RTTLC: Video Colorization with Restored Transformer and Test-time Local Converter

Jinjing Li[13]* Qirong Liang[1]* Qipei Li[1]* Ruipeng Gang[23]† Ji Fang[23]
Chichen Lin[1] Shuang Feng[1] Xiaofeng Liu[1]
[1]Communication University of China, Beijing 100024, China
[2]Academy of Broadcasting Science, NRTA, Beijing 100866, China
[3]UHDTV Research and Application Laboratory, NRTA, Beijing 100176, China
1792418414@qq.com gangruipeng@abs.ac.cn

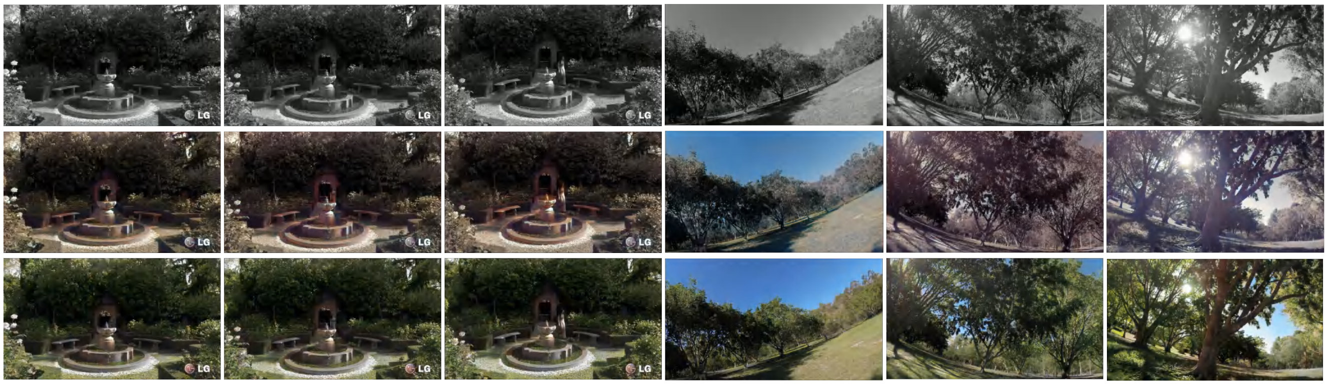

Figure 1. The video is colored by our method. First row: original video frames. Second row: the colored frames by DeOldify (baseline). Third row: the colored frames by our RTTLC.

Abstract

Video colorization is a highly challenging and ill-posed problem that suffers from severe flickering artifacts and color distribution inconsistency. To resolve these issues, we propose a Restored Transformer and Test-time Local Converter network(RTTLC). Firstly, we introduce a Bidirectional Recurrent Block and a Learnable Guided Mask to our network. This leverages hidden knowledge from adjacent frames that include rich information about occlusion, resulting in significant enhancements in visual quality. Secondly, we integrate a Restored Transformer that enables the network to utilize more spatial contextual information and capture multi-scale information more accurately. Thirdly, during inference, we utilize the Test-time Local Converter(TLC) strategy to alleviate distribution shift and enhance the performance of the model. Experimental results show good performance of FID and CDC. Notably, RTTLC achieves second prize in both tracks of the NTIRE23 video colorization challenges.

1. Introduction

The rapid advancement of digital technology increases demand for high quality video materials [30]. Many legacy videos in grayscale cannot meet the demands of contemporary audiences, due to the low resolution [21] and distracting artifacts. To address this issue, video colorization [25] has emerged as a novel technique that seeks to transform grayscale videos into fully colored ones. This process has great potential to enhance the visual quality of legacy videos and provide an immersive experience for audiences. Furthermore, colorization can also facilitate other computer vision tasks such as detection [3], tracking [32], and action recognition [14] in videos.

Manual colorization is an expensive and meticulous work, which requires lots of effort. Hence, automatic colorization methods are necessary. Recently, learning-based methods have been explored and shown remarkable progress. For instance, Richard Zhang *et al.* [31] propose a convolutional neural network-based video colorization algorithm that utilizes the temporal continuity between adjacent frames in a video. However, it is limited by the dataset used during training and may introduce color artifacts. Ja-

†corresponding author
*equal contribution

son Antic *et al*. [1] propose the DeOldify, which employs generative adversarial networks and recurrent neural networks to perform automatic colorization. Liu *et al*. [18] propose a Temporally Consistent Video Colorization framework(TCVC) that combines deep feature propagation and self-regularization learning to overcome the temporal inconsistency issue. BiSTNet [27] capitalizes on the semantic image prior to facilitating bidirectional temporal feature fusion. However, some researchers argue that classic optical flow algorithms could lead to considerable misalignment errors [20]. Meanwhile, the RTN [23] model employs a dual-stream architecture for deep learning. Despite these advancements, existing methods are still suffered from severe flickering artifacts and color distribution inconsistency [13].

