Temporal Consistent Automatic Video Colorization via Semantic Correspondence

Yu Zhang[1], Siqi Chen[1]*, Mingdao Wang[1], Xianlin Zhang[2], Chuang Zhu[1], Yue Zhang[2], Xueming Li[2]
[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications
Beijing, China
{zhangyu_03, sqchen, wmingdao, zxlin, czhu}@bupt.edu.cn,
zhangyuereal@163.com, lixm@bupt.edu.cn

Abstract

Video colorization task has recently attracted wide attention. Recent methods mainly work on the temporal consistency in adjacent frames or frames with small interval. However, it still faces severe challenge of the inconsistency between frames with large interval. To address this issue, we propose a novel video colorization framework, which combines semantic correspondence into automatic video colorization to keep long-range consistency. Firstly, a reference colorization network is designed to automatically colorize the first frame of each video, obtaining a reference image to supervise the following whole colorization process. Such automatically colorized reference image can not only avoid labor-intensive and time-consuming manual selection, but also enhance the similarity between reference and grayscale images. Afterwards, a semantic correspondence network and an image colorization network are introduced to colorize a series of the remaining frames with the help of the reference. Each frame is supervised by both the reference image and the immediately colorized preceding frame to improve both short-range and long-range temporal consistency. Extensive experiments demonstrate that our method outperforms other methods in maintaining temporal consistency both qualitatively and quantitatively. In the NTIRE 2023 Video Colorization Challenge, our method ranks at the 3rd place in Color Distribution Consistency (CDC) Optimization track. Code will be available online at https://github.com/bupt-ai-cz/TCVC

1. Introduction

As a well-known ill-posed problem, video colorization task owns serious ambiguity that a grayscale object could be plausible in various colors. This characteristic usually results in temporal inconsistency that the colors of an object may change in different frames. In order to resolve such inconsistency, mainly three kinds of methods are proposed: task-independent, fully-automatic and exemplar-based methods.

The task-independent [1, 2, 3, 4] methods aim to enhance temporal consistency between image colorization results via post-processing. They formulate a temporal filter and punish the warping errors computed by optical flow between adjacent frames. However, their results are still not consistent enough when the generated colors are extremely different in adjacent frames, and they have to process each frame twice. Based on the conception of task-independent methods, automatic colorization methods [5, 6, 7, 8] are proposed. They directly map the feature embedding of grayscale images to their color representations by learning from large datasets. For instance, Lei et al. [6] divide the video colorization into a single frame colorization subnet and a smoothing subnet. However, it is difficult to generate colorful results. To integrate both image and video colorization, Zhao et al. [5] propose an end-to-end network using two step training, and introduce a dense long-term loss to minimize flickers of generated frames. However, the long-term loss only covers few frames and is dependent on the quality of optical flow. For long videos, it still suffers temporal inconsistency in wide frame interval. Fig. 4 illustrates the different colorization strategy.

The exemplar-based methods utilize a colorized reference image to supervise the colorization process for all frames [9, 10, 11, 12, 13, 14]. A semantic correspondence network is usually adopted to find the pixel-wise correspondence between reference and grayscale images. For example, Zhang et al. [13] propose a recurrent network where the non-local operation [15] is responsible to find semantic correspondence between reference and grayscale images,

