DOI: 10.19884/j.1672-5220.202501012

ColorAlignNet: a Reference-Based Video Colorization Network with Temporal Aggregation

ZHU Wenzhi[1,2], WANG Tong[1,2*]
1. College of Information Science and Technology, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitized Textile & Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China

Abstract: Video colorization is an important technique to breathe life back into old movies. While current colorization methods work well on still images and low-motion video data, they often struggle with complex dynamic scenes. To address this problem, this study proposes ColorAlignNet, a reference-based video colorization network with temporal aggregation. The network uses source-reference attention to propagate color information from reference frames to grayscale frames, guaranteeing color accuracy, and uses deformable convolution to align features of adjacent frames to enhance temporal consistency. Finally, we use the cyclic transformer module to reconstruct the final prediction results. Extensive experimental results demonstrate that ColorAlignNet achieves excellent performance on the DAVIS and Videvo datasets, outperforming other state-of-the-art methods on both the learned perceptual image patch similarity (LPIPS) and color distribution consistency (CDC) metrics.

Keywords: deformable convolution; video colorization; Swin-transformer
CLC number: TP391.4 **Document code**: A
Article ID: 1672-5220(2026)02-0094-09

Open Science Identity (OSID)

0 Introduction

Video colorization is a rapidly growing field, widely used in film restoration, historical material enhancement, and creative media production, which has attracted a lot of attention. By turning grayscale lenses into color, video colorization has breathed life back into archival videos, making them more appealing to modern audiences. In recent years, a large number of methods for image colorization[1-4] have been proposed, and remarkable progress has been made. Compared with still images, video colorization faces unique challenges due to its temporal consistency. The main goal is to ensure the temporal consistency, achieve vivid natural colors, and maintain color stability in a variety of visual conditions.

In general, two video colorization methods have been proposed to solve the above problems. The first method is to consider temporal consistency in a colorization network, which requires a more complex network structure and greater domain knowledge. For example, fully automatic video colorization (FAVC)[5] proposes a variety of self-regularization methods for automatic video colorization. However, the method places too much emphasis on temporal consistency, which results in less favorable color vividness. In the context of the video colorization framework, while temporal consistency is undoubtedly a crucial aspect, it is not the only determining factor in achieving compelling visual results. The second method is to color the gray video frame by frame[6-9], and then post-process the color result. This method can achieve good temporal consistency. However, due to the differences in color information between adjacent frames, the post-processing phase is essentially equivalent to the reconstruction of the color video frames in the output. This can lead to significant differences between the predicted results and the colorization results, making the results uncontrollable.

In order to solve the conflict between colorization performance and temporal consistency, we propose a colorization network, ColorAlignNet, based on the convolutional neural network (CNN) and transformers[10] of reference frames. The network also belongs to frame-by-frame colorization methods. Unlike traditional post-processing networks, it improves temporal consistency and enhances controllability by introducing temporal alignment and color transfer strategies. Specifically, for the colorization problem, we design a color distribution network, ColorPreb. The network uses a source-reference attention[6] to make grayscale video frames pay close attention to the color of the reference frames, and supports input of different numbers of reference frames, which is very helpful for specific colorization tasks.

For the temporal inconsistency problem that is common in video tasks, we design a temporal aggregation network, FlexAlign, using deformable convolution[11]. The network aligns the features of adjacent frames to the current frame very well. Unlike feature alignment

Received date: 2025-01-19
* Correspondence should be addressed to WANG Tong, email: wangtong@dhu.edu.cn
Citation: ZHU W Z, WANG T. ColorAlignNet: a reference-based video colorization network with temporal aggregation[J]. *Journal of Donghua University (English Edition)*, 2026, 43(2): 94-102.

methods that use optical flow[12], deformable convolution uses fewer computational resources and has better results in processing large moving video data. To compensate for the loss of feature alignment and colorization, ColorAlignNet uses Swin-transformer[10] to compensate for unaligned depth features through self-attention.