In this paper, we propose a Restored Transformer and Test-time Local Converter network(RTTLC) for video colorization. Our approach evolved from RTN [23], and innovated on: (1) a Restored Transformer, and (2) a Test-time Local Converter(TLC).

The Restored Transformer is inspired by Restormer [28], which is originally used for image super-resolution. Our Restored Transformer presents a noteworthy innovation in its ability to capture a greater amount of multi-scale information during training. This is achieved through the utilization of the multiDconv head transposed attention (MDTA) module, which implicitly models global context by applying self-attention across channels instead of the spatial dimension. This results in linear complexity, as opposed to quadratic, enabling more efficient and effective processing.

The TLC is inspired by [8], which uses a method called local aggregation during testing to replace the information aggregation region from the entire spatial dimension to a local window. The input image is cropped into patches and fed into the model in the training phase. Conversely, most methods use the full-resolution image in inference [29], resulting in a train-test inconsistency problem. To better consider the potential distribution shift between the input images during model training and testing, we use TLC strategy to minimize this discrepancy during the inference phase. Additionally, we utilize a multi-model ensemble approach during inference to improve the model's performance on complex video clips.

We participate in both tracks in the NTIRE23 video colorization challenges, including color diversity evaluation and color distribution consistency evaluation. RTTLC is awarded second prize in both tracks, indicating its effectiveness and generalizability. In addition, we conduct comparative experiments on existing video colorization benchmarks and found that RTTLC outperforms state-of-the-art methods in various video restoration tasks.

Our main contribution can be summarized as follows:

1) We propose a new video colorization method named RTTLC to reduce color artifacts and flickering artifacts effectively. As shown in Fig. 1, our method achieves an even better visual performance than baseline.

2) We introduce the Restored Transformer which contains a multi-Dconv head transposed attention(MDTA) module that is capable of aggregating local and non-local pixel interactions.

3) We adopt the TLC which uses local aggregation during testing to handle spatial information. This not only leads to improvements in test-time performance but also enhances the consistency of reasoning during both training and testing.

4) Experimental results demonstrate that our method achieves good performance in terms of FID [9] and CDC [18], while also exhibits significant improvements in visual quality.

2. Related Work

Video Colorization. Recently, researchers have been using diverse neural network architectures based on deep learning for colorization [31]. Lei *et al*. [16] propose a fully automatic method of FAVC based on a self-regularization technique and a diversity-promoting term. Jason Antic *et al*. [1] propose DeOldify, using GAN and RNN to automatically colorization. Liu *et al*. [18] propose a temporally consistent video colorization framework(TCVC) with deep feature propagation and self-regularization learning. Wan *et al*. [23] present a Recurrent Transformer Network(RTN) to solve the mixed degradations of video by leveraging the temporal modeling of recurrent neural networks. However, some models may fail to differentiate between contaminants and actual frame content, resulting in undesirable flickering artifacts.

Vision Transformers. In video colorization, ViTs [22] are effective in video colorization due to their ability to capture long-range dependencies. However, the self-attention mechanism in Transformers can increase computational complexity [24]. To address this issue, Cao *et al*. [2] propose VSR-Transformer that uses the self-attention mechanism for better feature fusion in video SR. Swin Transformer [19] applies self-attention within local image regions. Furthermore, Liang *et al*. [17] design a Swin Transformer-based image restoration model SwinIR. However, these methods confine context aggregation to the local neighborhood, undermining the main advantage of using self-attention over convolutions.

