OmniFusion: Exemplar-based Video Colorization Using OmniMotion and DifFusion Priors

Xiaoyuan Fang[1], Longquan Dai[1], and Jinhui Tang[1]

Nanjing University of Science and Technology, Nanjing 210094, China

Abstract. Exemplar-based video colorization is a challenging task that involves the consistent propagation of colors across frames and the reasonable inference of colors from grayscale within frames. This paper proposes a novel video colorization method called OmniFusion, which iteratively completes the video colorization through two following steps. In the inter-frame propagation step, OmniMotion establishes correspondences between pixels across video frames. Any grayscale pixel can be queried whether a corresponding pixel and color are available from the exemplar according to such correspondences. Consequently, the processed images may still contain regions lacking color. In the intra-frame inpainting step, diffusion model provides these grayscale regions in a frame with plausible colors. The colorized frame is then fed into the first step as an exemplar, accepting queries from all uncolored pixels. This iterative process continues until all pixels are colorized. Evaluations indicate that OmniFusion achieves excellent performance in video colorization, surpassing existing methods in terms of color fidelity and visual quality.

Keywords: Video Colorization · Temporal Consistency · Diffusion Model

1 Introduction

Due to the immaturity of the photography technology in the last century, a large number of excellent films could only be captured in black and white, gradually fading out of modern society's consciousness. Video colorization revitalizes these grayscale works by imbuing them with realistic colors, enhancing their visual appeal and facilitating their dissemination in contemporary society.

Traditional methods of video colorization[36] depend on the collaborative efforts of an interdisciplinary team comprised of skilled colorists, rotoscoping animators, artists and historians. Achieving a satisfying and coherent result necessitates a substantial investment of hours and efforts. Consequently, automation of the video colorization is highly desired.

The simplest approach for automation of the video colorization [23,62,26] is to employ an image-based colorization model to process each grayscale frame independently with no temporal modeling. In recent years, a large number of single-frame colorization methods [66,13,62,14,15,19,49] propose have achieved significant progress. Su *et al.* [41] proposed to leverage an off-the-shelf object detector to obtain cropped object images and combine object features with complete features to realize the colorization of the entire image. However, these works are often short of temporal consistence constraints, resulting in colorized videos accompanied by severe color flickering, even artifacts and discontinuities. Therefore, generating high-quality color videos imposes higher demands on the colorization task: it not only requires reasonable colorization for each frame but also necessitates maintaining color consistency between frames during the colorization process.

To enable the model to perceive and focus on temporal correlations, various methods [35,5] stack features of each video frame along the time dimension, feeding them into the network to maintain temporal consistence. This approach significantly improves colorization performance. However, such temporal features are sparse and contain substantial redundancy, making it challenging to efficiently utilize the spatiotemporal priors between frames through simple stacking. More importantly, stacking redundant features imposes a significant burden on GPU memory when processing longer video tasks.

Exemplar-based video colorization[57,52,68] is a crucial sub-task in the field of automated video colorization. The objective is to accurately propagate chromatic information from a reference image to other grayscale video frames. These systems[8,58,62] often consist of two sub-nets: the similarity sub-net obtains the coarse chromaticity map via matching the basic feature statistics of the input pairs; the colorization sub-net refines the generated chromaticity map to produce the final colorful result. However, these methods lack the understanding of the image content, resulting in coloring parts that do not appear in the instance into unreasonable colors.

Generative models[9,37] continue to evolve and demonstrate remarkable performance in image restoration, but they have yet to provide a robust solution for the challenge of temporary consistency in video colorization. Subsequent researches utilize pre-trained GANs[71,11,44,18] or diffusion models[24,4,45,50] with stronger representation capabilities to enhance model representation ability and improved coloring effect. However, such models, especially those based on Stable Diffusion[37,25,48], often experience color flickering of the same object across frames due to the lack of temporal consistency modeling.