*Corresponding author

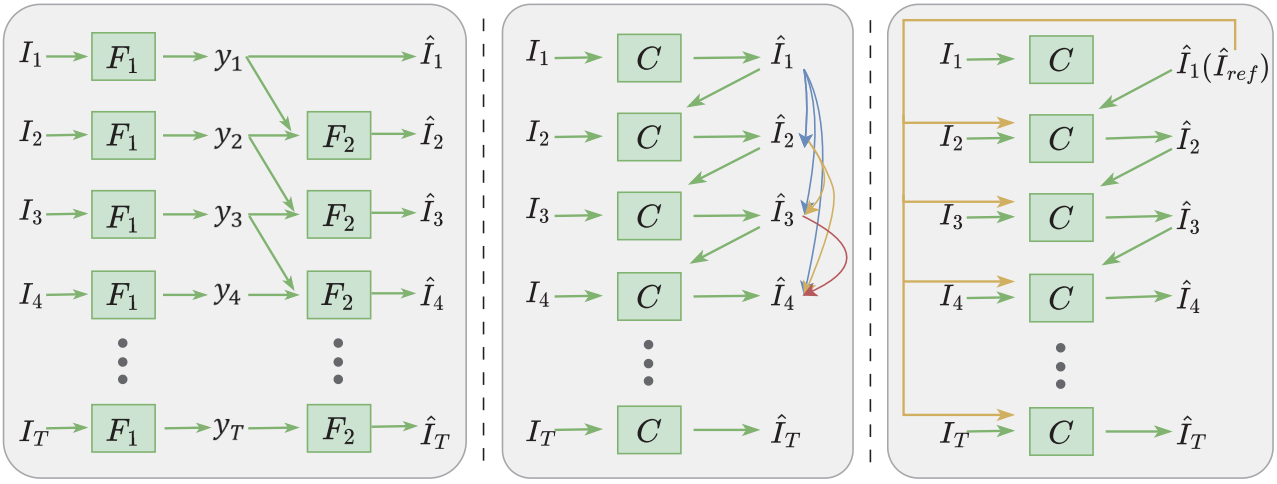

Figure 1. Comparison of different frameworks in video colorization: (a) Colorization with post-processing [1, 2, 3, 4], (b) Colorization with dense long-term loss [5], (c) Colorization with semantic correspondence (Ours), where I is the input, r is the reference image, $\hat{I}$ is the video colorization result and y is the image colorization result. F_1, F_2, C represent CNNs, and the color lines in (b) indicate dense long-term loss.

and the previous colored frame is also leveraged to increase temporal consistency. To further enhance spatiotemporal long-term dependency in videos, Chen et al. [16] propose a double-head non-local operation and an attention [17] based linkage subnet to improve the representation ability. However, the behavior of exemplar-based method is highly dependent on the selection of reference image, and the manual selection of reference is usually experience required and time-consuming.

Under this circumstance, this paper proposes Temporal Consistent Automatic Video Colorization with Semantic Correspondence, which combines semantic correspondence network into automatic video colorization. As difficult for automatic methods to keep long-range consistency, a reference image together with semantic correspondence network is leveraged to supervise the whole colorization process; and as complicated to manually select the reference image, a prior reference colorization network is leveraged to generate the reference image by automatically coloring the first frame in video. Such direct colorization of reference can not only avoid manual selection, but also increase the similarity of reference and grayscale images (since they belongs to the same video), which is beneficial for semantic correspondence. Our contributions can be summarized as:

- A novel framework combines automatic video colorization with semantic correspondence is proposed to keep long-range consistency.
- We leverage an automatically generated reference image to supervise the colorization of remaining frames. Each frame is supervised by both the reference image and the immediately colorized preceding frame.
- Experiments demonstrate that our method can better maintain temporal consistency, and outperforms recent state-of-the-arts both qualitatively and quantitatively.

2. Related Works

In this section, we will introduce the two main methods in video colorization: automatic and exemplar-based.

2.1. Automatic Colorization

Automatic methods [5, 6, 7, 8] are proposed to further optimise temporal coherence. They map grayscale images directly to color embedding using deep neural networks, while maintaining frame continuity. Lei et al. [6] propose a multi-modal automatic framework that can generate four diverse colorization results simultaneously. To maintain spatio-temporal consistency, they impose similarity between pixel pairs by K Nearest Neighbor (KNN) search in feature space or by optical flows. Zhao et al. [5] propose a hybrid recurrent network that integrates both image and video colorization and meanwhile leverage a dense long-term loss which considers not only adjacent but long-term continuity to optimize it. Nevertheless, it is as yet difficult to generate a colorful result with the help of these methods. Especially in practical applications like old movie restoration, there are certain colors in specific scenarios for objects such as clothes, skin, house, which have historical basis and are difficult to generate by fully-automatic approaches.

2.2. Exemplar-based Colorization

Exemplar-based methods generally utilize one or more colored frames in a video as reference images to guide

