NTIRE 2023 Video Colorization Challenge

Xiaoyang Kang* Xianhui Lin* Kai Zhang* Zheng Hui* Wangmeng Xiang* Jun-Yan He* Xiaoming Li* Peiran Ren* Xuansong Xie* Radu Timofte* Yixin Yang Jinshan Pan Zhongzheng Peng Qiyan Zhang Jiangxin Dong Jinhui Tang Jinjing Li Chichen Lin Qipei Li Qirong Liang Ruipeng Gang Xiaofeng Liu Shuang Feng Shuai Liu Hao Wang Chaoyu Feng Furui Bai Yuqian Zhang Guangqi Shao Xiaotao Wang Lei Lei Siqi Chen Yu Zhang Hanning Xu Zheyuan Liu Zhao Zhang Yan Luo Zhichao Zuo

Abstract

This paper reviews the video colorization challenge on the New Trends in Image Restoration and Enhancement (NTIRE) workshop, held in conjunction with CVPR 2023. The target of this challenge is converting grayscale videos into color videos with better colorization performance and temporal consistency. The challenge consists of two tracks. For Track 1, the goal is achieving the best FID (Fréchet Inception Distance) while being constrained to maintain or improve over the baseline method in terms of the temporal-consistency metric. The Color Distribution Consistency (CDC) index is used as the temporal consistency evaluation metric in this challenge. For Track 2, the target is to obtain a solution with the best CDC result while being constrained to maintain or improve over the baseline method in terms of FID. We use DeOldify-video as the baseline method for two tracks. For the final testing phase of both tracks, six teams submitted fact sheets and executable code of their solutions. This report brings together descriptions and discussions of all these solutions. Both tracks use the same data and the datasets are available at this url.

1. Introduction

Video colorization aims to transform multiple consecutive single-channel grayscale video frames into three-channel color video frames, and has received increasing attention in recent years. Its applications are vast and varied, spanning across the film industry, art, and visual media. Unlike image colorization, video colorization not only demands high-fidelity single-frame results but also necessitates maintaining temporal consistency between frames. In addition, instance consistency must be ensured in video colorization, *e.g.*, objects that appear in the previous frames should retain the same semantic color in subsequent ones. Thus, video colorization is a challenging problem in visual enhancement and restoration.

Recently, a large number of image colorization methods [7, 11, 16, 18, 19, 20, 35, 44, 45, 59] have been proposed and achieved impressive results. One possible solution for video colorization is to directly use the image colorization model to independently colorize each frame of the video. However, due to the lack of modeling of temporal information between frames, these image-based methods often result in temporal flickering and discontinuity.

In order to introduce temporal constraints, FAVC [22] first uses deep learning methods to achieve automatic video colorization by using self-regularization and diversity loss. TCVC [28] propagates frame-level deep features in a bidirectional manner, achieving better single-frame colorization results while enhancing temporal consistency. To achieve better flexibility and colorization results, some exemplar-based video colorization methods [14, 38, 48, 58] have been developed. These methods typically transfer colors from reference sample images to grayscale image frames. BiSTNet [53] uses bidirectional temporal feature fusion with the guidance of semantic image prior to achieve progressive colorization in a coarse-to-fine manner.

The goal of the NTIRE 2023 Video Colorization challenge is to promote further research in the video colorization field and to establish the current state-of-the-art. As part of the challenge, the participants were required to generate continuous color video frames giving multiple grayscale video frames as input. The challenge contained two tracks, namely Track 1 and Track 2. For Track 1, the aim is to obtain a solution with the best FID (Fréchet Inception Distance) [12] while being constrained to maintain or improve over the baseline method in terms of temporal-consistency metric. We use the Color Distribution Consistency (CDC)

*Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie and Radu Timofte are the NTIRE 2023 video colorization challenge organizers. The other authors participated in the challenge.
Appendix contains the authors' team names and affiliations.
NTIRE 2023 Webpage: https://cvlai.net/ntire/2023.

described in [28] as the temporal consistency evaluation metric. For Track 2, the aim is to obtain a solution with the best CDC result while being constrained to maintain or improve over the baseline method in terms of FID. Both tracks use the famous open-source video colorization model DeOldify-video [4] as the baseline method.

The challenge has 93 and 130 registered participants for two tracks, respectively. Among them, 6 participating teams submitted valid models and fact sheets in the final testing stage of each track. They introduce new technologies in network architectures, loss functions, ensemble methods, data augmentation methods, etc. We present detailed challenge results in Section 3.

This challenge is part of the NTIRE 2023 challenges on: night photography rendering [34], HR depth from images of specular and transparent surfaces [55], image denoising [26], video colorization [17], shadow removal [36, 37], quality assessment of video enhancement [27], stereo super-resolution [39], light field image super-resolution [42], image super-resolution ($\times 4$) [61], 360° omnidirectional image and video super-resolution [5], lens-to-lens bokeh effect transformation [8, 33], real-time 4K super-resolution [9, 56], HR nonhomogenous dehazing [3], efficient super-resolution [25].

2. Challenge

The NTIRE 2023 Video Colorization Challenge addresses the black-and-white video colorization task. To the best of our knowledge, this is the first challenge to focus on general video colorization. It aims to assess and advance the latest level of video colorization and sets up two tracks that emphasize high-fidelity and time-consistent solutions respectively. The rest of this section describes challenge settings, including the dataset, evaluation, as well as phases of challenges.