Global Information Aggregation. Inconsistency between the input image distribution during model training and testing could be a potential issue for poor performance in real world [15]. Chen *et al*. [7] propose a Half Instance Normalization Block(HIN), which can divide the input

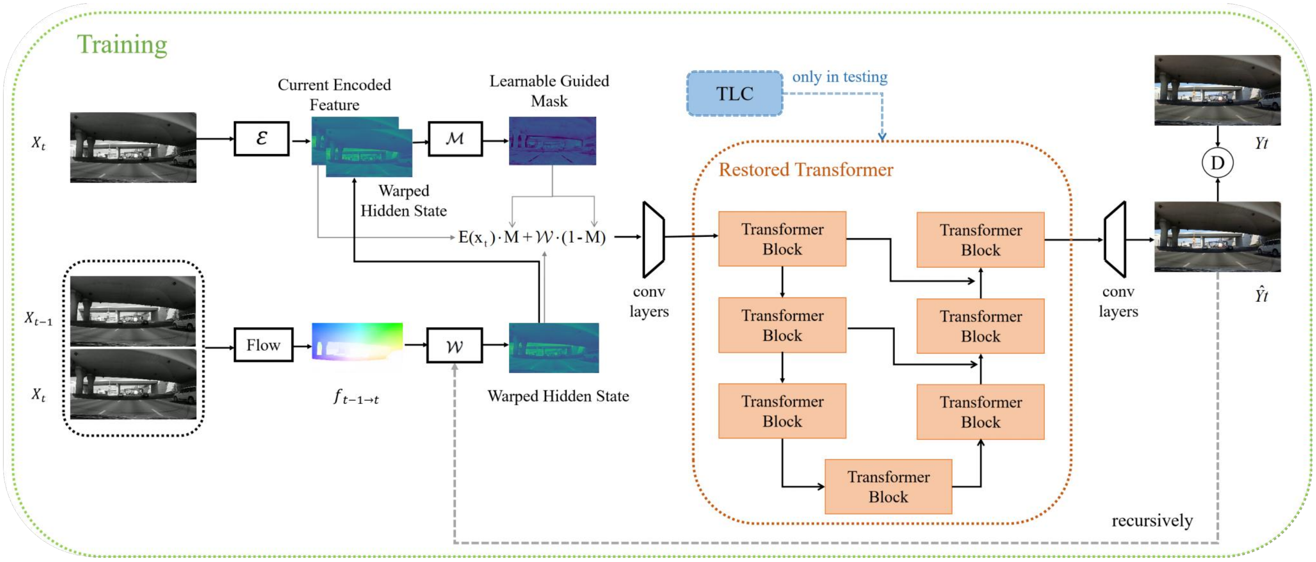

Figure 2. The Framework of Restored Transformer and Test-time Local Converter Network.

image into fixed-size patches. Chen *et al.* [6] employ an Image Processing Transformer(IPT) with multi-heads and multi-tails. The Test-time Local Converter(TLC) [8] can be employed to replace the information aggregation region from the entire spatial dimension to a local window. Overall, these methods give encouraging outcomes in tackling the aforementioned problems.

3. Method

3.1. Overview

Our algorithm consists of three modules: Bi-directional Recurrent Block, Learnable Guided Mask, and Restored Transformer. Assuming a video sequence $X = \{x_1, x_2, \cdots, x_T\}$, where T represents the length of a video. Firstly, the input video is processed by a Bi-directional Recurrent Block. It extracts features while obtaining sufficient temporal and spatial contextual information. Secondly, we use a soft mask guided by the optical flow method to locate spatial color change areas to suppress color artifacts. Finally, in order to improve the quality of video detail coloring, we utilize a Restored Transformer. Additionally, we adopt several strategies. Test-time Local Converter(TLC) can ensure training-test consistency. Multi-Model Ensemble Testing can improve the model's FID and CDC scores significantly. The overall system pipeline is shown in Fig. 2.

3.2. Network Structure

3.2.1 Bi-directional Recurrent Block

In order to suppress the flickering artifacts and fully utilize temporal information, we refer to [23] and introduce the Bi-directional Recurrent Block into our approach, which contains a forward hidden state and a backward hidden state. We estimate the optical flows $f_{t-1\to t}$ between x_{t-1} and x_t. The previous state s_{t-1} is aligned to time t by a warping function $\mathcal{W}(\cdot)$. x_t is extracted by the encoder $\mathcal{E}(\cdot)$, and $\mathcal{F}$ merge the states between history and current feature.