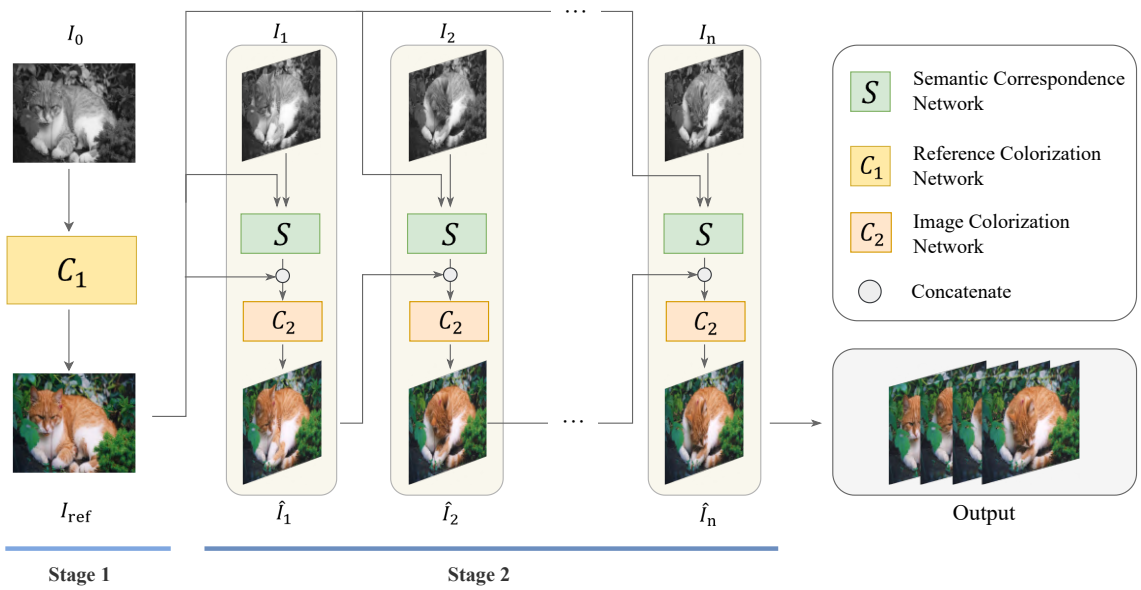

Figure 2. The overall framework of our method. There are mainly three components: a reference colorization network, an image colorization network and a semantic correspondence network. The reference colorization network generate a colorized reference image using the first grayscale frame of the video. The semantic correspondence network and the image colorization network then leverage the reference to supervise the whole colorization process.

the colorization process. These method [18, 19, 20] leverage handcrafted low-level features to find temporal correspondence between frames and colorize following frames in sequence. More recent methods tend to use deep neural networks to achieve temporal propagation [9, 10, 11]. While these approaches produce much more colorful results, their coloration depends only on the previous frame, which makes it easy to accumulate color errors as they propagate. Another type of method involves the reference images throughout the process [12, 13, 14], providing more stable results. For instance, Zhang et al. [13] propose a recurrent framework with novel loss functions where colorization depends on both the reference and the previous frames. Iizuka et al. [12] first propose a single framework for remastering vintage films. They adopt a source-reference attention that can handle multiple references, and utilize 3D-CNN for modeling temporal correspondence. Although favorable results are obtained, these approaches nonetheless lack long-term spatio temporal dependencies, likely to wash out color in motion areas. Different from previous methods, our method has strong ability in modeling long-term dependency both spatially and temporally.

3. Method

3.1. Problem formulation

Given consistent grayscale video frames $\{I_1^l, I_2^l, ..., I_n^l\}$, the colorization task aims to generate corresponding colorized frames $\{\hat{I}_1^{lab}, \hat{I}_2^{lab}, ..., \hat{I}_n^{lab}\}$, where l and ab denote the luminance and chrominance in CIELAB color space respectively. On the one hand, the generated result $\hat{I}_n^{lab}$ should be perceptually similar to the ground truth image I_n^{lab}. On the other hand, the current frame $\hat{I}_n^{lab}$ should not only be temporal consistent to its adjacent frames $\hat{I}_{n-1}^{lab}, \hat{I}_{n+1}^{lab}$, but also be similar to the frames with wide temporal interval (e.g. $\hat{I}_1^{lab}$). For recent automatic colorization methods [5, 6, 8], the colorization of I_n^l usually based on the previously colorized frame:

$$\hat{I}_n^{lab} = \mathcal{F}_{auto}(I_n^l, \hat{I}_{n-1}^{lab}), \tag{1}$$

where $\mathcal{F}_{auto}$ denotes the automatic colorization network. Such methods colorize the video in manner of a Markov Chain, while the consistency is established only for adjacent frame, and the frames in wide interval may be inconsistent. Meanwhile, exemplar-based methods [13, 16] usually colorize a frame depending on an additional reference image I_{ref}:

$$\hat{I}_n^{lab} = \mathcal{F}_{exemp}(I_n^l, \hat{I}_{n-1}^{lab}, I_{ref}), \tag{2}$$

where $\mathcal{F}_{exemp}$ represents the exemplar-based video colorization network. The reference image is responsible to supervise the colorization process. It determines the color style of images, thus reduce color ambiguity and enhance temporal consistency. However, the reference image usually needs manual selection which is experience required and time-consuming. Therefore, this paper propose a two-

stage colorization framework where the reference image is automatically generated and supervise the colorization.