2.1. Dataset

For the NTIRE 2023 Video Colorization Challenge, we employ a subset of Large-scale Diverse Video (LDV) dataset [49, 50, 51] as the training set and the validation set. The LDV dataset includes diverse categories of contents, various kinds of motion and different frame-rates. The original LDV dataset contains 240 high-quality videos with a resolution of 960×536. We use 200 of them as the training set and 15 of them as the validation set. The validation set is further divided into video frames that are publicly available to minimize the differences caused by different video decoding methods. The video frames are converted to grayscale using '`cv2.cvtColor()`'.

The test set contains 15 diverse videos collected from YouTube. Each video contains 100 grayscale frames in the size of 960×540. The videos contain multiple types of scenes, *e.g.*, animal, city, human, indoor, scenery, sports, and so on.

2.2. Evaluation

Following the experimental protocol of most existing colorization methods, we mainly use Fréchet Inception Distance (FID) [12] to evaluate the colorization performance of the methods, where FID measures the distribution similarity between generated images and ground truth images. Although colorization is an inverse problem, it is a widely held view that the pixel-level metrics such as Peak Signal-to-Noise Ratio (PSNR) [13] may not well reflect the actual colorization performance [7, 16, 18, 35, 45].

For temporal consistency, we adopt Color Distribution Consistency index (CDC) described in [28]. It is computed on the output colorized frames. Specifically, it computes the Jensen-Shannon (JS) divergence of the color distribution between consecutive frames:

$$CDC_t = \frac{1}{3 \times (N-t)} \sum_{c \in \{r,g,b\}} \sum_{i=1}^{N-t} JS(P_c(I^i), P_c(I^{i+t})), \tag{1}$$

where N is the video sequence length and $P_c(I^i)$ is the normalized probability distribution of color image I^i across c channel, which can be calculated from the image histogram. t denotes the time step. A smaller t indicates short-term temporal consistency, while larger t indicates long-term temporal consistency. The JS divergence can measure the similarity between two color probability distributions. The overall index can be calculated by:

$$CDC = \frac{1}{3}(CDC_1 + CDC_2 + CDC_4). \tag{2}$$

which considers the long-term and short-term temporal consistency together.

For Track 1, the aim is to obtain a solution with the best FID while being constrained to maintain or improve over the baseline in terms of CDC. For Track2, the aim is to obtain solutions with the best CDC while being constrained to maintain or improve over the baseline in terms of FID. For both tracks, we choose DeOldify, the famous open-source colorization method, as the baseline defining the maximum CDC / FID. The baseline evaluation code can be found in https://modelscope.cn/models/damo/CVPR2023_NTIRE_Video_Colorization/summary.

2.3. Challenge Phases

The whole challenge consists of three phases: the developing phase, the validation phase, and the testing phase.

In the developing phase, the participants can access to both grayscale and color videos of the training set. This period allows them to become familiar with data structure while also developing their algorithms.

In the validation phase, the participants can access the grayscale video frames of the validation set. The participants had the opportunity to test their solutions on the validation images and receive immediate feedback by uploading results onto the server. A validation leaderboard is available.

In the testing phase, the participants can access the grayscale video frames of the test set. A test server and a leaderboard are provided. At the end of the test phase, the participants need to submit the executable file and a detailed description file outlining their methods before receiving the final rank.

3. Results

From 93 and 130 registered teams in two tracks, 15 and 14 teams advanced to the final testing phase respectively. Among them, 6 teams submitted valid results and fact sheets for both tracks. These teams are ranked according to the evaluation metrics presented in Section 2.2.

3.1. Track 1: Fréchet Inception Distance (FID) Optimization

For Track 1, we use DeOldify-video [4] as the baseline method. The final results are ranked by FID, which means that top solutions are expected to achieve a lower FID score while maintaining its CDC score lower than DeOldify-video's. Table 1 reports the FID, CDC scores, and the final ranking of each team. Fig. 1 shows the qualitative results. The proposed methods of each team are described in Section 4.1.

Table 1. FID and CDC score over the test set and final rankings on Track 1. We denote in bold the main metric of the track.

Team	Author	**FID↓**	CDC↓
NJUSTer	Yixin Yang	21.5372	0.001717
CUCPLUS	Jinjing Li	26.7915	0.000963
MiAlgo	Shuai Liu	41.9539	0.001450
vectoria	Siqi Chen	55.9904	0.001714
ppzz	Hanning Xu	56.8085	0.001122
LVGroup HFUT	Zhao Zhang	60.0732	0.002548
baseline	-	61.2961	0.002149

3.2. Track 2: Color Distribution Consistency (CDC) Optimization

For Track 2, we use DeOldify-video [4] as the baseline method. The final results are ranked by CDC, which means that top solutions are expected to achieve a lower CDC score while maintaining its FID score lower than DeOldify-video's. Table 2 reports the FID, CDC scores, and the final ranking of each team. Fig. 2 shows the qualitative results. The proposed methods of each team are described in Section 4.2.

Table 2. FID and CDC score over the test set and final rankings on Track 2. We denote in bold the main metric of the track.

Team	Author	FID↓	**CDC↓**
MiAlgo	Shuai Liu	54.7238	0.000819
CUCPLUS	Jinjing Li	26.7934	0.000962
vectoria	Siqi Chen	63.7640	0.001017
NJUSTer	Yixin Yang	62.4467	0.001066
ppzz	Hanning Xu	56.8085	0.001122
LVGroup HFUT	Zhao Zhang	63.7058	0.001525
baseline	-	61.2961	0.002149

4. Teams and Methods

In this section, we briefly describe the methods proposed by teams participating in the final testing phase of the NTIRE 2023 Video Colorization Challenge.