In this paper, we explore leveraging the superior pixel-level tracking capabilities of OmniMotion for inter-frame color propagation, and employing the powerful generative abilities of Stable Diffusion to achieve comprehensive image colorization. By organically integrating these two works, we aim to address the challenge of exemplar-based video colorization effectively, minimizing the color inconsistency and flickering and producing a harmonious and visually appealing result.

The main contributions of this paper can be summarized as follows:

- We propose an exemplar-based video colorization method, which iteratively achieve the coloriztion of long videos without training a general network.
- We apply OmniMotion to establish correspondences between pixels across frames, enabling accurate pixel-level color propagation between frames.
- We propose a novel conditional image inpainting method with the modified Stable Diffusion, capable of providing rational colors for where color information from the reference frame is unavailable.

Experiments show that our proposed method achieves excellent results in both temporal consistency and accuracy.

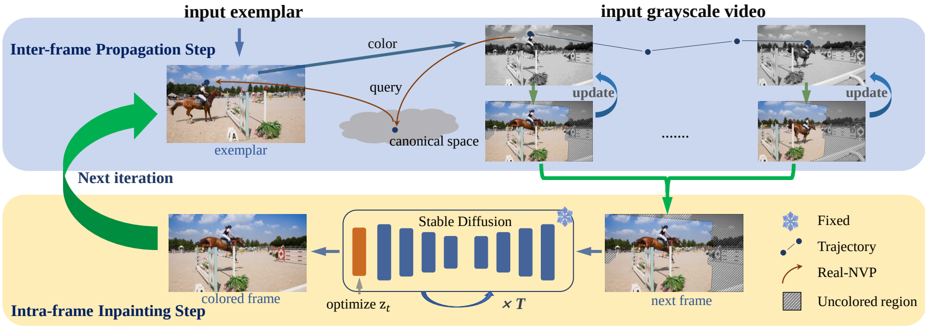

Fig. 1: Overview of our iterative colorization method. In the inter-frame propagation step (*blue*), we employ OmniMotion to establish correspondences between pixels across video frames. Any gray pixel in the video will be queried whether color is available in the exemplar and then be updated. Therefore, the image colored by propagation may still contain uncolored regions. In the intra-frame inpainting step (*yellow*), we utilize the diffusion prior to achieve complete colorization of the next frame which will be fed into inter-frame propagation step as an exemplar in the next iteration. The above two steps are iteratively repeated until the entire video is fully colorized.

2 Related Work

2.1 Video Colorization

Video colorization[59,43,71,25,26,23] requires to take color performance and temporal consistency in account. In order to suppress the color flicker of each frame, a general temporal filter[3,20] is used to introduce correlation in the temporal dimension, leading to color distortion. FAVC[21] proposed by Lei *et al.* is comprised of a colorization network for video frame colorization and a refinement network for spatiotemporal color refinement. TCVC [27] jointly considered colorization and temporal consistency in an unified framework optimized by a self-regularization learning scheme.

Exemplar-based video colorization[58,57,62,8,48] is an important sub-task in the field of automatic video colorization, which aims to get command of the color style of the other grayscale frames with a reference frame. Iizuka *et al.* [12] proposed a fully 3D convolution method that introduced a source-reference attention layer to colorize long videos while maintaining temporal consistency. Xu *et al.*[52] consists of two subnetworks: the similarity sub-net obtains the coarse chromaticity map via matching the basic feature statistics of the input pairs; the colorization sub-net refines the generated chromaticity map to produce the final colorful result. Liu *et al.*[25] explored an approach based on pre-trained diffusion model where adds designed Color Propagation Attention, but the consistency constraints across video frames were insufficient.