The forward hidden state $s_{t\uparrow}$ and backward hidden state $s_{t\downarrow}$ are obtained as Eqs. (1) and (2), respectively.

$$s_{t\uparrow} = \mathcal{F}\uparrow(\mathcal{E}(x_t), \mathcal{W}(s_{t-1}, f_{t-1\to t})) \tag{1}$$

$$s_{t\downarrow} = \mathcal{F}\downarrow(\mathcal{E}(x_t), \mathcal{W}(s_{t+1}, f_{t+1\to t})) \tag{2}$$

Finally, we input bidirectional temporal states to the Restored Transformer $\mathcal{R}(\cdot)$, and incorporate through the decoder $\mathcal{D}(\cdot)$.

$$y_t = \mathcal{D}(\mathcal{R}(s_{t\uparrow}), \mathcal{R}(s_{t\downarrow})) \tag{3}$$

3.2.2 Learnable Guided Mask

To ensure temporal consistency, we refer to [23] and introduce the Learnable Guided Mask. This is originally designed for scratch removal, which utilizes a learnable soft mask to locate scratch areas. However, we find that this module can position color changes well according to the motion of adjacent frames, thereby suppressing flickering artifacts. The mask M can be expressed as Eq. (4).

$$M = \mathcal{M}(\mathcal{E}(x_t), \mathcal{W}(s_{t-1}, f_{t-1\to t})) \tag{4}$$

Where $\mathcal{M}(\cdot)$ is a shallow convolutional neural network. We aggregate the temporal priors and current frames as Eq. (5).

$$\mathcal{F} = \mathcal{E}\left(\mathrm{x}_t\right) \cdot M + \mathcal{W}(s_{t-1}, f_{t-1\to t}) \cdot (1 - M) \tag{5}$$

3.2.3 Restored Transformer

In order to leverage more spatial contextual information and capture more multi-scale information, we introduce the Restored Transformer [28]. Compared with the original Swin Transformer [19], Restored Transformer has stronger spatial pixel learning ability and a larger receptive field. This is attributed to its proposed Multi-Dconv Head Transposed Attention(MDTA), which has linear complexity.

MDTA uses a depthwise separable convolution [10] to aggregate pixel-level cross-channel context information for Query(Q), Key(K), and Value(V) vectors. Then, Q and K are reshaped into a transposed attention map of size $\mathbb{R}^{C} \times \mathbb{R}^{C}$, instead of the huge regular attention map of size $\mathbb{R}^{H\times W} \times \mathbb{R}^{H\times W}$, where C represents the number of channels. This approach reduces computational complexity significantly. The above process can be defined as Eqs. (6) and (7).

$$\hat{X} = W_p Attention(\hat{Q}, \hat{K}, \hat{V}) + X \tag{6}$$

$$Attention(\hat{Q}, \hat{K}, \hat{V}) = \hat{V} \times softmax(\frac{\hat{K} \times \hat{Q}}{\alpha}) \tag{7}$$

In these formulas, $W_p(\cdot)$ is the 1×1 point-wise convolution of depthwise separable convolution. X and $\hat{X}$ represent the input and output feature maps. $\hat{Q}$, $\hat{K}$, and $\hat{V}$ are obtained by reshaping tensors from the original size $\mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$. The parameter α is a scaling factor.

After several transformer blocks, Restored Transformer can effectively perceive global context information and by taking advantage of its large receptive field.

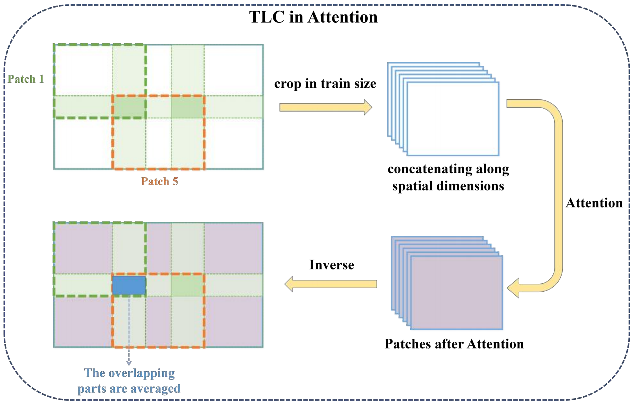

Figure 3. The process of TLC in attention, the green part indicates overlapping areas.

3.3. Tips and Tricks

3.3.1 Test-time Local Converter

In order to ensure model train-test consistency, we introduce Test-time Local Converter(TLC) [8], which converts global operations to local ones during testing. Researchers usually divide images into patches during training, but input the entire image during testing. This inconsistency between training and testing can greatly reduce the model's performance. To address this issue, We use TLC in the attention layer of Restored Transformer, and the process is shown in Fig. 3.