We have extensively evaluated our method on the DAVIS[13] and Videvo[14] datasets and found that our method has more favorable results and outputs more satisfying video frames compared to existing methods. In addition, ColorAlignNet using reference frames means that our model is more controllable compared to fully automatic colorization methods, which is helpful when we want to produce satisfactory results in a specific colorization task.

The main contributions of this study are as follows.

1) We propose a reference-based video colorization network, ColorAlignNet, which colorizes grayscale videos using reference color frames. Through the designed ColorPrep module, the network extracts key color information from the reference frames. It accurately transfers this information to the target grayscale frames, generating color videos with temporal consistency.

2) We design a temporal feature aggregation network based on deformable convolution to replace traditional optical flow[12] alignment. This method aligns features between adjacent frames more robustly. It avoids errors caused by large motion, occlusion, or edge discontinuities, improving temporal consistency in video colorization.

3) The network also integrates a Swin transformer-based[10] reconstruction module to capture wider contextual information. This module ensures structural consistency and color continuity across frames. It alleviates color drift and flicker in dynamic scenes, improving video restoration and enhancement.

1 Methodology

By inputting grayscale video frames and reference frames, the entire process of our method is performed in CIE-Lab space for channels a and b. The goal is to obtain color video frames with good vividness and temporal consistency.

1.1 ColorPrep

We design a color pre-processing module, ColorPrep, using source-reference attention[6]. By calculating the similarity between the grayscale frame and reference frame features, the attention mechanism dynamically selects the most relevant color information for propagation.

First, for a given grayscale frame, we use it and the reference frame as inputs for color propagation, and the method accepts an arbitrary number of reference color frames, which the model uses as cues during the colorization process. The input to the source-reference attention consists of two volumetric feature maps of different lengths, one representing the grayscale frame and the other representing the reference frame, allowing the model to capture the non-local similarity between the grayscale frame and the reference frame.

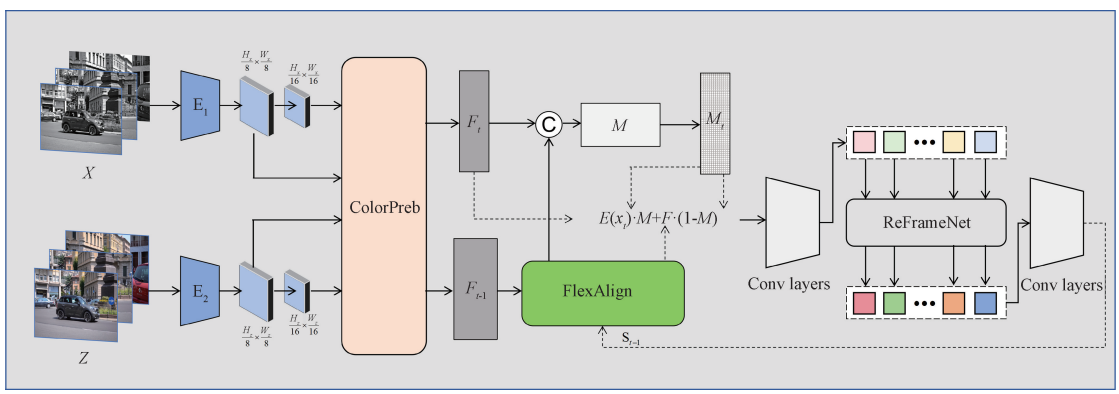
Specifically, given consecutive grayscale video frames $X = \{x_1, x_2, \cdots, x_n\}$, where $x_t \in \mathbf{R}^{H_x \times W_x \times C_x}$ denotes the tth frame, $H \times W$ denotes the spatial dimensions, C represents the number of channels, and the reference video frame $Z = \{z_1, z_2, \cdots, z_m\}$, where $z_t \in \mathbf{R}^{H_z \times W_z \times C_z}$. We use a convolutional encoder to map the grayscale video frame x_t and the reference video frame z_t to $\frac{H}{16} \times \frac{W}{16}$ spatial resolution after four downsamplings.