3.2. Two-stages Colorization

Our overall framework is illustrated in Fig. 2. The framework is divided into two stages. The first stage involves an automatic reference colorization network, and the second stage includes a semantic correspondence network and an image colorization network. In the first stage, the first frame of each videos is selected to be automatically colorized. And the resulting image is then regarded as the reference image in the second stage.

$$\hat{I}_{ref}^{lab} = \mathcal{C}_1(I_0^l), \tag{3}$$

where $\mathcal{C}_1$ represents the reference colorization network. I_i, $\hat{I}_{ref}$ denote the $i-th$ frame and the reference image respectively. For maintaining temporal consistency, rather than only correlated to the previous few frames, the colorization of the remaining grayscale frames also depends on their semantic correspondence with the reference image, which can be formulated as:

$$\hat{I}_n^{lab} = \mathcal{C}_2(\mathcal{S}(I_n^l, \hat{I}_{ref}^{lab}), \hat{I}_{n-1}^{lab}), \tag{4}$$

where $\mathcal{S}$ represents the semantic correspondence network, and $\mathcal{C}_2$ the image colorization network in the second stage. Thus, our approach is capable of better maintaining temporal consistency along time series.

3.3. Loss function

As an inherent ambiguous problem, it is improper to directly compare the color difference between the ground truth and generated image. Recently, the perceptual difference has been proved to be robust to appearance differences caused by two plausible colors [21]. It compares the difference between features $reluL\_2$ extracted by pretrained VGG-19 network [22]. In this paper, the coarse-to-fine perceptual loss is leveraged:

$$\mathcal{L}_{perc} = \sum_L \alpha_L \|\Phi_L(\hat{T}) - \Phi_L(T)\|_2^2, \tag{5}$$

here $L \in \{3, 4, 5\}$, and $\alpha_L \in \{0.02, 0.003, 0.5\}$ denotes corresponding weight coefficient. The coarse-to-fine strategy involves the comparison of both high-level and low-level feature representations.

Besides, we empirically find that the L_1 loss helps the convergence of network, and the smooth loss [13] helps to reduce color bleeding. Moreover, the PatchGAN [23] is also adopted to increase high-frequency color fidelity. It classifies each patch as real or fake rather than the whole image. For networks in the first stage, the overall objective loss can be written as:

$$\begin{aligned}\mathcal{L}_1 =& \lambda_{perc} L_{perc} + \lambda_{L_1} L_{L_1} + \lambda_{smooth} L_{smooth} \\ &+ \lambda_{patch} L_{patch}\end{aligned} \tag{6}$$

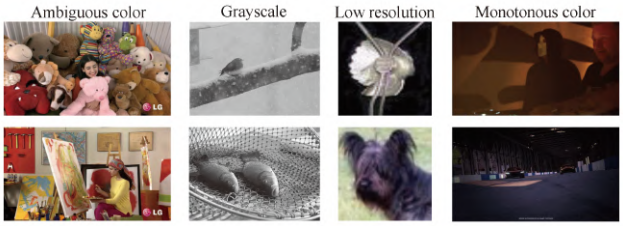

Figure 3. Examples of bad images removed from our training set. Including images with ambiguous colors, monotonous colors, low resolution or grayscale.

For networks in the second stage, the temporal warping loss [3] is further adopted to constraint temporal consistency. The corresponding objective loss is:

$$\begin{aligned}\mathcal{L}_2 =& \lambda_{perc} L_{perc} + \lambda_{L_1} L_{L_1} + \lambda_{smooth} L_{smooth} \\ &+ \lambda_{patch} L_{patch} + \lambda_{temp} L_{temp}\end{aligned} \tag{7}$$

4. Implementation

Network Structure. The reference colorization network is an encoder-decoder structure with skip connections, group convolutions and dilated convolutions [24]. The semantic correspondence network is a CNN-Transformer structure [25] with non-local operation [15]. And the image colorization network combines the encoder-decoder structure in the first stage with a Transformer branch. The network structure in the second stage is basically the same as in [16], and we recommend to check out more details from the original paper.

