segmentation while maintaining temporal consistency. Recently, FAVC [10] is proposed for automatic video colorization. This method regularizes its model with a KNN graph built on the ground-truth color video and simultaneously posed a temporal loss term for constraining temporal consistency. FAVC focuses mainly on temporal consistency. Lai et al. [8] propose an approach to enforce stronger temporal consistency of a video generated frame by frame by an image processing algorithm such as colorization.

3. Proposed Method

Our main goal is to develop an effective model that explores both spatial and temporal information to generate colorful videos with better temporal consistency. To this end, we introduce key designs such as, a bidirectional temporal fusion block employed to avoid the influence of inaccurate color information from reference exemplars and alleviate flicking artifacts, a mixed expert block to extract semantic image prior to better guide the colorization of the regions of object boundaries and a multi-scale recurrent block that is capable of performing color mapping in coarse-to-fine mechanism than a single-scale network [28, 30], which benefits the single-frame colorization performance.

In this section, we first present the overall pipeline of our video colorization network architecture in detail. Then we describe the core components of the proposed blocks: (a) bidirectional temporal fusion block (b) mixed expert block

and (c) multi-scale recurrent block. Finally, we introduce the loss functions.

Overall pipeline. Given a grayscale video $X = \{x_0, x_1, \dots, x_{N-1}\}$, where $x_t \in \mathbb{R}^{H \times W \times 1}$ denotes the frame at time t , $H \times W$ denotes the spatial resolution, our main goal is to estimate the colorized video frames $Z = \{z_0, z_1, \dots, z_{N-1}\}$, where $z_t \in \mathbb{R}^{H \times W \times 3}$. We use two reference images $r^f \in \mathbb{R}^{H \times W \times 3}$ and $r^b \in \mathbb{R}^{H \times W \times 3}$ for each video clip. r^f and r^b contain similar contents to the input frames at time 0 and time N , respectively.

Specifically, we first apply the correspondence subnet that contains a forward stage and a backward stage to calculate bidirectional semantic correspondence given input grayscale image x_t . The forward semantic correspondence is to obtain the relationship between x_t and r^f and the backward semantic correspondence is to obtain the relationship between x_t and r^b . We establish these semantic correspondence using the deep features F_x^t, F_r^f and $F_r^b \in \mathbb{R}^{H \times W \times C}$ that are extracted from x_t, r^f and r^b by the pre-trained VGG19 model [21]. Then we reshape these features into feature vectors $\mathbf{F}_x^t, \mathbf{F}_r^f$ and $\mathbf{F}_r^b \in \mathbb{R}^{HW \times C}$. We estimate the correspondence matrix by:

$$\begin{aligned} \mathbf{C}_t^f &= \text{softmax} \left(\frac{\mathbf{F}_x^t \mathbf{F}_r^f^\top}{\|\mathbf{F}_x^t\|_2 \|\mathbf{F}_r^f\|_2} \right), \quad \mathbf{C}_t^f \in \mathbb{R}^{HW \times HW} \\ \mathbf{C}_t^b &= \text{softmax} \left(\frac{\mathbf{F}_x^t \mathbf{F}_r^b^\top}{\|\mathbf{F}_x^t\|_2 \|\mathbf{F}_r^b\|_2} \right), \quad \mathbf{C}_t^b \in \mathbb{R}^{HW \times HW} \end{aligned} \quad (1)$$

where $\text{softmax}(\cdot)$ denotes the softmax operation that is applied to each row of the matrix. We then generate the warped results of reference images r^f and r^b by:

$$\begin{aligned} \mathbf{w}_t^f &= \mathbf{C}_t^f \mathbf{r}^f, \\ \mathbf{w}_t^b &= \mathbf{C}_t^b \mathbf{r}^b, \end{aligned} \quad (2)$$

where \mathbf{r}^f and \mathbf{r}^b denote the vector forms of the reference images r^f and r^b . By applying a reshaping function to \mathbf{w}_t^f and \mathbf{w}_t^b , we can get the warped color frames $w_t^f \in \mathbb{R}^{H \times W \times 3}$ and $w_t^b \in \mathbb{R}^{H \times W \times 3}$. Taking these two warped frames as the input, the bidirectional temporal fusion block is developed to reduce influences of the inaccurate colors of w_t^f and w_t^b and alleviate ghost artifacts or color contamination when large motion happens or severe occlusion occurs (see details in Section 3.1).