The architecture of ColorAlignNet is shown in Fig. 1. As shown in Fig. 1, ColorAlignNet is divided into three core parts: a) ColorPrep; b) FlexAlign; c) ReFrameNet.

Passing the preprocessed feature maps into the ColorPrep network, we aim to propagate the color information in the reference frame to the grayscale frames through an attention mechanism. The attention layer L_{SR} can be defined as

$$L_{\mathrm{SR}} = z_2 + \gamma d(x_1 \cdot \mathrm{softmax}(z_2, x_1)), \quad (1)$$

where $\gamma \in \mathbf{R}$ is a learnt parameter; d represents a norm function.



E_1, E_2— two convolutional encoders; C—concatenation; M—mask; Conv layers—convolutional layers; F_t, F_{t-1}—feature maps at time t and $t-1$, respectively; M_t—mask at time t; S_{t-1}—state feature at time $t-1$; $E(\cdot)$—function of the convolutional encoder; F—aligned temporal feature map.

Fig. 1 Architecture of ColorAlignNet

1.2 FlexAlign

Different from the traditional conventional feature alignment by optical-flow[12], FlexAlign utilizes deformable convolution[11] to complete the feature alignment and aggregation operations. Specifically, given an input feature map F_{t+i}, $i \in [-k, k]$, with k being the distance between the current frame and adjacent frames. To generate the required aligned feature map, we use bilinear interpolation[15] to perform × 2 upsampling (× 2 up) on the high-dimensional feature map F_t after ColorPrep preprocessing and colorization. In the process of alignment, the offsets Δp_n are obtained through a convolutional layer that learns adaptively from input features, the Δp_n allows the convolutional kernel to dynamically adjust its sampling point positions during feature alignment, and it can be defined as

$$\Delta p_n = f([F_{t+1}, F_t]), \ i \in [-k, k], \tag{2}$$

where $[\cdot, \cdot]$ denotes the concatenation operation.

Figure 2 illustrates the frame alignment process by deformable convolution. Given a preprocessed color frame F_t, the frame at time $t+i$ is aligned to frame t by deformable convolution. As shown in Fig. 2, after obtaining Δp_n, it is applied to the adjacent frames F_{t+i}, $i \in [-k, k]$, and a convolution is performed on the aligned adjacent frames to output the final aligned feature map F'_{t+i}, with the expression as

$$F'_{t+i} = \sum_{n=1}^{n} w(p_n) \cdot F_{t+i}(p_n + \Delta p_n) + b(p_n), \tag{3}$$

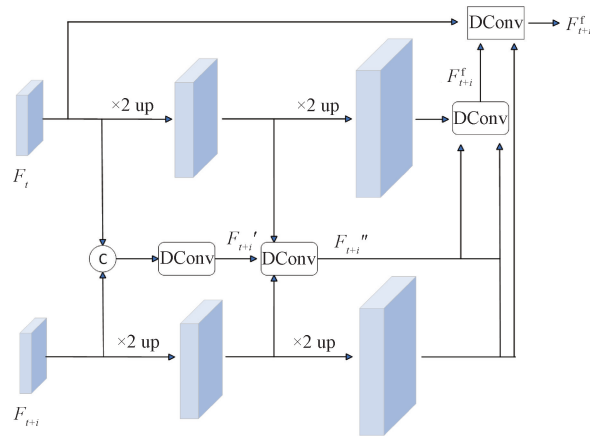
where $\omega(p_n)$ represents the weight at position k; $b(p_n)$ is the learnable parameter. Next, we feed the learned offset and the aligned reference feature map of this layer F'_{t+i} into the upsampled feature map. The offset for this layer is not only derived from the two input feature maps of this layer, but also from the offset of the previous layer. Moreover, the aligned feature map output by this layer depends not only on the output of the deformable convolution of this layer but also on the aligned feature map of the previous layer. The specific expressions are

$$\Delta p'_{t+i} = \mathrm{h}([F_{t+i}\uparrow, F_t\uparrow], \Delta p_n\uparrow), \tag{4}$$

$$F''_{t+i} = \mathrm{g}(\mathrm{D}(F_{t+1}, \Delta p'_{t+i}), F'_{t+i}\uparrow), \tag{5}$$

where $\uparrow$ represents × 2 upsampling using bilinear interpolation; $\mathrm{D}(\cdot)$ represents deformable convolution; $\mathrm{g}(\cdot)$ and $\mathrm{h}(\cdot)$ represent regular convolution. Repeat this operation to send the aligned feature map of the current layer to the next layer, thereby obtaining the next aligned feature map F^{f}_{t+i}.