2.2 Neural Video Representation

Neural Radiation Field (NeRF)[29] synthesizes novel views of 3D complex scenes by a neural network to remember the 3D space to implicitly represent the 3D model and calculating the pixel color along a single perspective according to volume rendering. Some methods apply such latent representation idea to video representation. CoDeF[32] compresses the video into a canonical image with optical flow information constraints. Some dynamic reconstruction methods [51,53,54] are also capable of generating 2D motion, but they typically center around objects, focusing primarily on articulated objects. Additionally, there are representations based on video decomposition, such as Layered Neural Atlases [16] and Deformable Sprites [60], which primarily concentrate on the mapping between each frame and a global texture atlas. Some other methods[22,34] including OmniMotion [46] make use of the reversible flow network Real-NVP[7] to map video coordinates to the canonical space with excellent properties for action estimation and tracking. To extend image editing methods to videos, NVEdit[56] employs video implicit neural representation embedded within diffusion model to enhance temporal consistency. This work employs such a scheme to propagate colors between frames.

2.3 Diffusion Model

Diffusion models [9,40,6,31,17] is probabilistically generative models meant to fit the data distribution p(x) by means of denoising the normally distributed variables progressively. Latent diffusion models [48,42,38,61] or Stable Diffusion [37] use the perceptual compression of the autoencoders $\mathcal{E}$ and $\mathcal{D}$ for efficient low-dimensional representation features. To avoid the tremendous computational cost of retraining Stable Diffusion[4,2], there are currently two main approaches for controlling generation of diffusion models. One [10,63,33,69] is to design appropriate adapters as a conditional entrance to the original denoiser and fine-tune these adapters with most network parameters fixed. The other training-free way [39,67,70,55,30] is to optimize the latent variables with the designed losses related to the condition during the SD inference process. Most of the existing colorization methods based on diffusion model[24,4,45,50] use the former to introduce grayscale constraints. In these works [25,48], gray and color reference features are fused through the color propagation attention proposed and fed into the designed adapter similar to ControlNet[63].

3 Method

Our iterative video colorization strategy comprises two primary steps, as illustrated in fig. 1. In the inter-frame propagation step, we employ OmniMotion to establish correspondences between pixels across video frames, ensuring temporal consistency in the generated video. Any grayscale pixel in the video will be queried whether color information is available in the exemplar. Since this step can only colorize pixels that appear in both the reference frame and gray frame, the intra-frame inpainting step applys appropriate colors to the uncolored parts of a grayscale frame with diffusion priors. The colorized frame from the second step becomes the exemplar for the first step in the next iteration, and this process repeats until all frames are fully colorized.

To give a more convenient explanation, we describe the simplified exemplar-based video colorization task as follows: Given an exemplar x_1^{lab} and a set of consistent video

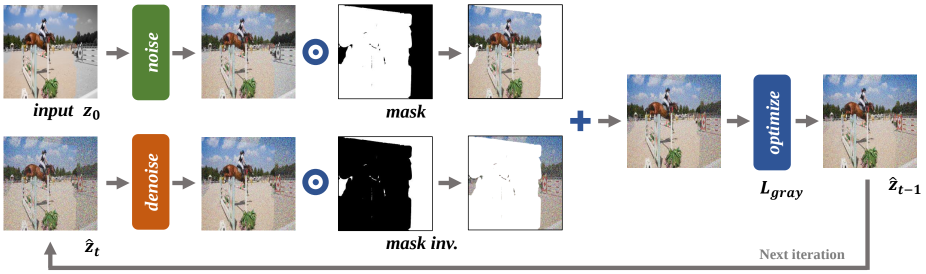

Fig. 2: Overview of the intra-frame inpainting step. To incorporate the colored region as a conditioning factor, we modify the standard denoising process of the diffusion model. At each denoising step, we sample the colored region from the noised input (*top*) and the uncolored region from the denoised output (*bottom*). Then we optimize the latent based on the grayscale loss.

frames $X = \{x_1^{lab}, x_2^l, \ldots, x_L^l\} \in R^{L\times H\times W\times 3}$ where grayscale frames are expanded to three channels. The goal is to generate a color video $\hat{X} = \{\hat{x}_1^{lab}, \hat{x}_2^{lab}, \ldots, \hat{x}_L^{lab}\} \in R^{L\times H\times W\times 3}$ where l and ab represent the luminance and chrominance in the CIELAB color space, and L, H and W represent the frame length, height and width. Moreover, $M \in R^{L\times H\times W}$ is the binary matrix indicating whether each pixel has been colorized or not. Below, we provide a detailed description of both steps in section 3.1 and section 3.2, respectively. Then we elucidate our complete iterative colorization algorithm in section 3.3.