We note that there usually exist color-bleeding artifacts around the boundaries of the important objects in videos. To overcome this problem and improve the quality of the colorized videos, we develop a color mapping subnet that contains a mixed expert block (MEB) and a multi-scale recurrent block (MSRB), see details in Section 3.2 and Section 3.3. This subnet obtains well-colorized image z_t by receiving two inputs: the grayscale input x_t and the fused color map p_t by the bidirectional temporal fusion block. In

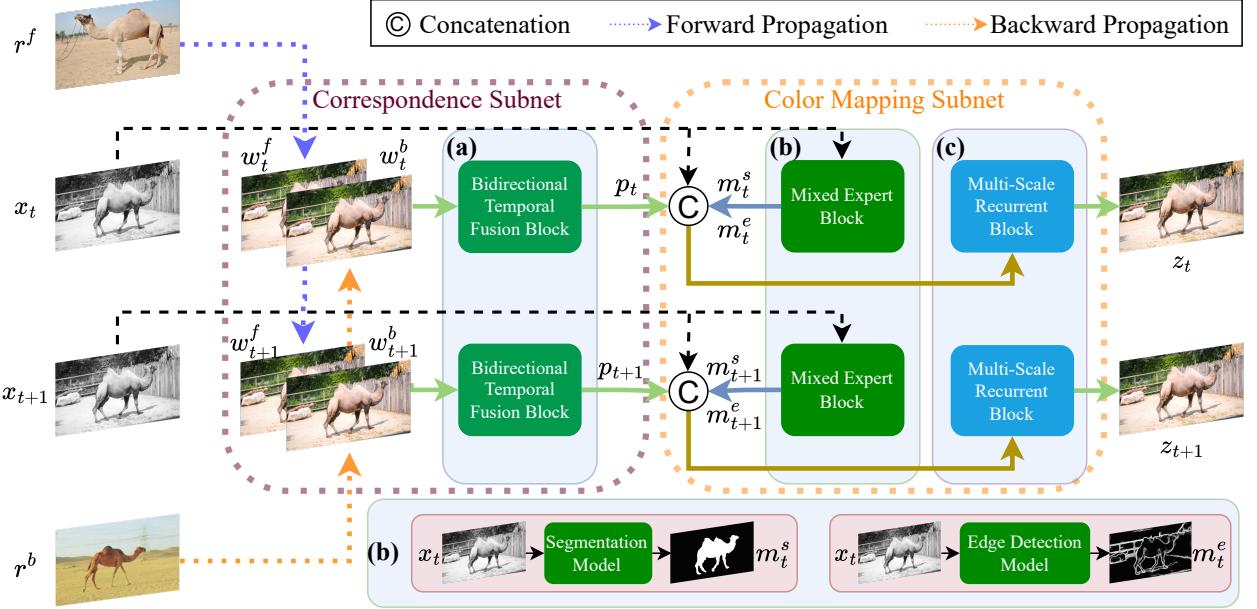


Figure 2. The architecture of BiSTNet for exemplar-based video colorization. The core components of our method include: (a) bidirectional temporal fusion block (BTFB), (b) mixed expert block (MEB) and (c) multi-scale recurrent block (MSRB).

the color mapping subnet, we first utilize MEB to extract the image prior from the given grayscale input x_t . MEB contains a semantic segmentation module that outputs segmentation mask m_t^s and an edge detection module that outputs edge detection mask m_t^e . We then propose MSRB to further enhance visual performance. Our MSRB is a three-level network based on Unet to obtain a well-colorized image z_t . It receives four inputs: the grayscale input x_t , the fused color map p_t by the bidirectional temporal fusion block, the segmentation mask m_t^s and the edge detection mask m_t^e . The pipeline for BiSTNet is shown in Figure 2.

3.1. Bidirectional Temporal Fusion Block

The quality of the colorized video frames often decays quickly when the future frames do not match the semantic contents of reference frames. This makes it a quite common scenario in exemplar-based video colorization tasks that the semantic correspondence is wrongly matched. Hence, it is necessary to utilize temporal clues to select well-colorized regions.

To this end, we develop a bidirectional temporal fusion block to obtain the fused color map p_t at time t based on the estimated warped color images w_t^f and w_t^b as detailed above. The fusion of w_t^f and w_t^b is mainly based on the temporal distance between the input frame x_t and the reference images r^f and r^b . The temporal distance d_t from input frame $x_t \in \{x_0, x_1, \dots, x_{N-1}\}$ to forward reference image r^f is defined as $d_t^f = t - 0$ and the temporal distance from input frame x_t to backward reference image r^b is defined as