After three stages of feature alignment across different dimensions, to further focus the adjacent frames on the features of the current frame, we perform another alignment on the feature maps F_{t+i} and F^{f}_{t+i} after two rounds of × 2 upsampling to get the final aligned features $F^{\mathrm{f}}_{t+i}{}'$, further refining the aligned features. Performing deformable convolution on different feature maps results in more complex transformations, allowing FlexAlign to learn how to align adjacent frames to the current frame under complex conditions.



DConv—deformable convolution.

Fig. 2 Frame alignment by deformable convolution

1.3 ReFrameNet

During the process of colorization and feature alignment, we have problems such as color propagation errors and loss of edge detail, which can lead to the generation of less desirable shading results. To address these problems, we propose a transformer-based image reconstruction network, ReFrameNet.

Specifically, in order to further learn the time information, the result of FlexAlign processing $F^{\mathrm{f}}_{t+i}{}'$ is connected with the previously recovered color frames s_{t-1} to calculate the soft mask between the current frame and the previous frame. The current frame and the time prior are weighted and mixed to combine the information of different time frames smoothly. The soft mask computing network M_t is denoted as

$$M_t = M[F^{\mathrm{f}}_{t+i}{}', \mathrm{D}(s_{t-1}, F^{\mathrm{f}}_{t+i})], \tag{6}$$

$$d_t = M_t \cdot F^{\mathrm{f}}_{t+i} + (1 - M_t) \cdot F^{\mathrm{f}}_{t+i}{}', \tag{7}$$

where d_t indicates the result of aggregating the temporal priori and the current frame.

After aggregating the temporal information of preceding frames, we reconstruct the results into vivid and temporally consistent color frames. Considering the local receptive field of CNNs, which limits their ability to model global context and long-range pixel dependencies, we utilize an efficient transformer network for reconstruction. Specifically, we adopt the Swin-transformer[10], which employs a sliding window attention mechanism. This divides the input feature map into multiple local windows and computes attention within each window, significantly reducing computational costs. To address the limitation of local window attention in capturing global dependencies, a shifted window mechanism enables cross-window interactions and smooth information flow. More specifically, the feature map is divided into non-overlapping $N \times N$ windows, within

which tokens are projected into query Q, key K, and value V by using multilayer perceptrons (MLPs). The attention is then calculated as

$$\mathrm{Atten}(Q, K, V) = \mathrm{Softmax}(QK^{\mathrm{T}}/\sqrt{d} + B)V, \quad (8)$$

where B is a learnable relative position embedding. Through the operation of ReFrameNet, we succeeded in generating the desired results, and the effectiveness of our module would be specifically analyzed in the ablation experiment.

1.4 Loss function

To better constrain the model training, we use three loss functions, $L1$ loss[16], perceptual loss[17] and discriminator loss[18], to train the video colorization model. The $L1$ loss is used to directly measure the pixel-level difference between the generated frames and the real frames to ensure the accuracy of the generated color; the perceptual loss is computed by the middle-layer features of the pre-trained visual geometry group (VGG) network to ensure that the generated results are semantically similar to the real frames at a high level; the discriminator loss helps the model generate more realistic video frames through adversarial training, thus improving visual quality. Combining these three loss functions, our model can generate high-quality colorization videos while maintaining structural consistency. The $L1$ loss L_1 is denoted as

$$L_1 = \frac{1}{T}\sum_{t=1}^{T} \| x_t - s_t \|. \quad (9)$$

Perceptual loss L_{perc}:

$$L_{\mathrm{perc}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{p\in P} \lambda_p \| \Phi_p(x_t) - \Phi_p(s_t) \|, \quad (10)$$

where Φ_p is the activation value of the pth layer of the pretrained network; λ_p is the weight of the layer.