3.1 Inter-frame Propagation Step

We employ OmniMotion[46], which currently offers the best tracking performance, to better achieve color information transfer between frames. This work introduces depth into each video frame, treating each frame as a local volume. The OmniMotion representation comprises a canonical 3D space and a set of bijections that map the 3D coordinate of each frame's local volume to the 3D canonical space. This approach also involves a coordinate-based density network over the canonical space, which maps each canonical 3D coordinate to a density, thereby indicating the locations of surfaces within the frame volume. Once these bijections using Real-NVP [7] are optimized by noisy optical flow estimates, the representation can be queried with any pixel from any frame to produce a smooth and accurate motion trajectory throughout the entire video, as illustrated in the blue section of the fig. 1.

Before colorizing a grayscale video, we first need to train this neural latent video representation. We then query the 3D coordinate from every grayscale pixel in the exemplar and determine whether the target coordinate is on the surface of the exemplar volume (visible) based on these 3D coordinates (including depth). If it is visible, we assign the color channels to the grayscale pixel, achieving pixel-level colorization. We represent inter-frame propagation step with the following function:

$$X, M = InterFramePropagationStep(x_i^{lab}, X, M, i) \tag{1}$$

Algorithm 1: Pseudocode of the Overall Algorithm

```
  Input: A reference color image $x_1^{lab} \in R^{H \times W \times 3}$; The video frames
         $X = \{x_1^{lab}, x_2^l, \ldots, x_L^l\} \in R^{L \times H \times W \times 3}$
  Output: The colored video $\hat{X} = \{x_1^{lab}, x_2^{lab}, \ldots, x_L^{lab}\} \in R^{L \times H \times W \times 3}$.
  Train OmniMotion
  Initial the mask $M = \{M_1, ..., M_L\} \in R^{L \times H \times W}$
  for i=1 to L do
      $X, M = InterFramePropagationStep(x_i^{lab}, X, M, i)$
      if every frame is full colored then
          break
      $x_{i+1}^{lab}, M_{i+1} = IntraFrameInpaintingStep(x_{i+1}^l, M_{i+1})$
  The colored video $\hat{X} = X$
```

where x_i^{lab} denotes the exemplar in the current iteration, i indicates its index, X and M represent the video and the binary matrix to be updated. It is important to note that the video contains partially colored frames. For instance, in the first iteration, all pixels in exemplar have color while the remaining frames do not. After processing, all frames will be imbued with some color.

3.2 Intra-frame Inpainting Step

After the inter-frame propagation step, we obtain partially colorized frames. The subsequent challenge lies in seamlessly blending the remaining grayscale portions of a frame with the appropriate colors. To achieve this, we propose utilizing Stable Diffusion[37], known for its robust generative capabilities, as a prior to guide the color inpainting process, as illustrated in fig. 2. Our modifications to SD are primarily divided into the following two main points: 1) Integrate color conditions to keep the colors for boundary regions harmoniously with the assigned areas; 2) Introduce grayscale latent optimization to maintain the grayscale values of the generated colors consistent with the original ones. Below, we provide a detailed explanation of our approach.

Color conditions. After inter-frame propagation step, we obtain the frame where some regions contain color information. We need to infer the color for the uncolored parts based on the already colorized regions, similar to the process of inpainting. Inspired by RePaint [28], we modify the standard denoising process of the Stable Diffusion [37], as illustrated in fig. 2. Since every reverse step depends solely on z_t, we sample the colored region from the noised input and the uncolored region from the denoised output, where we keep the correct properties of the corresponding distribution.