$d_t^b = N - 1 - t$. We normalize these temporal distances based on the maximum distance $N - 1$ between the input frames, $\tilde{d}_t^f = \frac{t-0}{N-1}$ and $\tilde{d}_t^b = \frac{N-1-t}{N-1}$. Then we utilize \tilde{d}_t^f and \tilde{d}_t^b as weight parameters to obtain the fused color frame p_t . When the input frame x_t is closer to $r^f(r^b)$, the corresponding weights for $w_t^f(w_t^b)$ will be larger so that a more accurate fused color map p_t can be obtained. Based on such consideration, we can obtain the fused color map p_t by:

$$p_t = \frac{t-0}{N-1}w_t^f + \frac{N-1-t}{N-1}w_t^b. \quad (3)$$

3.2. Mixed Expert Block

Existing methods based on deep neural networks for video colorization often suffer from the color-bleeding artifact, a problematic color spreading near the boundaries between adjacent objects. Such color-bleeding artifacts affect the quality of generated videos, limiting the applicability of colorization models in practice. Therefore, it is necessary to explore prior knowledge from frames, which can better model the boundaries between adjacent objects to guide colorization.

In this paper, we develop a mixed expert block (MEB) to model the boundaries between adjacent objects. Our MEB contains two models, a semantic segmentation model and an edge detection model. MEB takes the grayscale frame x_t as input and outputs two masks, a probability mask of semantic segmentation (see m_t^s in Figure 2) and an edge detection mask (see m_t^e in Figure 2). To enhance visual performance near the object boundaries, we use an off-line

pre-trained semantic segmentation model [34] that takes the grayscale frame x_t as input to obtain $m_t^s \in \mathbb{R}^{H \times W \times C_{Seg}}$, where C_{Seg} is the number of semantic labels. In addition, we use an off-line pre-trained edge detection model [26] to obtain $m_t^e \in \mathbb{R}^{H \times W \times 1}$. We use the obtained m_t^s and m_t^e to better guide the colorization.

3.3. Multi-Scale Recurrent Block

To further improve the quality of the colorized videos, we employ a multi-scale recurrent block (MSRB) based on the commonly used coarse-to-fine strategy [14, 15]. MSRB contains three levels, each level is built by the same Unet structure (denoted as $\mathcal{N}, \mathcal{N}_2, \mathcal{N}_3$). It takes the concatenation results of the grayscale input x_t , the fused color map p_t , the semantic segmentation mask m_t^s and the edge detection mask m_t^e as the input. Specifically, let I_t denote the input of MSRB, we first downsample I_t to the 1/2 resolution feature $I_t^{\frac{1}{2}}$ and a 1/4 resolution feature $I_t^{\frac{1}{4}}$. Second, we apply a \mathcal{N}_3 to the feature $I_t^{\frac{1}{4}}$ and obtain a coarse colorized frame $z_t^{\frac{1}{4}}$. In this level, our target is to train \mathcal{N}_3 good at controlling the global color style. For the second level, the network \mathcal{N}_2 takes the concatenation of $I_t^{\frac{1}{2}}$ and upsampled colorized frame $z_t^{\frac{1}{4}}$ as input and generates the finer colorized frame $z_t^{\frac{1}{2}}$. Similarly, the network \mathcal{N} takes the concatenation of the original resolution feature I_t with the upsampled result of $z_t^{\frac{1}{2}}$ as input and restores the colorized frame z_t at the original resolution.

3.4. Loss Function

To better constrain the network training, we use the edge-enhancing loss by [6] to alleviate the color-bleeding artifacts and the problematic color spreading near the boundaries between adjacent objects:

$$\mathcal{L}_{edge} = \|\mathcal{S}(x_t) - \mathcal{S}(z_t)\|_2, \quad (4)$$

where \mathcal{S} is the Sobel filter used in [6].

We note that the majority of the colorized areas are smooth, artifacts usually happen at object boundaries. Simply using the L_1 -norm-based content aware loss function (i.e., $\|y_t - z_t\|_1$, where z_t denotes the colorized frame produced by our BiSTNet and y_t denotes the ground truth frame.) would make the trained colorization model mainly focus on well-colorized regions and ignore the regions where artifacts exist. To overcome this problem, we further employ the hard example mining loss by [15, 27].

In addition to those above loss functions, we further use the same loss function based on the content, perception and temporal consistency by [30] to constrain the network training. Finally, the loss function for the network training is:

$$\mathcal{L}_{total} = \lambda_{edge}\mathcal{L}_{edge} + \lambda_{hem}\mathcal{L}_{hem} + \lambda_c\mathcal{L}_c, \quad (5)$$

where \mathcal{L}_c is the loss function by [30]; \mathcal{L}_{hem} is the hard example mining loss function by [15]; λ_{edge} , λ_{hem} , and λ_c are weight parameters.