Firstly, TLC uses a sliding window to crop the feature map and get n patches. The size of the window depends on the patch size during training. Then, the attention layer is calculated on each patch. Finally, the output of each patch is merged into an entire feature map. Because the computation is independent in each window, unnatural dividing lines may appear at the window's boundaries. To mitigate this situation, there is a partial overlap between each window, and the overlapped areas are averaged when output. The process of local information aggregation can be formulated as Eq. (8).

$$\Psi(X, f) = \overset{n}{\underset{i,j}{\cup}} f(X_{i,j}) \tag{8}$$

Where $X_{i,j}$ indicates the $(i,j)^{th}$ patch. f indicates the calculation of the attention layer. $\overset{n}{\underset{i,j}{\cup}}$ indicates the operation. The size of $\Psi(X, f)$ is the same as the output of attention layer without TLC.

We find that the artifacts of edge splicing are reduced with little additional computational. Meanwhile, TLC can enable the model to generate more color textures with no additional computational cost.

3.3.2 Multi-Model Ensemble Testing

Since the learning ability of the model is different on different data, we finetune them using different training sets. During testing, we select the best-performing model on the specific category of the test set. In the competition, we use six models in total, which significantly improve the model's score in testing.

3.4. Loss Function

The entire loss function L_{total} contains L_1 loss, perceptual loss, and Spatial-Temporal Adversarial Loss. The Spatial-Temporal Adversarial Loss refers to the Temporal-PatchGAN [5]. It can enhance perceptual quality and spatial-temporal coherence. The discriminator D is aimed to distinguish between real and fake spatial-temporal fea-

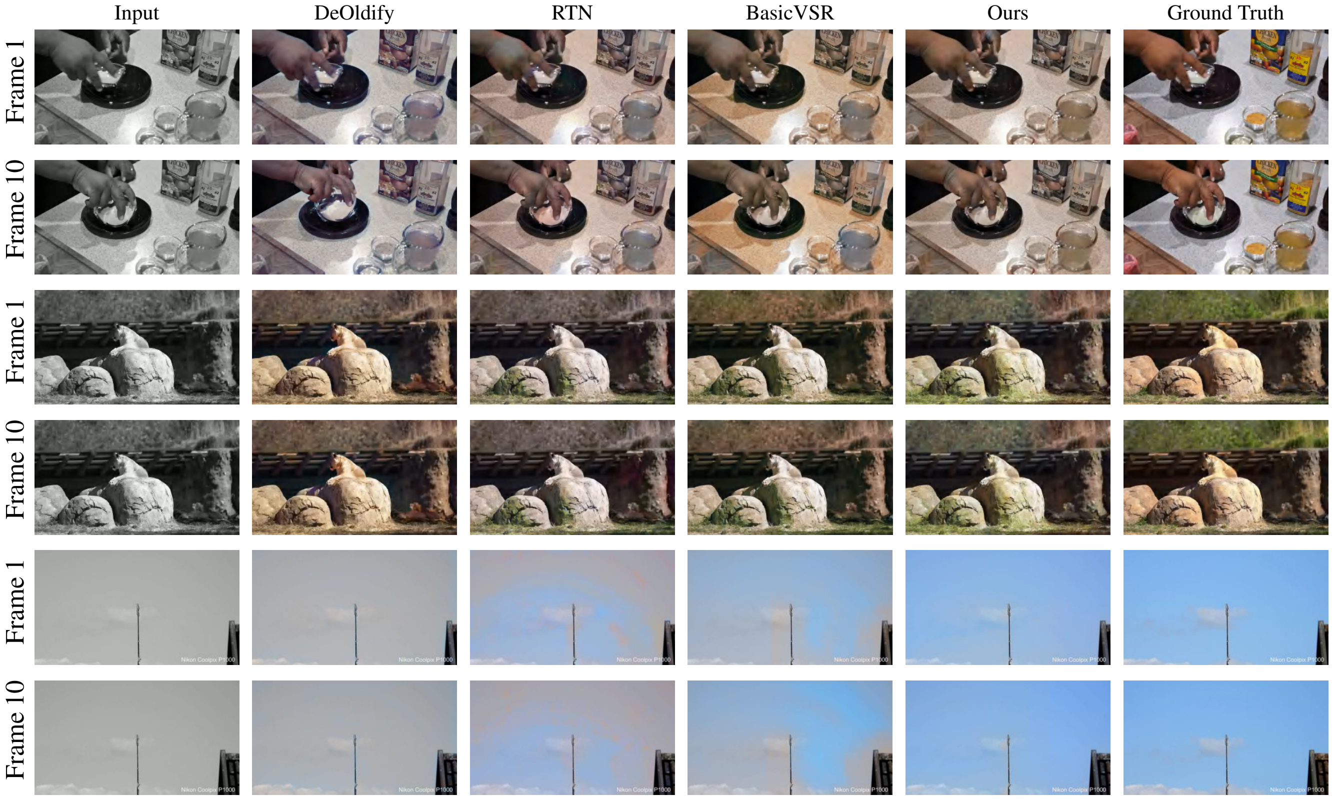

Figure 4. Qualitative comparison on LDV(NTIRE23 Video Colorization Challenge validation set) for video coloring.