4.1. Track 1

4.1.1 NJUSTer

The NJUSTer team adopted BiSTNet [53] as their baseline model. BiSTNet is a deep video colorization method that leverages semantic image prior to guide bidirectional temporal feature fusion. It can effectively exploit the color information of reference exemplars and propagate it to colorize each frame. BiSTNet consists of several core components: (a) bidirectional temporal fusion block (BTFB), which fuses the features of adjacent frames in both forward and backward directions; (b) mixed expert module (MEB), which selects different colorization strategies based on the semantic image prior; (c) multi-scale recurrent framework (MSRB), which progressively colorizes each frame from coarse to fine. BiSTNet has been evaluated on multiple datasets and demonstrated its superiority in both quantitative and qualitative aspects.

For example-based video coloring methods, high-quality reference frames are crucial. The NJUSTer team first experimented DISCO [46], an image colorization method, to generate colorful reference frames. They fine-tuned this model with the NTIRE2023 Video Colorization training dataset, and colored 'f001.png', 'f050.png', and 'f100.png' frames (key colorful frames) required by BiSTNet. Experimental results show that DISCO generates reference frames with good visual effects but overall colorfulness is far from satisfactory. Moreover, since DISCO does not consider temporal consistency between frames, selecting reference frames will cause color inconsistency of the same objects. They discovered that the accuracy of the color greatly impacts the calculation of the FID score. Even if the generated image color is reasonable, there is a significant difference from the ground truth, which can result in a high FID value (the lower, the

Figure 1. Qualitative Results for Track 1. Best viewed in color.

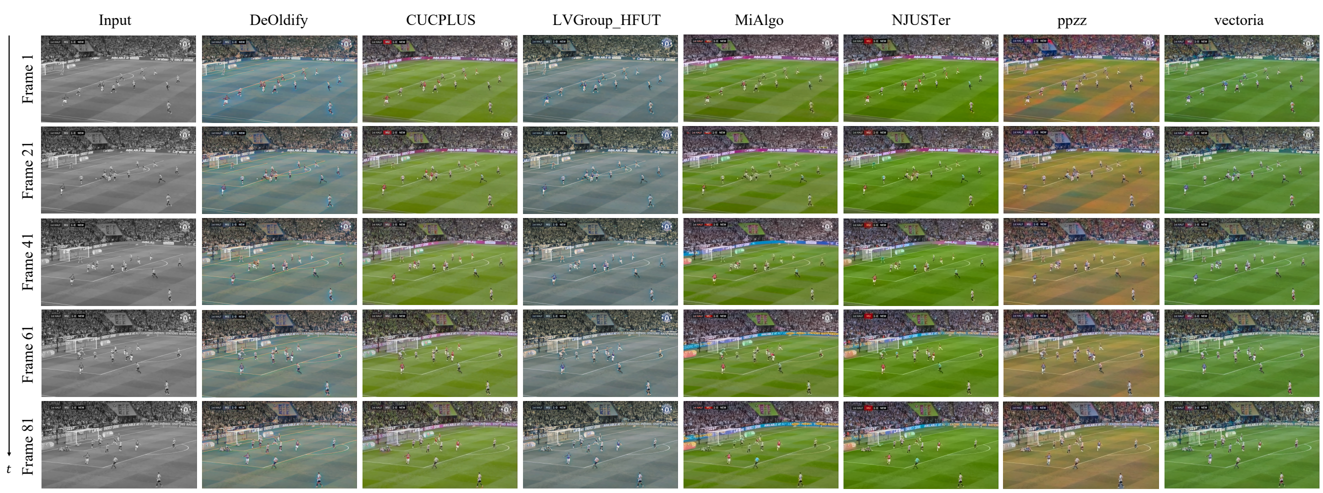

Figure 2. Qualitative Results for Track 2. Best viewed in color.

better). Therefore, they ultimately chose to search for the closest possible color images from the internet as reference frames for BiSTNet. The video colorization model based on these reference frames performs very competitively (please see in Table 3). In conclusion, their research demonstrates that our model performs well enough when there are high-quality reference frames available. When high-quality reference frames are not accessible, image colorization methods (like DISCO) equipped with human manual coloring are also a good alternative solution.

Table 3. Team NJUSTer: the impact of the colorful reference.

source of key frames	**FID↓**	**CDC↓**
from the internet	21.5372	0.001717
from the DISCO [46]	73.4874	0.001716

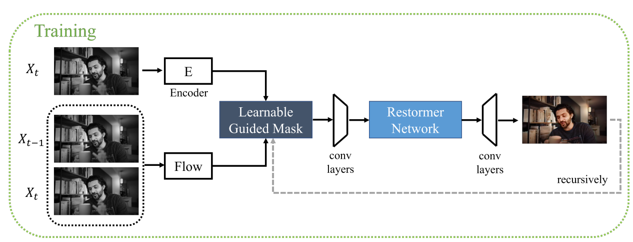

Figure 3. The pipeline of BRT proposed by Team CUCPLUS.

4.1.2 CUCPLUS

The CUCPLUS team proposed a Bi-directional Recurrent Transformer Network for Video Colorization [24]. As shown in Fig. 3, the proposed network BRT is based on RTN [38], with the difference being that the transformer adopted in it is the Restormer Network [57]. BRT takes a series of video frames as input, a shallow feature, and optical flow information will be obtained after Encoder and Flow, respectively. Learnable Guided Mask [38] module will fusion this information and feed the output feature maps into Restormer Network, which can capture more multi-scale semantics during training and reduce the occurrence of color artifacts. The Learnable Guided Mask block adopted in BRT is with the same setting as in RTN [38]. In addition, the proposed BRT employs adversarial training as the same as RTN [38] during the training process.