4. Experimental Results

In this section, we evaluate the proposed method against state-of-the-art ones. Due to the page limit, we include more results in the supplemental material. The training code and test model will be available to the public.

Datasets. We adopt the DAVIS dataset [17] and the Videvo dataset [8] as the benchmark datasets for training and testing. Our training set includes 60 clips from DAVIS and 80 clips from Videvo, whereas our testing set includes 30 clips from DAVIS. Before training, we resize each frame of all training videos into 384×224 pixels.

To obtain the reference color images, we adopt a similar strategy in [4]. We provide the first and the last frames as the reference images for each 90-frame length video clip. Then, we utilize these reference images to colorize the remaining part (the consecutive video frames from frame 6 to frame 85) of the video clips.

Evaluation metrics. To evaluate the quality of the colorized videos, we use the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM) [25], and the learned perceptual image patch similarity (LPIPS) [32] matrices. The color distribution consistency index (CDC) [12] is used to measure the temporal consistency of videos.

Implementation details. We train all models using the PyTorch framework on a machine with four RTX-A6000 GPUs. We adopt the CIE LAB color space for each frame in our experiments. In the training process, we use the Adam algorithm [7] with default parameters and train our models with 50 epochs. The batch size is set to 8. The learning rate is set to a constant 2×10^{-4} . We employ the off-the-shelf method ProtoSeg [34] for semantic segmentation. We also employ HED [26] for edge detection with default parameters. For the weights in the loss function, we empirically set $\lambda_{edge} = 2$, $\lambda_{hem} = 2$, $\lambda_c = 1$.

4.1. Comparisons with State-of-the-Art Methods

We compare the proposed method with state-of-the-art ones including both the automatic colorization ones [10, 12] and exemplar-based video colorization ones [4, 30]. In addition, to make the paper more self-contained, we include comparisons with state-of-the-art image colorization methods [5, 22, 31, 33]. We note that exemplar-based methods [4, 30] require a reference exemplar as guidance. For fair comparisons, we choose the first ground truth color frame as the reference exemplar for these methods.

Table 1 shows the quantitative evaluation results on the commonly used benchmark [17], where the proposed BiSTNet generates better-colorized results. In particular,

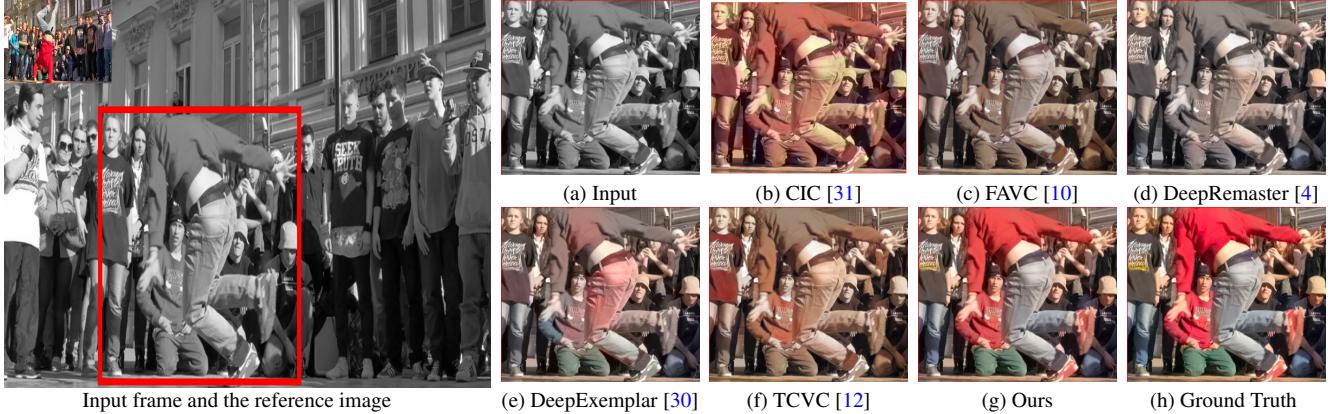


Figure 3. Qualitative colorization comparisons on clip *breakdance* from the DAVIS dataset [17]. The evaluated methods do not generate colorful frames, and the colors by these methods are not estimated correctly in (b)-(f). Our method generates well-colorized image.