Discriminator loss L_{disc}:

$$L_{\mathrm{disc}} = -[\mathrm{E}_{x_t}[\lg P(x_t)] + \mathrm{E}_{s_t}[\lg(1 - P(s_t))]], \quad (11)$$

where P is a function of the discriminator, which outputs the probability value of the input sample, indicating the probability that the sample is a true sample.

Full objective Combining the above losses, our total loss function L_{total} is

$$L_{\mathrm{total}} = \lambda_1 L_1 + \lambda_{\mathrm{perc}} L_{\mathrm{perc}} + \lambda_{\mathrm{disc}} L_{\mathrm{disc}}, \quad (12)$$

where λ_1, λ_{perc}, and λ_{disc} are the weights of L_1, L_{perc}, and L_{disc}, respectively.

With the L_{total}, our model is not only capable of accurate color recovery at the pixel level, but also performs well in terms of overall visual effect. The weights are separately set as $\lambda_1 = 1$, $\lambda_{\mathrm{perc}} = 1$, $\lambda_{\mathrm{disc}} = 0.01$, which are experimentally verified to effectively balance the role of each loss in the training process, ensuring that the generated color video achieves a high standard of detail and structure.

2 Experiment and Analysis

Datasets To train and evaluate ColorAlignNet, we use the DAVIS[13] and Videvo[14] datasets. DAVIS provides finely labelled high-quality videos suitable for verifying the accuracy of the model, while Videvo contains diverse videos of realistic scenes to enhance the generalisation ability of the model. All videos are scaled to 384 × 224 pixels to balance computational efficiency with detail preservation. In addition, we perform grayscaling, frame sampling, and data enhancement to ensure the robustness and performance consistency of the model across a wide range of scenes and video qualities.

Evaluation metrics 1) We use the peak signal-to-noise ratio (PSNR)[19] to compare the difference in pixel values between the original image and the generated image to measure the degree of image distortion; the higher the PSNR value, the less distorted the image. 2) We use the structural similarity index measure (SSIM)[20] to simulate human perception of image structure to assess image quality. 3) Learned perceptual image patch similarity (LPIPS)[21] is used to compare the distance between the original image and the generated image in the feature space to assess the image quality. 4) Color distribution consistency (CDC)[22] is used to measure the color distribution consistency between the original video and the colorization video.

Implementation details We train the model using the PyTorch framework on a machine with four RTX-3090 GPUs, with a total training time of three days. The entire experiment is conducted in the CIE-LAB color space for processing video frames, which helps to handle luminance and color information more naturally and accurately. During training, we use the default Adam optimizer for 1 000 000 iterations. The Adam configuration parameters are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 2×10^{-4}. The batch size is set to 16.

To enhance the model's generalization ability for the color propagation task, we introduce a data augmentation strategy during the training phase. Before inputting the deep features into the network, we artificially add Gaussian noise or feature noise and deliberately weaken the color information. This is done to improve the model's robustness and color restoration ability in situations where color information is missing or incomplete.

2.1 Comparisons with state-of-the-art methods

We have compared our proposed method with state-of-the-art methods, including fully automatic video colorization[5, 23] and reference frame-based video colorization[6-7, 24]. The quantitative results evaluated on the DAVIS[13] and Videvo[14] datasets are shown in Tables 1 and 2, respectively. The best items and the second-best items are highlighted in bold and underlined, respectively; ↓ indicates that a lower evaluation metric value corresponds to better performance, while ↑

indicates that a higher evaluation metric value corresponds to better performance. As shown in Table 1, in the DAVIS dataset evaluation results, ColorAlignNet significantly outperforms the most advanced methods on LPIPS and CDC metrics, including scribble-based video colorization network (SVCNet)[7] and temporally consistent video colorization (TCVC)[25]. In addition, as shown in Table 2, ColorAlignNet works similarly on the Videvo dataset to the DAVIS dataset, also achieving the best results on LPIPS and CDC, and is better than most methods on the SSIM.