During the inference phase, considering the potential discrepancy between the input image distribution during model training and testing, the proposed BRT employed the TLC [47] strategy to alleviate such a gap in the inference phase. As shown in Fig. 4, BRT leverages different model weights to perform inference predictions on subsets of the test set. At the same time, some video clips in the test set contain rich content, such as numerous characters, and the generalization ability of BRT is poor on such clips. To address this issue, firstly, BRT utilized the baseline [4] method to generate synthetic data, and then trained the model on these data. Secondly, BRT is trained with the extra clips which are collected from YouTube for finetuning.

4.1.3 MiAlgo

The MiAlgo team proposes a multi-model fusion strategy for Track 1. As shown in Fig. 5, the approach involves training CT2 [44] on a large amount of unfiltered YouTube data [52] and naming the model as "CT2 classic". Subsequently, the data is cleaned and filtered to select video frames of common scenes using the official training set. Another CT2 model is trained on the filtered data, and named "CT2 vivid". During testing, a content-based image retrieval system (CBIR) is utilized to match the test video and the filtered training set. If the distance exceeds a threshold, the vivid result is used, otherwise, the classic result is used. This design is intended to increase the robustness of the method by using different models for common and uncommon scenarios.

4.1.4 vectoria

The vectoria team proposes Temporal Consistent Automatic Video Colorization with Semantic Correspondence [60], which combines semantic correspondence network into automatic video colorization. As illustrated in Fig. 6, the proposed framework is divided into two stages. The first stage involves an automatic image 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 video is selected to be automatically colorized. And the resulting image is then regarded as a reference image in the second stage.

$$I_{ref}^{lab} = \mathcal{C}_1(I_0^l) \tag{3}$$

In which $\mathcal{C}_1$ represents the image colorization network in the first stage. I_i, I_{ref} denote the i^{th} frame and the reference image respectively. For maintaining temporal consistency, rather than only correlating 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 denoted by:

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

Where $\mathcal{S}$ represents the semantic correspondence network, and $\mathcal{C}_2$ the image colorization network in the second stage. Thus, this approach is capable of better maintaining temporal consistency along time series. They train another model without the semantic correspondence network to represent its effectiveness, and the visual comparison is illustrated in Fig. 7. Without a semantic correspondence network, the object can have diverse colors in different frames. With the semantic correspondence network, the frames with large intervals still maintain pleasant temporal consistency.

The image colorization network in the first stage is an encoder-decoder structure with skip connections, group convolutions, and dilated convolutions [54]. The semantic correspondence network is a CNN-Transformer structure [30] with non-local operation [41]. And the image colorization network in the second stage combines the encoder-decoder structure in the first stage with a Transformer branch.

The training of the networks in two stages is independent. For the network in the first stage, the image colorization network is trained on images from ImageNet [10], REDS [29], DAVIS [31], SportMOT [1] and the official training set in the competition. The images in odd colors, low resolution, or low contrast are removed. About 1.1 million images are involved in training. Image-based objectives: L1 loss, perceptual loss, generator loss, and smoothness loss [58] are adopted. And for networks in the second stage, the training set includes DAVIS [31], Videvo [2], and FVI [6] dataset. 2090 videos in total are collected. Moreover, The pre-trained models in [43, 58] are used to initialize the parameters. Besides the image-based objectives, video-based objective temporal warping loss [28] is also adopted.

4.1.5 ppzz

The ppzz team proposed a method that uses two pretrained models to generate the final test results. They use a ColorFormer [16] pretrained on ImageNet to generate the exemplar images regarding each video clip. These exemplar