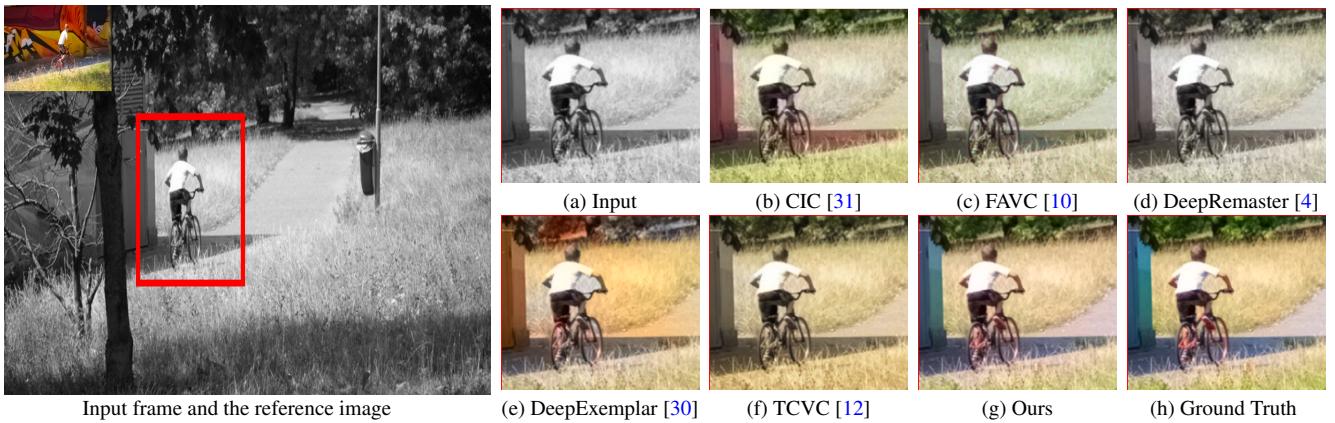


Figure 4. Qualitative colorization comparisons on clip *bmx-trees* from the DAVIS dataset [17]. Our method in (g) is able to avoid color contamination near the boundaries of lawns and roads.

BiSTNet outperforms DeepExemplar [30], a state-of-the-art exemplar-based method, by up to 1.37 dB in terms of PSNR, and achieves 18.52% improvement in terms of the temporal consistency index CDC. Compared to the TCVC method [12] which mainly focuses on improving the temporal consistency of the colorized videos, our BiSTNet still shows an improvement of 10.9% in terms of CDC. All comparisons in Table 1 demonstrate that the proposed BiSTNet generates better-colored frames while the generated videos have better temporal consistency.

Figure 3 and Figure 4 show qualitative comparisons of the proposed BiSTNet and the methods with top performance in Table 1. We note that the DeepExemplar method [30] and the DeepRemaster method [4] do not generate the results with the correct colors. In contrast, the proposed BiSTNet generates vivid frames whose colors are visually close to the ground truths. In particular, the proposed BiSTNet restores the color of the cloth in Figure 3. In addition, the proposed BiSTNet can avoid color contamination near the boundaries of lawns and roads in Figure 4,

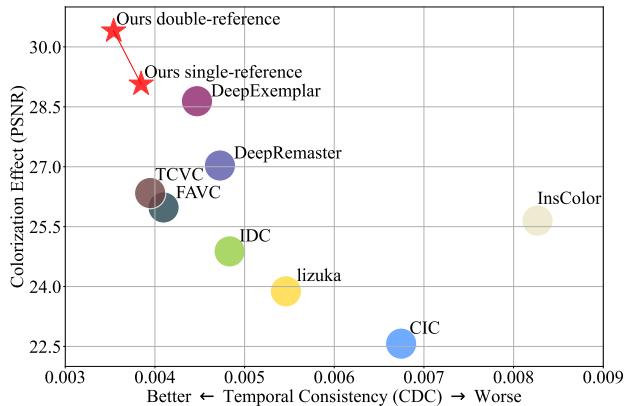


Figure 5. Model performance and temporal consistency comparison between our proposed BiSTNet family and other colorization methods on the DAVIS dataset [17]. Our BiSTNet family achieves both a higher PSNR and a lower CDC among all other methods.

suggesting the effectiveness of the proposed MEB on video colorization tasks.