Training. The training process of the networks in two stages is independent. For network in the first stage, the reference colorization network is trained on images from ImageNet [26], REDS [27], DAVIS [28], SportMOT [29] and the official training set in NTIRE 2023 Video Colorization Challenge [30]. We remove the images with ambiguous colors, monotonous colors, low resolution or grayscale (Fig. 3). About 1.1 million of images are involved in training. For networks in the second stage, the training set includes DAVIS [28], Videvo [31] and FVI [32] dataset. 2090 videos in total are collected. And we train the networks in manner of frame propagation (i.e. the first frame in each video is regarded as the reference image). Moreover, The pretrained models in [13, 33] are used to initialize the parameters. One can refer to our published code for more implementation details.

5. Experiment

5.1. Comparisons with state-of-the-arts

In this section, state-of-the-art methods including ChromaGAN [34], DVP [4], FAVC [6] and VCGAN [5] are

Figure 4. Visual comparison with the state-of-the-art methods on Videvo test set. From top to bottom are the methods of ChromaGAN [34], ChromaGAN+DVP [4], FAVC [6] and VCGAN [5] respectively. Our method achieves the most colorful while consistent result.

Table 1. Quantitative comparison with state-of-the-art methods on DAVIS and Videvo dataset. Our method gets the best FID, while maintains comparable CDC.

	DAVIS		Videvo		
Method	FID↓	CDC↓	FID↓	CDC↓	Model type
ChromaGAN [34]	52.97	0.008771	50.57	0.004565	Fully-automatic (image)
ChromaGAN+DVP [4]	58.94	**0.003672**	58.85	0.001967	Task-independent
FAVC [6]	58.33	0.003682	57.08	**0.001575**	Fully-automatic
VCGAN [5]	59.58	0.008951	67.48	0.003208	Fully-automatic
Ours	**46.28**	0.003836	**49.02**	0.001681	Fully-automatic

compared with our method both quantitatively and qualitatively. The official published code of the methods are used for comparison.

Quantitative comparison. For quantitative comparison, the image quality metric FID (Fr´echet Inception Distance) [35] and temporal metric CDC (Color Distribution Consistency index) [3] are adopted, as which are widely used in previous works [3, 13, 16, 21, 30]. The FID measures the semantic distance between generated and ground truth images. The lower the FID, the more natural the image result. And the CDC computes the Jensen-Shannon (JS) divergence of the color distribution between consecutive frames. More consistent video will get lower CDC. We experiment on the test set of DAVIS [28] and Videvo [31] dataset, the quantitative result is illustrated in Tab. 1. The ChromaGAN gets excellent FID, but with bad CDC since it is an image colorization method without temporal modeling. With DVP, the CDC of ChromaGAN obviously declines, but its FID gets worse at the same time. FAVC achieves the second best and the best CDC in the two datasets respectively. However, it gets high FID. In both of the two datasets, our method achieves the best FID, while maintains comparable CDC with the best results.

Qualitative comparison. The visual comparison of the methods on Videvo test set is illustrated in Fig. 4. ChromaGAN generates colorful result, but the object color can be very different from frame to frame, as in the case of the ducks in the left video. The DVP distinctly removes the temporal inconsistency of ChromaGAN. However, it also washes out the colors in images (like in the right video), as it is tend to remove the bright but inconsistent colors rather than propagate the bright colors to other frames. The result of FAVC is quite consistent, but it is not colorful enough compared with other methods. VCGAN models dense temporal consistency in small frame interval. But with large frame interval, distinct inconsistency can be observed. Such as the grass in the left video and the water in the right video. Though the dense consistency in VCGAN with large interval is feasible, it requires huge computational consumption that dozens or hundreds of optical flows are required for a