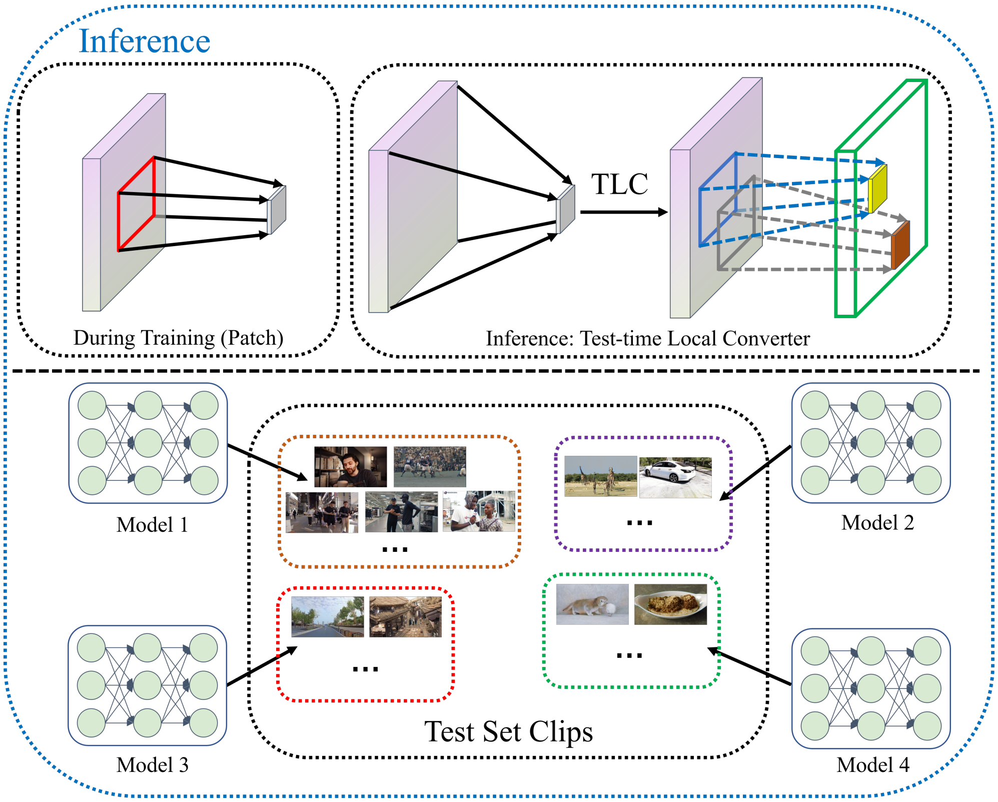

Figure 4. Team CUCPLUS: Test phase inference strategy. Firstly, BRT uses TLC [47] to alleviate the gap between training and testing. Secondly, different clips of the test set are predicted by different model weights.

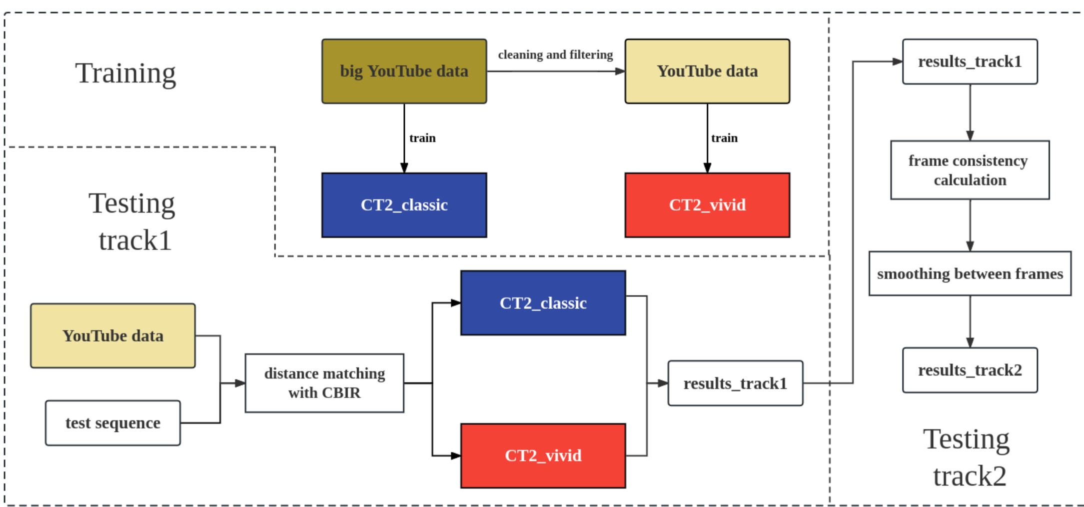

Figure 5. Overview of the approach used by Team MiAlgo.

images are further utilized by a Deep-Exemplar [58] pre-trained on Videvo and Hollywood2 datasets to produce the colorized frames.

Compared with other SOTA images colorization models [44, 45], employing ColorFormer [16] as the exemplar-generation backbone has three advantages: 1) stability to produce highly-coefficient images when given frame sequences; 2) fast inference speed; 3) low memory consumption.