Table 1 Results on DAVIS dataset

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CDC ↓
DeepExemplar[24]	20.63	0.846	0.325	4.006×10^{-3}
DeOldify[23]	23.99	0.885	0.306	4.901×10^{-3}
DeepRemaster[6]	21.95	0.848	0.354	5.098×10^{-3}
AutoColor[5]	24.41	0.915	0.264	3.734×10^{-3}
TCVC[25]	25.17	0.921	0.239	3.649×10^{-3}
SVCNet[7]	**25.71**	**0.956**	0.257	4.086×10^{-3}
ColorAlignNet (ours)	24.96	0.930	**0.216**	$\mathbf{3.629 \times 10^{-3}}$

As shown in Fig. 3, our proposed ColorAlignNet is qualitatively compared with different methods listed in Tables 1 and 2, focusing on colorization accuracy and temporal consistency. It can be observed that the DeOldify method exhibits significant color fluctuations between frame 1 and frame 208, indicating weak temporal consistency constraints. Although AutoColor and DeepExemplar emphasize temporal coherence, their colorization results tend to be overly uniform, lacking realistic color restoration. TCVC and SVCNet generate visually pleasing colorized frames but suffer from limited controllability, failing to accurately learn the color information from the ground truth. Additionally, the reference-based DeepRemaster method partially captures the correct colors but falls short in details. For instance, the color of the duck's beak is not correctly reproduced. In contrast, our proposed ColorAlignNet effectively integrates information from adjacent frames through the FlexAlign module. It produces vivid, naturally colorized frames that are visually close to the ground truth, clearly demonstrating its superior performance in video colorization tasks.

Table 2 Results on Videvo dataset

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CDC ↓
DeepExemplar	20.63	0.831	0.348	2.011×10^{-3}
DeOldify	24.31	0.895	0.325	3.134×10^{-3}
DeepRemaster	21.88	0.856	0.358	3.607×10^{-3}
AutoColor	25.90	0.925	0.277	1.668×10^{-3}
TCVC	25.18	0.929	0.273	1.629×10^{-3}
SVCNet	**26.30**	**0.961**	0.289	2.704×10^{-3}
ColorAlignNet (ours)	25.46	0.958	**0.263**	$\mathbf{1.509 \times 10^{-3}}$

Fig. 3 Qualitative comparison with different video colorization methods

2.2 Ablation study

In our method, color propagation is performed by ColorPreb, followed by post-processing operations with FlexAlign and ReFrameNet. In the ablation study, we mainly focus on the impact of these parts on the whole method. Specifically, we conduct ablation experiments from the following four aspects, each corresponding to a specific structure setting in Table 3: removing the temporal prior (structure 1); replacing deformable convolution with a traditional optical flow method (structure 2); substituting the transformer-based structure with a CNN for frame reconstruction (structure 3); replacing ColorPrep with the AdaIN technique for color transfer (structure 4). The results for each setting on the two public datasets, DAVIS and Videvo, are summarized in Table 3.

Table 3 Ablation experiments based on DAVIS and Videvo datasets

Ablation experiment setting	DAVIS				Videvo			
	PSNR ↑	SSIM ↑	LPIPS ↓	CDC ↓	PSNR ↑	SSIM ↑	LPIPS ↓	CDC ↓
Structure 1	23.73	0.921	0.402	4.225×10^{-3}	24.16	0.939	0.325	1.807×10^{-3}
Structure 2	23.95	0.915	0.308	4.128×10^{-3}	25.40	0.943	0.336	1.913×10^{-3}
Structure 3	24.87	**0.934**	0.280	4.357×10^{-3}	25.33	0.946	0.329	1.746×10^{-3}
Structure 4	24.21	0.913	0.230	4.323×10^{-3}	24.87	0.913	0.367	1.514×10^{-3}
ColorAlignNet (ours)	**24.96**	0.930	**0.216**	$\mathbf{3.629 \times 10^{-3}}$	**25.46**	**0.958**	**0.263**	$\mathbf{1.509 \times 10^{-3}}$