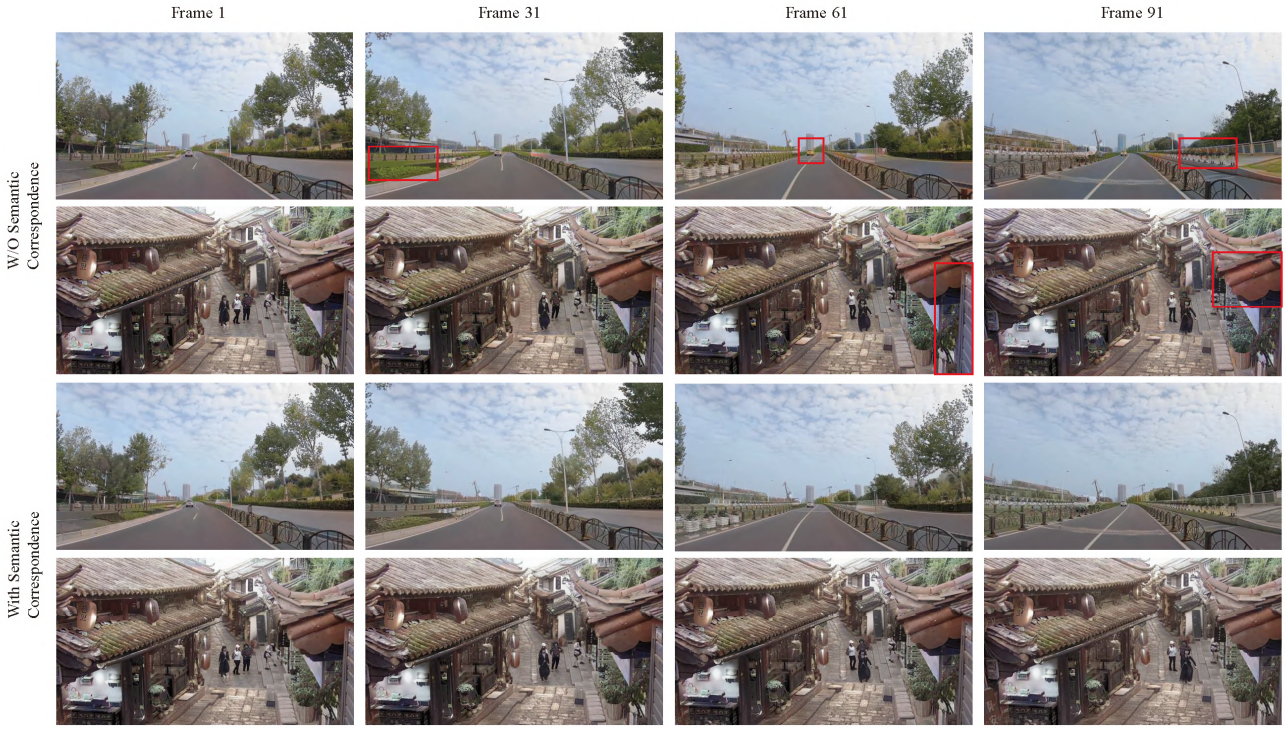

Figure 5. Visual comparison of colorization results for networks with or without semantic correspondence on the test set of NTIRE 2023 Video Colorization Challenge [30]. Each interval of the adjacent frames is 30.

Table 2. Comparison of our method with task-independent method DVP [4] on DAVIS test set. Our method maintains better color perceptual quality while obtaining the better temporal consistency compared to DVP. Moreover, we try to combine DVP with our method (the last row), and the CDC further improved, but with apparent performance drop in FID.

Image colorization	Semantic correspondence	DVP	FID↓	CDC↓	Time (s)	Parameters (M)
✓			41.12	0.005045	0.4648	32.80
✓		✓	58.85	0.004189	0.4648+0.9884	32.80+23.93
✓	✓		46.28	0.003836	0.4718	32.82+148.21
✓	✓	✓	59.92	0.003708	0.4718+0.9884	32.82+23.93+148.21

single image's training. With semantic correspondence, our method achieves the most colorful while consistent result.

5.2. Ablation study

Effect of semantic correspondence. We train another model without the semantic correspondence network in order to represent its effectiveness, and the task-independent method DVP [4] is also adopted for comparison. The quantitative result is reported in Tab. 2. Combining DVP with image colorization, the CDC gets improved, but with huge decline of FID (from 41.12 to 58.85). While with semantic correspondence, our method achieves better CDC and less drop in FID (from 41.12 to 46.28), which represents that our method maintains better color perceptual quality while obtaining better temporal consistency. Moreover, we try to combine DVP with our method, and the CDC further improved, but with apparent performance drop in FID (from 46.28 to 59.92).

Beside, the processing time per image and number of network parameters in Tab. 2 represents that our method consumes less time than task-independent method DVP (0.4718 sec. compared to 0.4648+0.9884 sec. per image) though with more parameters (32.82+148.21 M compared to 32.80+23+93 M). Moreover, the visual comparison for networks with or without semantic correspondence is illustrated in Fig. 5. Without semantic correspondence network, the object can have diverse colors in different frames (e.g. the color of the car could be white in frame 1 and yellow

Figure 6. Comparison of the colorization via reference images obtained by different methods on DAVIS test set. The methods including GCP (Generative Color Prior) [36], a retrieval method [21] and our reference colorization network. Each interval of the adjacent frames is 20. Our method obtains more colorful and realistic results.