Compared with other SOTA video colorization mod-

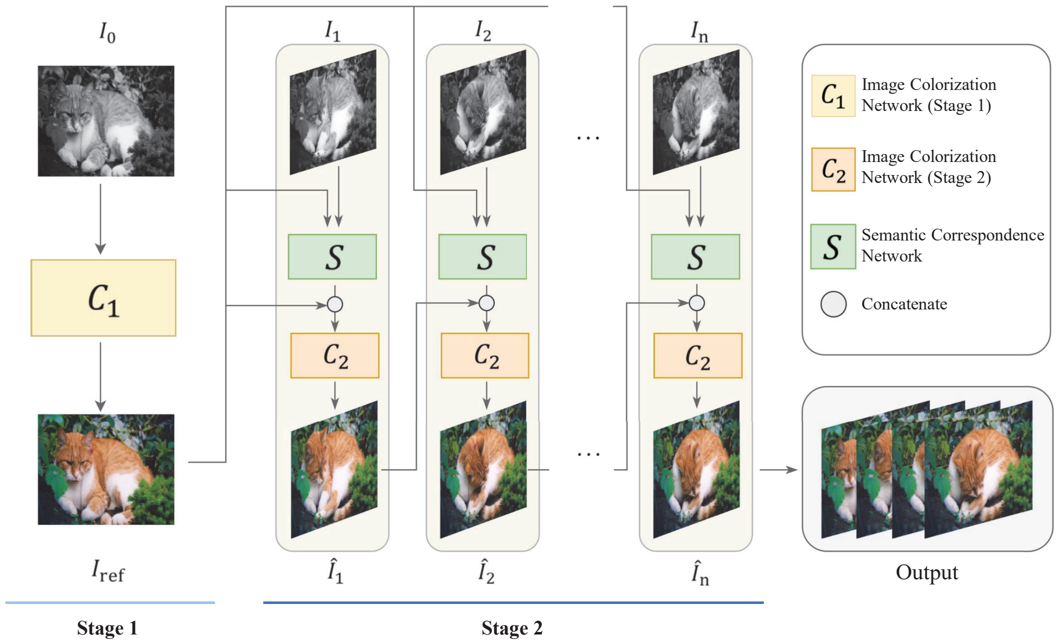

Figure 6. The overall framework used by Team vectoria

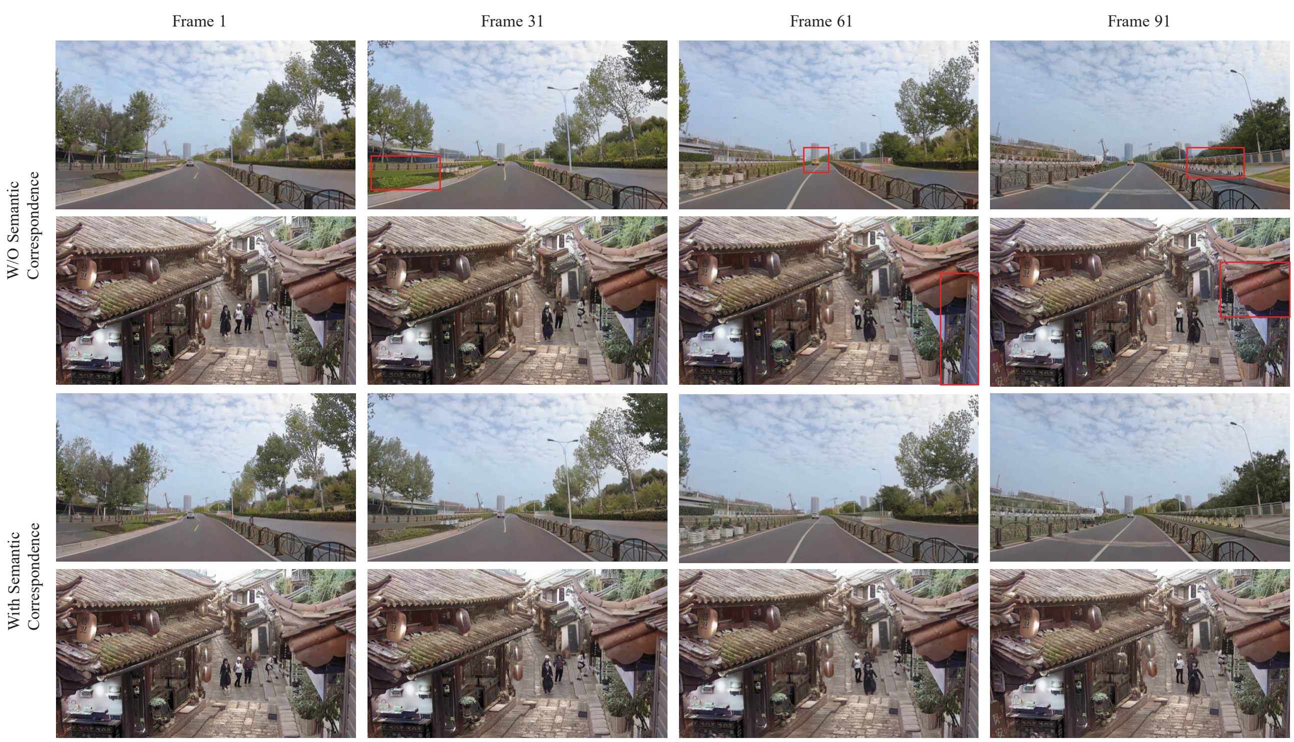

Figure 7. Team vectoria: visual comparison of colorization results with or without semantic correspondence network. The images are selected from the official test set. Each interval of the adjacent frames is 30.

els which are based on single image colorization methods [21, 22, 23, 28], employing Deep-Exemplar [58] as the video colorization backbone also has three advantages: 1) astonishing high temporal consistency between generated frames especially when the exemplar image has the similar structure as the gray frames; 2) fast inference speed; 3) low

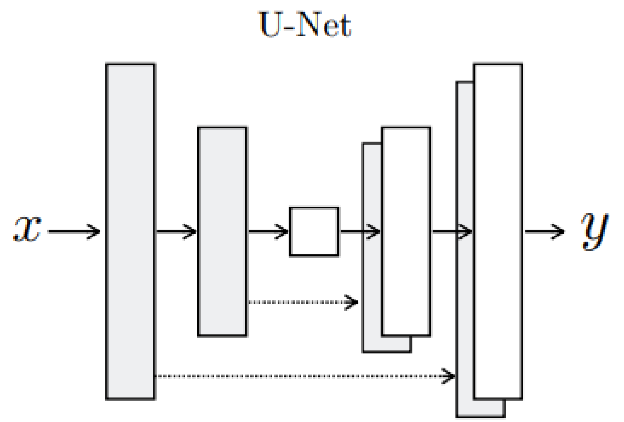

Figure 8. The network architecture employed by team LVGroup_HFUT for Track 1.

memory consumption.

4.1.6 LVGroup_HFUT

The LVGroup_HFUT team uses U-Net [32] as the backbone of the proposed method for Track 1. The skip connection [15] is applied between mirrored layers in the encoder and decoder stacks as shown in Fig. 8. They add skip connections between each layer i and layer $n - i$, where n is the total number of layers. Each skip connection simply concatenates all channels at layer i with those at layer $n - i$. This approach promotes the decoder to preserve low-level details and facilitates the convergence of the whole system since the gradients easily pass to encoder layers.

4.2. Track 2

4.2.1 MiAlgo

As illustrated in Fig. 5, the models used for testing Track 2 are the same as those employed in Track 1. These models are used to calculate the CDC of each test sequence. After the CDC values have been calculated, inter-frame smoothing is applied with varying strengths based on the CDC value. The purpose of this approach is to improve the performance of the method by reducing the impact of camera motion on the visual content of the test sequence.

4.2.2 CUCPLUS

The CUCPLUS team proposes the same method for both tracks, which is described in Section 4.1.2.