Structure 1 We introduce a temporal prior in FlexAlign to enhance correlation between adjacent video frames. It leverages aligned features from the previous stage to guide temporal consistency. To verify its effectiveness, we remove this temporal prior in the ablation experiment. FlexAlign then processes features from only the current stage without previous alignment results. This limits temporal awareness and weakens color consistency across frames.

Structure 2 In the FlexAlign module, we originally adopted deformable convolution[11] to handle feature alignment under large motion scenarios. To verify its effectiveness, we replace it with an optical-flow-based[12] alignment method, keeping other components unchanged. As shown in Table 3, our model declines on all metrics of DAVIS and Videvo. This shows that deformable convolution can better adapt to data characteristics and obtain more satisfactory results when facing complex motion scenes.

Structure 3 In ReFrameNet, we use a transformer-based[10] structure for frame reconstruction, leveraging its powerful global modeling ability to capture long-range pixel dependencies. To evaluate the necessity of the transformer, we replace it with a U-Net-based reconstruction network[7] and assess its reconstruction performance. As shown in Table 3, CNN maintains the same level of PSNR and SSIM metrics compared to the method using a transformer. But because CNN lacks the ability to capture global dependencies and detailed recovery, LPIPS and CDC score significantly lower. This suggests that CNN has more difficulty dealing with complex visual perception and color consistency issues.

Structure 4 We use the source-reference attention-based ColorPrep to extract semantically consistent color information from the reference color video and accurately transfer it to the target grayscale frames. To evaluate its effectiveness, we replace it with the commonly used AdaIN[26] method in image colorization, which only matches the mean and variance of the color distribution. As shown in Table 3, all indicators have decreased. This indicates that AdaIN is insufficient for precise color propagation and temporal consistency, confirming the superiority of source-reference attention in video colorization tasks.

The visualization results from ablation experiments are shown in Fig. 4. It can be seen that our proposed ColorAlignNet achieves the best and most correct colorization effect, demonstrating excellent color recovery. This fully reflects the ability of ColorAlignNet to achieve colorization in different scenarios, emphasizing the controllability of colorization while also fully learning from the features in the reference frame.

Fig. 4 Visualization results from ablation experiments with different settings

3 Conclusions

This study presents an efficient ColorAlignNet model. The model combines reference frame guidance with temporal consistency processing by propagating color information through the source-reference attention mechanism. It also uses a feature aggregation module based on deformable convolution to align the features of adjacent frames. The circulating transformer module further improves the vividness of color recovery, resolving features lost during colorization and feature alignment. ColorAlignNet effectively resolves the conflict between color vividness and temporal consistency in video colorization. The method has been extensively validated through experiments on DAVIS and Videvo datasets, producing compelling results.

Future work mainly involves two directions. The most urgent challenge is real-time performance improvement, especially for large-scale video colorization tasks. As video resolution and length increase, optimizing the model's speed and efficiency becomes essential. Another important direction is handling complex and dynamic scenes. This includes cases with fast motion or extreme lighting conditions. Extending the model's capacity in these areas is necessary to improve robustness and adaptability.

References

[1] WENG S C, SUN J M, LI Y, et al. CT2: colorization transformer via color tokens[C]// European Conference on Computer Vision. Berlin: Springer-Verlag Berlin, 2022: 1-16.

[2] KANG X Y, YANG T, OUYANG W Q, et al. DDColor: towards photo-realistic image colorization via dual decoders[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2024: 328-338.

[3] CONG X Y, WU Y, CHEN Q F, et al. Automatic controllable colorization via imagination[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2024: 2609-2619.

[4] WU Y Z, WANG X T, LI Y, et al. Towards vivid and diverse image colorization with generative color prior [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2022: 14357-14366.