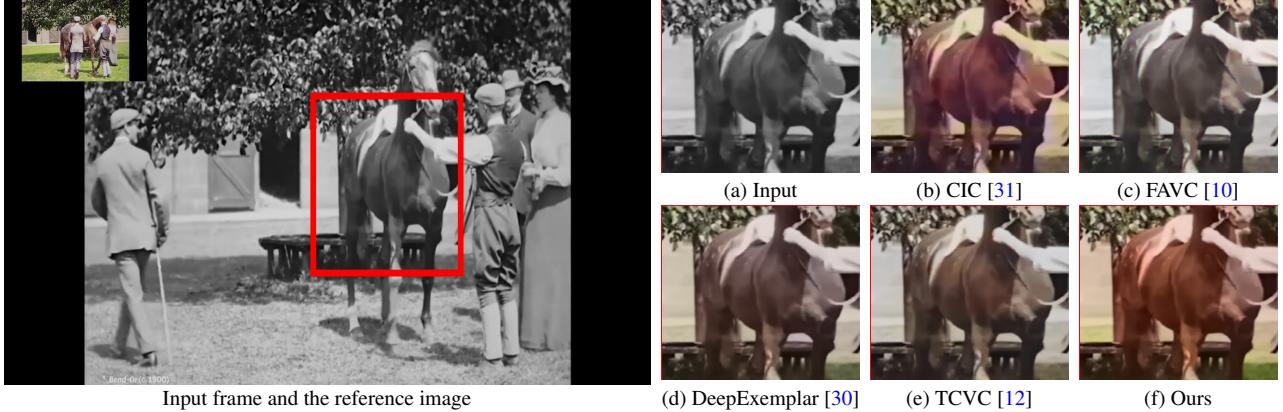


Figure 6. Qualitative colorization comparisons on real-world film *Bend-Or*. The colorization methods [10, 12, 30] do not colorize the horse and the backgrounds well. Our method colorizes the frame well, where the colors look better.

4.2. Evaluations on Real-World Grayscale Videos

We further evaluate the proposed method on real-world grayscale videos, where the ground truth colors videos are not available. Figure 6 shows an example from a challenging real-world old film *Bend-Or*. We select the well-colorized frames from the internet as the reference exemplar for all the exemplar-based methods and compare the proposed method with state-of-the-art ones [4, 10, 12, 30, 31]. Figure 6 shows that state-of-the-art methods do not colorize the objects (e.g., horse) well. In contrast, our method generates a better-colorized frame, where the colors of the horse look natural (see Figure 6(g)).

4.3. Temporal Consistency Evaluations

Existing methods mainly focus on improving the quality of each colorized frame. As temporal consistency is one of the important factors that determines the quality of colorized videos, we further demonstrate whether colorized videos by the proposed method have a better temporal consistency or not. We plot the relationship between the quality of each colorized frame and the temporal consistency in Figure 5.

Overall, the proposed method generates high-quality colorized videos with better temporal consistency.

5. Analysis and Discussion

In this section, we further conduct extensive ablation studies to demonstrate the effectiveness of the proposed temporal fusion model, mixed expert block, and multi-scale recurrent block. We train all the possible baseline models using the same settings as the proposed BiSTNet and evaluate them on the DAVIS dataset [17] for fair comparisons. Table 2 shows the quantitative evaluation results.

Effectiveness of the bidirectional temporal fusion block. The proposed BTFB is used to utilize temporal clues to select well-colorized regions from the reference exemplars.

Table 1. Quantitative comparisons of the proposed method against state-of-the-art ones on the synthetic DAVIS dataset [17]. Our method achieves the best performance in terms of PSNR, SSIM, LPIPS, and CDC.

Method	PSNR↑	SSIM↑	LPIPS↓	CDC↓
Input	19.84	0.953	0.230	/
Iizuka et al [5]	23.88	0.947	0.176	0.005461
CIC [31]	22.57	0.947	0.106	0.006746
IDC [33]	24.88	0.949	0.116	0.004833
InsColor [22]	25.65	0.951	0.082	0.008267
FAVC [10]	25.98	0.967	0.172	0.004096
TCVC [12]	26.34	0.962	0.175	0.003947
DeepRemaster [4]	27.03	0.964	0.057	0.004725
DeepExemplar [30]	28.64	0.972	0.041	0.004469
Ours single-reference	29.07	0.973	0.036	0.003846
Ours double-reference	30.40	0.976	0.030	0.003540

Table 2. Quantitative evaluations of each component in the proposed BiSTNet on the DAVIS dataset [17].

Method	PSNR↑	CDC↓
Ours w/o BTFB	28.32	0.004359
Ours w/o m_t^s	27.58	0.004484
Ours w/o m_t^e	27.99	0.004078
Ours w/o m_t^e, m_t^s	26.61	0.003972
Ours w/o MSRB	27.29	0.004640
Ours w/o \mathcal{L}_{edge}	28.15	0.004660
Ours w/o \mathcal{L}_{hem}	27.08	0.004525
Ours w/o $\mathcal{L}_{edge} \& \mathcal{L}_{hem}$	27.00	0.004682
Ours	30.40	0.003540

To demonstrate its effectiveness, we replace the BTFB with average weighting to fuse the bidirectional features. Table 2 shows that our model without TFB does not perform well.