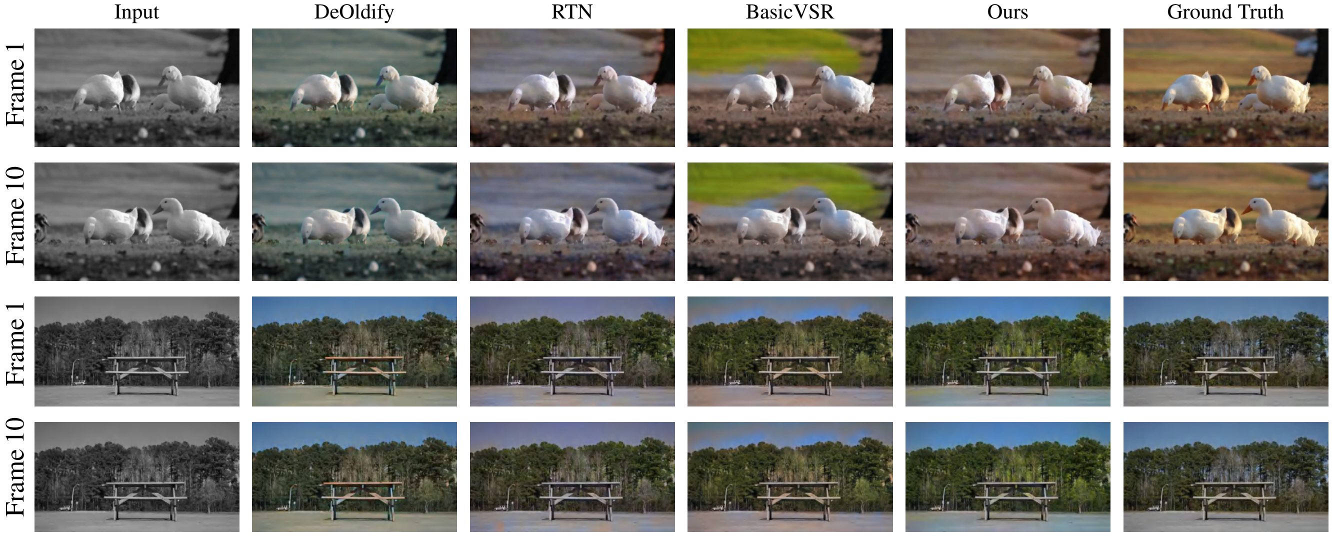

Figure 5. Qualitative comparison on the Videvo20 dataset for video coloring.

tures, using hinge loss as shown in Eqs. (9) and (10).

$$\mathcal{L}_D = \mathbb{E}y \sim Y[\text{ReLU}(1-D(y))]+\mathbb{E}\hat{y} \sim \hat{Y}[\text{ReLU}(1+D(\hat{y}))] \tag{9}$$

$$\mathcal{L}_G = -\mathbb{E}_{y \sim Y}[D(y)] \tag{10}$$

Here, Y and $\hat{Y}$ are the output video and ground truth video, respectively. The expression is as Eq. (11).

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_1 + \lambda_{per} \mathcal{L}_{per} + \lambda_{adv} \mathcal{L}_{adv} \tag{11}$$

4. Experiments

4.1. Training Datasets and Details

Training datasets. Our dataset is obtained from the Large-scale Diverse Video(LDV) dataset [26], which is the official dataset of NTIRE23 Video Colorization Challenge [11]. The dataset contains 240 high-quality videos and exhibits a high degree of diversity. Specifically, we select 200 color videos with a resolution of 960×536 as the training set. The validation set contains 15 videos. Additionally, we select 5 videos from the internet that are similar to the dataset as extra training data. The video frames are converted to grayscale using 'cv2.cvtColor()'.