[5] LEI C Y, CHEN Q F. Fully automatic video

colorization with self-regularization and diversity [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 3748-3756.
[6] IIZUKA S, SIMO-SERRA E. DeepRemaster: temporal source-reference attention networks for comprehensive video enhancement [J]. *ACM Transactions on Graphics*, 2019, 38(6): 1-13.
[7] ZHAO Y Z, PO L M, LIU K C, et al. SVCNet: scribble-based video colorization network with temporal aggregation[J]. *IEEE Transactions on Image Processing*, 2023, 32: 4443-4458.
[8] YANG Y X, PAN J S, PENG Z Z, et al. BiSTNet: semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(8): 5612-5624.
[9] CASEY E, PÉREZ V, LI Z R. The animation transformer: visual correspondence via segment matching [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2022: 11303-11312.
[10] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2022: 9992-10002.
[11] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 764-773.
[12] PAN L Y, LIU M M, HARTLEY R. Single image optical flow estimation with an event camera [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 1669-1678.
[13] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video object segmentation [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 724-732.
[14] LAI W S, HUANG J B, WANG O, et al. Learning blind video temporal consistency[C]// Computer Vision-ECCV 2018. Cham: Springer, 2018: 179-195.
[15] PARSANIA P P S, VIRPARIA D P V. A review: image interpolation techniques for image scaling[J]. *International Journal of Innovative Research in Computer and Communication Engineering*, 2015, 2(12): 7409-7414.
[16] DE SANTIS M, LUCIDI S, RINALDI F. Fast active-set-type algorithms for *L*1-regularized linear regression [J]. *SIAM Journal of Optimization*, 2016, 26(1):781-809.
[17] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution [C]//Computer Vision-ECCV 2016. Cham: Springer, 2016: 694-711.
[18] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [J]. *Advances in neural information processing systems*, 2014, 27: 261560300.
[19] SZE V, BUDAGAVI M, SULLIVAN G J. High efficiency video coding (HEVC): algorithms and architectures[M]. Cham: Springer International Publishing, 2014.
[20] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600-612.
[21] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 586-595.
[22] OREL R, LUO X, SHAN M Y, et al. StyleSDF: high-resolution 3D-consistent image and geometry generation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 13493-13503.
[23] SALMONA A, BOUZA L, DELON J. DeOldify: a review and implementation of an automatic colorization method [J]. *Image Processing on Line*, 2022, 12: 347-368.
[24] ZHANG B, HE M M, LIAO J, et al. Deep exemplar-based video colorization [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 8044-8053.
[25] LIU Y H, ZHAO H Y, CHAN K C K, et al. Temporally consistent video colorization with deep feature propagation and self-regularization learning [J]. *Computational Visual Media*, 2024, 10(2): 375-395.
[26] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization [C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 1510-1519.

ColorAlignNet：基于参考帧的时间聚合视频着色网络

朱文志[1,2]，王　彤[1,2*]

1. 东华大学 信息科学与技术学院，上海 201620

2. 东华大学 数字化纺织服装技术教育部工程研究中心，上海 201620

摘　要：视频着色是一项为老旧视频注入新生命的技术。尽管现有的着色方法在静态图像和低动态视频上表现出色，但它们通常难以处理复杂的动态场景。为此，本研究提出了一种基于参考帧的时间聚合视频着色网络 ColorAlignNet。该网络使用源−参考注意力机制，将参考帧中的颜色信息有效传播至灰度帧，确保色彩还原的准确性。同时，通过设计基于可变形卷积的特征对齐模块，对相邻帧进行特征对齐，以提升时序一致性。最后，结合循环 transformer 模块来重构最终的预测结果。大量实验结果表明，ColorAlignNet 在 DAVIS 和 Videvo 数据集上取得了优异性能，在感知图像块相似度（learned perceptual image patch similarity，LPIPS）和色彩分布一致性（color distribution consistency，CDC）指标上均优于现有主流方法。

关键词：可变形卷积；视频着色；Swin-transformer