4.2.3 vectoria

The vectoria team proposes the same method for both tracks, which is described in Section 4.1.4.

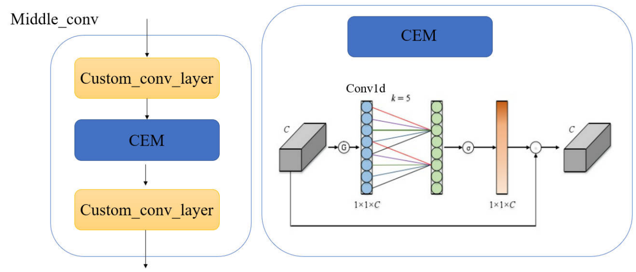

Figure 9. The network architecture proposed by team LVGroup_HFUT for Track 2.

4.2.4 NJUSTer

The NJUSTer team proposes the same method for both tracks, which is described in Section 4.1.1.

4.2.5 ppzz

The ppzz team proposes the same method for both tracks, which is described in Section 4.1.5.

4.2.6 LVGroup_HFUT

The LVGroup_HFUT team proposes a Channel Enhancement Module to enhance the performance of Deoldify [4] for Track 2. As shown in Fig. 9, the proposed Channel Enhancement Module (CEM) mainly adopts ECA [40] to enhance channel information locally. A local cross-channel interaction strategy without dimensionality reduction is used to avoid dimension reduction for channel attention learning. Additionally, the appropriate cross-channel interaction can significantly reduce model complexity while maintaining performance.

Acknowledgments

We thank the NTIRE 2023 sponsors: Sony Interactive Entertainment, Meta Reality Labs, ModelScope, ETH Zürich (Computer Vision Lab) and University of Würzburg (Computer Vision Lab). We thank ModelScope for sponsoring this challenge. We thank all organizers and all the participants for their great work.

Appendix

A. Teams and Affiliations

NTIRE2023 Video Colorization Challenge Organizers

Members:

Xiaoyang Kang[1] (*kangxiaoyang.kxy@alibaba-inc.com*)
Xianhui Lin[1] (*xianhui.lxh@alibaba-inc.com*)
Kai Zhang[2] (*kai.zhang@vision.ee.ethz.ch*)

Zheng Hui[1] (*huizheng.hz@alibaba-inc.com*)
Wangmeng Xiang[1] (*wangmeng.xwm@alibaba-inc.com*)
Jun-Yan He[1] (*leyuan.hjy@alibaba-inc.com*)
Xiaoming Li[3] (*csxmli@gmail.com*)
Peiran Ren[1] (*peiran.rpr@alibaba-inc.com*)
Xuansong Xie[1] (*xingtong.xxs@alibaba-inc.com*)
Radu Timofte[4,2] (*radu.timofte@uni-wuerzburg.de*)
Affiliations:
[1] DAMO Academy, Alibaba Group, China
[2] Computer Vision Lab, ETH Zürich, Switzerland
[3] Nanyang Technological University, Singapore
[4] University of Würzburg, Germany

NJUSTer

Title:
BiSTNet: Semantic Image Prior Guided Bidirectional Temporal Feature Fusion for Deep Exemplar-based Video Colorization
Team Leader:
Yixin Yang[1] (*yixin.yang@email.ucr.edu*)
Members:
Jinshan Pan[1], *Zhongzheng Peng*[1], *Qiyan Zhang*[1], *Jiangxin Dong*[1], *Jinhui Tang*[1]
Affiliation:
[1] Nanjing University of Science and Technology

CUCPLUS

Title:
Bi-directional Recurrent Transformer Network for Video Colorization
Team Leader:
Jinjing Li[1] (*1792418414@qq.com*)
Members:
Chichen Lin[1], *Qipei Li*[1], *Qirong Liang*[1], *Ruipeng Gang*[2 3], *Xiaofeng Liu*[1], *Shuang Feng*[1]
Affiliations:
[1] Communication University of China, Beijing 100024, China
[2] Academy of Broadcasting Science, NRTA, Beijing 100866, China
[3] UHDTV Research and Application Laboratory, Beijing 100176, China

MiAlgo

Title:
Multi-model fusion strategy
Team Leader:
Shuai Liu[1] (*liushuai21@xiaomi.com*)
Members:
Hao Wang, Chaoyu Feng, Furui Bai, Yuqian Zhang, Guangqi Shao, Xiaotao Wang, Lei Lei
Affiliation:
[1] Xiaomi Inc., China

vectoria

Title:
Temporal Consistent Automatic Video Colorization with Semantic Correspondence
Team Leader:
Siqi Chen[1] (*sqchen@bupt.edu.cn*)
Member:
Yu Zhang[1]
Affiliation:
[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

ppzz

Title:
Deep Exemplar Feature Propagation for Video Colorization
Team Leader:
Hanning Xu[1] (*hanningxu@zjut.edu.cn*)
Member:
Zheyuan Liu[1]
Affiliation:
[1] ZJUT

LVGroup_HFUT

Title of Track 1:
Skip connection of the Unet
Title of Track 2:
Channel Enhancement Module for Deoldify
Team Leader:
Zhao Zhang[1] (*cszzhang@gmail.com*)
Members:
Yan Luo[1], *Zhichao Zuo*[1]
Affiliation:
[1] Hefei University of Technology (HFUT)

References

[1] Sportsmot. *https://deeperaction.github.io/datasets/sportsmot.html*. 5
[2] Videvo. *https://www.videvo.net/*. 5
[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2023 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2
[4] Jason Antic. jantic/deoldify: A deep learning based project for colorizing and restoring old images (and video!).