Training details. In our joint loss function, $(\lambda_1, \lambda_{per}, \lambda_{adv}) = (1.0, 1.0, 0.01)$. We use Adam optimizer [12] with $(\beta_1, \beta_2) = (0.9, 0.99)$. To facilitate convergence, we use Cosine Annealing. The learning rate for both the generator and discriminator are set to 1×10^{-4} for the first 10 epochs, and then linearly decays to zero. We train our model on the 200 videos of LDV for 8 epochs. The patch size is 192×192, and the number of input frames is 8. Subsequently, we finetune our model with 60 videos from the training set that is similar to the validation set and 5 extra dataset. During finetuning, the patch size is 256×256 and the number of input frames is 5, with a total of 50,000 iterations. Our model is implemented using PyTorch and trained for 4 days on four NVIDIA GeForce RTX 3090 GPUs.

4.2. Comparison with Other Methods

We compare our proposed method with several state-of-the-art approaches, including DeOldify [1], RTN [23] and BasicVSR++ [4]. For a fair comparison, all approaches are evaluated on the LDV validation set and Videvo20 [18]. Qualitative results on the validation set and Videvo20 are presented in Fig. 4 and Fig. 5, respectively. We use FID [9] and CDC [18] as evaluation metrics. The FID measures the similarity between the colorized videos and ground truth videos. The CDC evaluates the temporal consistency of the colorization results over time. In addition, we compare the number of parameters and FLOPs. It can be seen from Tab. 1 that the proposed method is almost optimal in both evaluation indexes, which is slightly worse than DeOldify at FID of Videvo20.

4.3. Ablation Study

4.3.1 Restored Transformer

We retrain on our training set using the original training setup of RTN [23], which uses Swin Transformer as the coloring module. As shown in Fig. 6, the Restored Transformer produces more vibrant colors in the colored videos. Swin Transformer struggles to accurately color the contents of the images. The FID and CDC scores are shown in Tab. 2, where Restored Transformer's score increased.

Method	LDC validation set		Videvo20		Params	FLOPs
	FID↓	CDC↓	FID↓	CDC↓		
DeOldify	47.1719	0.003475	37.1743	0.002584	218.22M	140.82G
BasicVSR++	54.4334	0.005795	53.6191	0.005224	6.98M	243.10G
RTN	54.7994	0.003018	48.9520	0.002113	6.17M	162.61G
RTTLC(ours)	43.3332	0.002594	48.4203	0.002092	17.80M	164.50G

Table 1. Comparison with SOTA methods. Red and blue indicate the best and the second best performance, respectively. The input size is (3, 256, 256) when calculating Params and FLOPs.

	(A)	(B)	RTTLC
Restored Transformer	×	✓	✓
TLC	×	×	✓
FID↓	54.7994	43.6007	43.3332
CDC↓	0.003018	0.002921	0.002594

Table 2. Ablation experiment for the Restored Transformer and TLC.

4.3.2 Test-time Local Converter

We conduct a comparative experiment on whether to use TLC. As shown in Fig. 6, TLC makes the colored videos more realistic and closer to the ground-truth(GT) videos. The FID and CDC scores are shown in Tab. 2. This indicates the importance of consistent data distribution between training and testing.

4.3.3 Result of NTIRE23 Video Colorization Challenge

We participated in NTIRE23 Video Colorization Competition which contains Track 1 of FID and Track 2 of CDC, and got the second prize on both tracks. The results of our competition are shown in Tabs. 3 and 4.

Team	Author	FID↓	CDC↓
NJUSTer	Yixin Yang	21.5372	0.001717
CUCPLUS(ours)	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

Table 3. Results of the top5 methods in the NTIRE 2023 Video Colorization Challenge Track 1.