Table 3. Comparison of different methods to automatically obtain reference image on DAVIS test set. Note that we have not specially trained the stage 2 network for reference images in stage 1.

	DAVIS		Videvo	
Reference	FID↓	CDC↓	FID↓	CDC↓
Retrieval [21]	53.38	0.003939	48.24	0.001793
GCP [36]	50.19	0.003843	48.51	0.001710
Ours-r	46.17	0.004179	45.35	0.001684

in frame 61). While with the semantic correspondence network, the frames with large interval still maintain pleasant temporal consistency.

Different reference selection strategy. We further introduce two methods to automatically obtain the reference image for video colorization, which are: (1) GCP (Generative Color Prior) [36], a generation method based on BigGAN [37]. The GCP learns a mapping from a grayscale image to a embedding which acts as the condition of BigGAN to generate a colorized image similar with the grayscale image. (2) a retrieval method [21] using PCA-based compression [38]. This method compares the PCA embedding of the grayscale image with embeddings of images from large dataset (e.g. ImageNet). And the image with largest correlation will be selected. The two methods are compared with our reference colorization network on DAVIS and Videvo datasets. The reference image obtained by three methods are used in our stage 2 network respectively. As the stage 2 networks are trained by frame propogation, we train another network mostly leveraging the retrieved images as references like in [13]. As shown in Tab. 3, in DAVIS dataset, our method gets the best FID. And in Videvo dataset, our method obtains both the best FID and CDC. Besides, the

Table 4. Test results on NTIRE 2023 Video Colorization Challenge [30]. Our method obtains the 3rd place in CDC track with 52.68% improvement over the baseline.

Team	FID↓	CDC↓
MiAlgo	54.72	0.000819
CUCPLUS	26.79	0.000962
Ours	**63.76**	**0.001017**
NJUSTer	62.45	0.001066
ppzz	56.81	0.001122
LVGroup HFUT	63.71	0.001525
baseline	61.30	0.002149

visual comparison is shown in Fig. 6. It can be noticed that our results are more colorful and realistic. We believe this is because our method obtains a reference image that is more similar to the grayscale frames, so we are able to transfer colors more precisely and obtain more vivid results.

5.3. NTIRE 2023 Challenge

We have proposed our method for NTIRE 2023 Video Colorization Challenge [30]. Our entry obtains the 3rd place in Track 2: Color Distribution Consistency (CDC) Optimization and the CDC score is very close to the second method. The goal of this track is to obtain the best CDC result while being constrained to maintain FID. The benchmark results of our model and the other teams in NTIRE 2023 are shown in Tab. 4.

5.4. Limitations

Despite the promising progress of our method for maintaining video temporal consistency, there are still some limitations.

Since we include the semantic correspondence network, our method is not robust enough when the scene changes in videos, which is an inherent weakness of the exemplar-based video colorization [16]. Meanwhile, as the reference image is automatically colorized, it is not likely to generate diverse image results.

Moreover, the performance of the final colorization result is highly dependent on the quality of the reference image. As shown in Fig. 7, we can apparently observe that the reference image has a high similarity to the subsequent colored images in terms of the color style. The reason is that the most of the colors from the reference image are considered plausible and will be transferred directly to the grayscale images. Even the incorrectly colored regions may still be considered as supervision information to guide the colorization of subsequent grayscale images. That is, the colorization of the reference image will affect the style of the videos (e.g., the color styles of the first and third rows are quite different due to the different reference images.), or even lead to unexpected colors (e.g., the ground in the second row is colorized to unpleasant green).

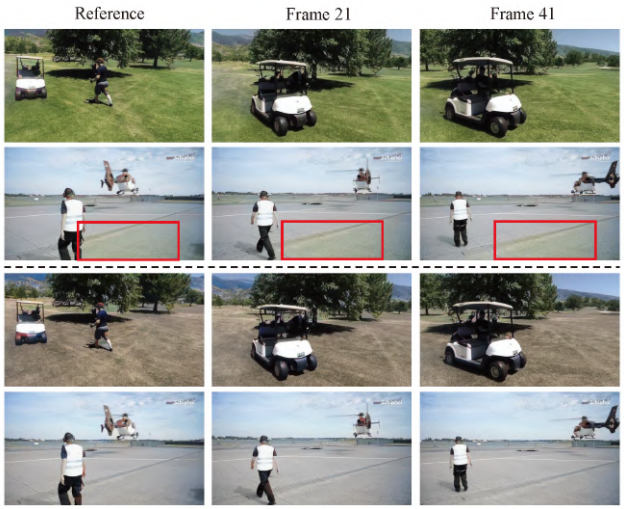

Figure 7. Colorization with different reference images obtained from our reference colorization network. The first and the second row show the results of the original model. The third and the fourth row show the results of another model trained without ImageNet [26] dataset.