https://github.com/jantic/DeOldify, 2019. 2, 3, 5, 8

[5] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Ying Shan, Gen Li, Radu Timofte, et al. NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[6] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *International Conference on Computer Vision*, pages 9066–9075, 2019. 5

[7] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *International Conference on Computer Vision*, pages 415–423, 2015. 1, 2

[8] Marcos V Conde, Manuel Kolmet, Tim Seizinger, Thomas E. Bishop, Radu Timofte, et al. Lens-to-lens bokeh effect transformation. NTIRE 2023 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[9] Marcos V Conde, Eduard Zamfir, Radu Timofte, et al. Efficient deep models for real-time 4k image super-resolution. NTIRE 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[11] Shuhang Gu, Radu Timofte, and Richard Zhang. Ntire 2019 challenge on image colorization: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[12] 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. *Annual Conference on Neural Information Processing System*, 30, 2017. 1, 2

[13] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 2

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

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 8

[16] Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. ColorFormer: Image colorization via color memory assisted hybrid-attention transformer. In *European Conference on Computer Vision*, pages 20–36, 2022. 1, 2, 5, 6

[17] Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie, Radu Timofte, et al. NTIRE 2023 video colorization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[18] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. DDColor: Towards photo-realistic and semantic-aware image colorization via dual decoders. *arXiv preprint arXiv:2212.11613*, 2022. 1, 2

[19] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. BigColor: Colorization using a generative color prior for natural images. In *European Conference on Computer Vision*, pages 350–366, 2022. 1

[20] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *International Conference on Learning Representations*, 2021. 1

[21] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *European Conference on Computer Vision*, pages 170–185, 2018. 7

[22] Chenyang Lei and Qifeng Chen. Fully automatic video colorization with self-regularization and diversity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3753–3761, 2019. 1, 7

[23] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. *Annual Conference on Neural Information Processing System*, 33:1083–1093, 2020. 7

[24] Jinjing Li, Qirong Liang, Qipei Li, Ruipeng Gang, Ji Fang, Chichen Lin, Shuang Feng, and Xiaofeng Liu. Rttlc: Video colorization with restored transformer and test-time local converter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 4

[25] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[26] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[27] Xiaohong Liu, Xiongkuo Min, Wei Sun, Yulun Zhang, Kai Zhang, Radu Timofte, Guangtao Zhai, Yixuan Gao, Yuqin Cao, Tengchuan Kou, Yunlong Dong, Ziheng Jia, et al. NTIRE 2023 quality assessment of video enhancement challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[28] 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, 7

[29] 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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2019. 5

[30] 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 *International Conference on Computer Vision*, pages 367–376, 2021. 5
[31] 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. 5
[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015. 8
[33] Tim Seizinger, Marcos V Conde, Manuel Kolmet, Tom E Bishop, and Radu Timofte. Efficient multi-lens bokeh effect rendering and transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2
[34] Alina Shutova, Egor Ershov, Georgy Perevozchikov, Ivan A Ermakov, Nikola Banic, Radu Timofte, Richard Collins, Maria Efimova, Arseniy Terekhin, et al. NTIRE 2023 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2
[35] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7968–7977, 2020. 1, 2
[36] Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. Wsrd: A novel benchmark for high resolution image shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2
[37] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. NTIRE 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2
[38] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17694–17703, 2022. 1, 4, 5
[39] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2
[40] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11534–11542, 2020. 8
[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 5
[42] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2
[43] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 5
[44] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. CT2: Colorization transformer via color tokens. In *European Conference on Computer Vision*, 2022. 1, 5, 6
[45] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *International Conference on Computer Vision*, 2021. 1, 2, 6
[46] Menghan Xia, Wenbo Hu, Tien-Tsin Wong, and Jue Wang. Disentangled image colorization via global anchors. *ACM Transactions on Graphics*, 41(6):204:1–204:13, 2022. 3, 4
[47] Chengpeng Chen Xiaojie Chu, Liangyu Chen and Xin Lu. Improving image restoration by revisiting global information aggregation. *arXiv preprint arXiv:2112.04491*, 2021. 5, 6
[48] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9363–9372, 2020. 1
[49] Ren Yang and Radu Timofte. NTIRE 2021 challenge on quality enhancement of compressed video: Dataset and study. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2
[50] Ren Yang, Radu Timofte, et al. AIM 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. In *European Conference on Computer Vision Workshops*, 2022. 2
[51] Ren Yang, Radu Timofte, et al. NTIRE 2022 challenge on super-resolution and quality enhancement of compressed video: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 2
[52] Ren Yang, Radu Timofte, Xin Li, Qi Zhang, Lin Zhang, Fanglong Liu, Dongliang He, Fu Li, He Zheng, Weihang Yuan, et al. AIM 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. In *European Conference on Computer Vision*, pages 174–202, 2023. 5
[53] Yixin Yang, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, Jinhui Tang, and Jinshan Pan. BiSTNet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *arXiv preprint arXiv:2212.02268*, 2022. 1, 3
[54] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 5
[55] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, et al. NTIRE 2023 challenge on hr depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2
[56] Eduard Zamfir, Marcos V Conde, and Radu Timofte. Towards real-time 4k image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[57] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5718–5729, 2022. 4

[58] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8061, 2019. 1, 5, 6, 7

[59] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666, 2016. 1

[60] Yu Zhang, Siqi Chen, Mingdao Wang, Xianlin Zhang, Chuang Zhu, Yue Zhang, and Xueming Li. Temporal consistent automatic video colorization via semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 5

[61] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, et al. NTIRE 2023 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2