Effectiveness of the mixed expert block. The MEB is used to explore image prior such as semantic information and object edges to guide the colorization of areas at the boundaries between adjacent objects. To demonstrate the effectiveness of MEB, we compare with the baselines by replacing the segmentation mask m_t^s and the edge mask m_t^e

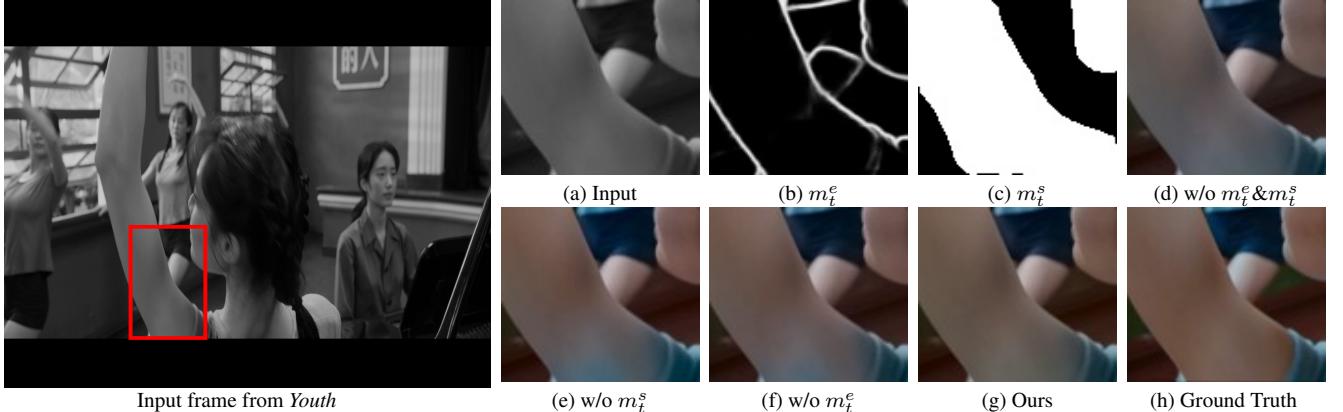


Figure 7. Effectiveness of the proposed MEB module. (b) and (c) denote the edge detection mask m_t^e and the semantic segmentation mask m_t^s . (d) denotes the colorized image by the proposed model without using both m_t^e and m_t^s (i.e., the proposed method without using the MEB module). (e) denotes the colorized image by the proposed model without using m_t^s . (f) denotes the colorized image by the proposed model without using m_t^e . Note that the method without using the MEB module does not colorize the frame well, where the color of the arm is contaminated by the color of the T-shirt. Our model with the MEB module restores better colors in (g).



Figure 8. Effectiveness of the proposed MEB module. As shown in (b), the result by the method without using the MSRB contains uneven color distribution from left to right. The method using the MSRB restores the better-colorized frame in (c).

one by one. We also conduct an experiment to remove both m_t^s and m_t^e . Table 2 shows that the colorization model without semantic image prior does not colorize the videos well. Figure 7 shows the effectiveness of the proposed MEB visually. We note that the BiSTNet without using the m_t^s or the m_t^e does not colorize the frame well where the results contain significant color-bleeding artifacts. For example, the colors of the T-shirt spread to the arm. In constraint, using the proposed MEB is able to reduce this phenomenon and generates a better-colorized frame (Figure 7(g)).

Effectiveness of the multi-scale recurrent block. The MSRB is based on the coarse-to-fine strategy that is widely used in image and video restoration [14, 15]. However, its effect is not clear in video colorization. To validate its effectiveness on video colorization, we compare the proposed method without this block. The comparisons of “Ours w/o MSRB” and “Ours” in Table 2 show that using the MSRB is able to improve the performance of video colorization. Specifically, as shown in Figure 8, the global coloring effect is not good and the color of the fence is contaminated by the color of the roof around the right area of the image. The model with the proposed MSRB generates a better-colorized frame (Figure 8(c)).

Effectiveness of the edge loss and hard example mining

loss. We further examine the effect of the edge loss and the hard example mining on video colorization. The comparisons of the results in Table 2 (see “Ours w/o \mathcal{L}_{edge} ”, “Ours w/o \mathcal{L}_{hem} ”, “Ours w/o $\mathcal{L}_{edge} \& \mathcal{L}_{hem}$ ”, and “Ours”) show that using these two loss functions helps video colorization.