6. Conclusion

In this paper, we propose a novel automatic video colorization method via semantic correspondence, which utilize an automatically generated reference image to supervise the colorization process and preserve temporal consistency. Our intuition is to fully exploit the semantic correspondence between video frames to improve the colorization consistency of the network. Experiment also demonstrated that our method is capable of better maintaining color consistency in large frame interval than recent methods. Finally, ablation studies show the effectiveness of the network components.

Acknowledgements

This work was supported by the National Key R&D Program of China (2021ZD0109802) and the National Natural Science Foundation of China (81972248).

References

[1] N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister. Blind video temporal consistency. *ACM Transactions on Graphics*, 34(6cd):1–9, 2015. 1, 2

[2] W. S. Lai, J. B. Huang, O. Wang, E. Shechtman, E. Yumer, and M. H. Yang. Learning blind video temporal consistency. In *European Conference on Computer Vision*, 2018. 1, 2

[3] 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, 4, 5

[4] Chenyang Lei, Yazhou Xing, Hao Ouyang, and Qifeng Chen. Deep video prior for video consistency and propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022. 1, 2, 4, 5, 6

[5] Yuzhi Zhao, Lai-Man Po, Wing Yin Yu, Yasar Abbas Ur Rehman, Mengyang Liu, Yujia Zhang, and Weifeng Ou. Vcgan: Video colorization with hybrid generative adversarial network. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 1, 2, 3, 4, 5

[6] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3753–3761, 2019. 1, 2, 3, 4, 5

[7] Panagiotis Kouzouglidis, Giorgos Sfikas, and Christophoros Nikou. Automatic video colorization using 3d conditional generative adversarial networks. In *International Symposium on Visual Computing*, pages 209–218. Springer, 2019. 1, 2

[8] H. Thasarathan, K. Nazeri, and M. Ebrahimi. Automatic temporally coherent video colorization. In *2019 16th Conference on Computer and Robot Vision (CRV)*, 2019. 1, 2, 3

[9] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3

[10] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018. 1, 3

[11] Sifei Liu, Guangyu Zhong, Shalini De Mello, Jinwei Gu, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Switchable temporal propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018. 1, 3

[12] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 1, 3

[13] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8052–8061, 2019. 1, 3, 4, 5, 7

[14] Yaxin Liu, Xiaoyan Zhang, and Xiaogang Xu. Reference-based video colorization with multi-scale semantic fusion and temporal augmentation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1924–1928. IEEE, 2021. 1, 3

[15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1, 4

[16] Siqi Chen, Xueming Li, Xianlin Zhang, Mingdao Wang, Yu Zhang, Jiatong Han, and Yue Zhang. Exemplar-based video colorization with long-term spatiotemporal dependency, 2023. 2, 3, 4, 5, 8

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[18] Vivek George Jacob and Sumana Gupta. Colorization of grayscale images and videos using a semiautomatic approach. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 1653–1656. IEEE, 2009. 3

[19] Nir Ben-Zrihem and Lihi Zelnik-Manor. Approximate nearest neighbor fields in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5233–5242, 2015. 3

[20] Sifeng Xia, Jiaying Liu, Yuming Fang, Wenhan Yang, and Zongming Guo. Robust and automatic video colorization via multiframe reordering refinement. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 4017–4021. IEEE, 2016. 3

[21] Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018. 4, 5, 7

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 4

[24] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 4

[25] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 367–376, 2021. 4

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4, 8

[27] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, June 2019. 4

[28] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 4, 5

[29] Sportsmot. *https://deeperaction.github.io/datasets/sportsmot.html*. 4

[30] Xiaoyang Kang, Xianhui Lin, et al. Ntire 2023 video colorization challenge. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 4, 5, 6, 8

[31] Videvo. *https://www.videvo.net/*. 4, 5

[32] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9066–9075, 2019. 4

[33] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 4

[34] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2445–2454, 2020. 4, 5

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[36] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14377–14386, 2021. 7

[37] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 7

[38] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 584–599. Springer, 2014. 7