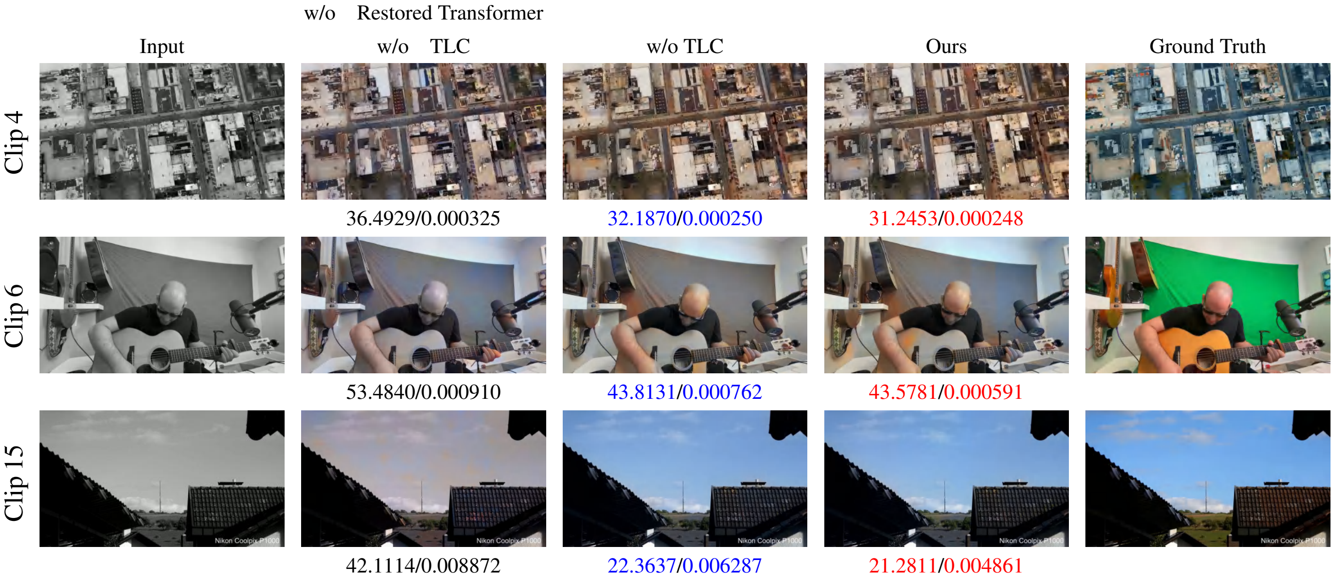

Figure 6. The result of ablation experiments(FID/CDC). Red and blue indicates the best and the second best performance, respectively. First column: input frames. Second column: the output without TLC and Restored Transformer. Third column: the output without TLC. Fourth column: the output of the complete network. Fifth column: the Ground Truth frames.

Team	Author	FID↓	CDC↓
MiAlgo	Shuai Liu	54.7238	0.000819
CUCPLUS(ours)	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

Table 4. Results of the top5 methods in the NTIRE 2023 Video Colorization Challenge Track 2.

5. Conclusion

In conclusion, video colorization is a complex task that has the potential to enhance the quality of legacy videos. While learning-based methods have made remarkable progress, they still suffer from flickering artifacts and temporal inconsistency. To address these issues, we propose RTTLC that uses both spatial and temporal information to achieve accurate and high quality colorization results. Our method incorporates a Restored Transformer to aggregate local and nonlocal pixel interactions. In addition, we introduce the TLC to address the potential distribution shift between the input images during model training and testing. We also participate in the NTIRE23 video colorization challenges and conduct comparative experiments, which demonstrate the effectiveness and generalizability of our method. Overall, our RTTLC reduces color artifacts and represents a significant advancement in video colorization.

6. Acknowledgement

This paper is funded by the ”Video Super-Resolution Algorithm Design and Soft- ware Development for Face Blur Problem” (JBKY20220210), “Research and Simulation Experiment of Lightweight Sports Event Remote Production System” (ZZLX-2020-001) and “Research on intelligent synthesis technology of video scenes” (JBKY20230100) projects of the Academy of Broadcasting Science, National Radio and Television Administration.

References

[1] Jason Antic. Deoldify. https://github.com/jantic/DeOldify, 2019.

[2] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022.

[5] 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.

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021.

[7] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021.

[8] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 53–71. Springer, 2022.

[9] 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.

[10] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

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

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018.

[14] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6874–6883, 2017.

[15] Nam-Yong Lee. Block-iterative richardson-lucy methods for image deblurring. *EURASIP Journal on Image and Video Processing*, 2015:1–17, 2015.

[16] 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.

[17] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[18] 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.

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[20] Syed Tafseer Haider Shah and Xiang Xuezhi. Traditional and modern strategies for optical flow: an investigation. *SN Applied Sciences*, 3:1–14, 2021.

[21] Bin Sheng, Hanqiu Sun, Marcus Magnor, and Ping Li. Video colorization using parallel optimization in feature space. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(3):407–417, 2013.

[22] 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.

[23] Ziyu Wan, Bo Zhang, Dongdong Chen, and Jing Liao. Bringing old films back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17694–17703, 2022.

[24] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022.

[25] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-based architecture for fast deep exemplar colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9363–9372, 2020.

[26] Ren Yang. Ntire 2021 challenge on quality enhancement of compressed video: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2021.

[27] 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.

[28] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[29] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021.

[30] 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.

[31] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

[32] Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4591–4600, 2019.