Two or single reference exemplars. As our method uses two reference frames, one may wonder whether the proposed BiSTNet works well if only a single reference frame is used. To answer this question, we build a baseline model that only uses the first color frame of a video (*i.e.*, the frame at time 0) and train this baseline model using the same settings as the proposed method for fair comparisons. Table 1 shows that although using a single reference does not generate better results compared to the method using double reference frames, it still achieves better performance than the state-of-the-art methods.

Limitations and future work. Although the proposed method achieves favorable performance on video colorization, there are still some limitations. As the colors of the old videos are unknown and subjective, only exploring the colors of several exemplars are not enough to colorize videos. How to explore the domain knowledge of the videos and establish their relations with the well-known knowledge for the colors of scenarios would provide more convincing color information for video colorization.

6. Conclusion

In this paper, we propose an effective BiSTNet to better explore and propagate colors from reference exemplars for video colorization. We first establish the semantic correspondence between each frame and the reference exemplars in a deep feature space and develop a simple yet effective bidirectional temporal fusion block to better propagate

the colors of reference exemplars and avoid the inaccurately matched colors from exemplars. We develop a mixed expert block to guide the colorization of the regions around object boundaries. By progressively colorizing frames in a coarse-to-fine manner, we show that the proposed BiSTNet performs favorably against state-of-the-art methods on the benchmark datasets.

References

- [1] Xiaowu Chen, Dongqing Zou, Qinping Zhao, and Ping Tan. Manifold preserving edit propagation. *ACM TOG*, 31(6):1–7, 2012. [2](#)
- [2] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *ICCV*, 2015. [2](#)
- [3] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM TOG*, 37(4):1–16, 2018. [3](#)
- [4] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM TOG*, 38(6):1–13, 2019. [1, 2, 3, 5, 6, 7](#)
- [5] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM TOG*, 35(4):1–11, 2016. [2, 5, 7](#)
- [6] Eungyeup Kim, Sanghyeon Lee, Jeonghoon Park, Somi Choi, Choonghyun Seo, and Jaegul Choo. Deep edge-aware interactive colorization against color-bleeding effects. In *ICCV*, 2021. [5](#)
- [7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [5](#)
- [8] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018. [3, 5](#)
- [9] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016. [2](#)
- [10] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *CVPR*, 2019. [1, 2, 3, 5, 6, 7](#)
- [11] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. *ACM TOG*, 23(3):689–694, 2004. [2](#)
- [12] Yihao Liu, Hengyuan Zhao, Kelvin CK Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong. Temporally consistent video colorization with deep feature propagation and self-regularization learning. *arXiv preprint arXiv:2110.04562*, 2021. [1, 2, 5, 6, 7](#)
- [13] Qing Luan, Fang Wen, Daniel Cohen-Or, Lin Liang, Ying-Qing Xu, and Heung-Yeung Shum. Natural image colorization. In *ESRT*, 2007. [2](#)
- [14] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. [5, 8](#)
- [15] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *CVPR*, 2020. [5, 8](#)
- [16] Somdyuti Paul, Saumik Bhattacharya, and Sumana Gupta. Spatiotemporal colorization of video using 3d steerable pyramids. *IEEE TCSVT*, 27(8):1605–1619, 2016. [3](#)
- [17] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. [5, 6, 7](#)
- [18] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. Manga colorization. *ACM TOG*, 25(3):1214–1220, 2006. [2](#)
- [19] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*, 2017. [2](#)
- [20] Bin Sheng, Hanqiu Sun, Marcus Magnor, and Ping Li. Video colorization using parallel optimization in feature space. *IEEE TCSVT*, 24(3):407–417, 2013. [3](#)
- [21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. [3](#)
- [22] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *CVPR*, 2020. [2, 3, 5, 7](#)
- [23] Harrish Thasarathan, Kamyar Nazeri, and Mehran Ebrahimi. Automatic temporally coherent video colorization. In *CRV*, 2019. [2](#)
- [24] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *CVPR*, 2022. [2](#)
- [25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. [5](#)
- [26] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. [5](#)
- [27] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *CVPR*, 2019. [5](#)
- [28] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *CVPR*, 2020. [2, 3](#)
- [29] Liron Yatziv and Guillermo Sapiro. Fast image and video colorization using chrominance blending. *IEEE TIP*, 15(5):1120–1129, 2006. [2](#)
- [30] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *CVPR*, 2019. [1, 2, 3, 5, 6, 7](#)
- [31] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016. [1, 2, 5, 6, 7](#)
- [32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [5](#)
- [33] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *ACM TOG*, 36(4):1–11, 2017. [2, 5, 7](#)
- [34] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, 2022. [5](#)