This approach is formalized in the following equation:

$$\hat{z}_{t-1} = m \odot Noise(z_0, t) + (1 - m) \odot Denoise(\hat{z}_t, t) \tag{2}$$

where m represents the binary mask M_i interpolated to the size in the latent space, and z_0 represents the latent variable of the selected semi-colored image.

Grayscale latent optimization. In the inference stage of continuous denoising, the color of the previously uncolored areas will be generated based on prior information

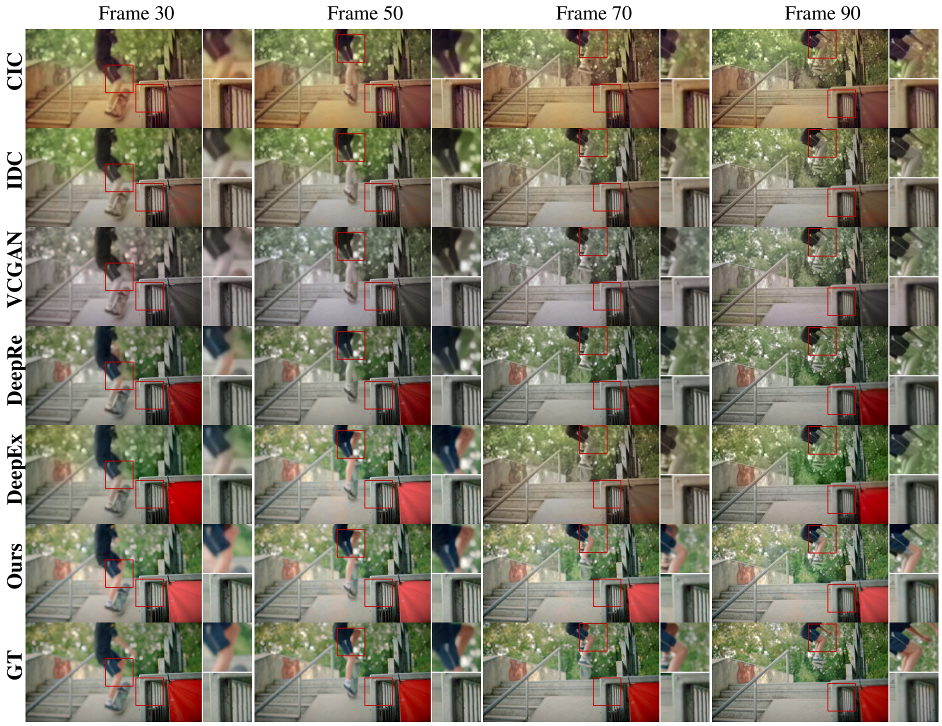

Fig. 3: Comparison with existing colorization methods on the video in the Videvo dataset [1]. They are corresponding colorization results generated by CIC[64], IDC[66], VCGAN[71], Deep Remaster[12],Deep Exemplar[62] and the proposed method respectively.

embedded within the model. To ensure that the generated color is constrained by the original grayscale information, we draw inspiration from previous works[39,67,55] and optimize latent variables during the inference stage using the following grayscale loss function:

$$\mathcal{L}_{gray}(\hat{z}_t) = ||(F(\hat{z}_t) - x_i) \odot (1 - M_i)||_1 \tag{3}$$

where, x_i denotes the frame to color. $\odot$ signifies counter multiplication. The binary mask M_i indicates whether each pixel in the frame has been colorized or not. The variable $\hat{z}_t$ represents the latent variable after the inpainting process. The process denoted by $F(\cdot)$ encompasses the transformation of z_t to z_0, followed by the decoding of z_0 into pixel space, and ultimately degrading the resulting image to a grayscale representation. In each iteration, $\hat{z}_t$ is updated by taking one gradient descent step to minimize $\mathcal{L}_{gray}$:

$$\hat{z}_t = \hat{z}_t - \eta \cdot \frac{\partial \mathcal{L}_{gray}(\hat{z}_t)}{\partial \hat{z}_t} \tag{4}$$

where η is the learning rate for latent optimization.

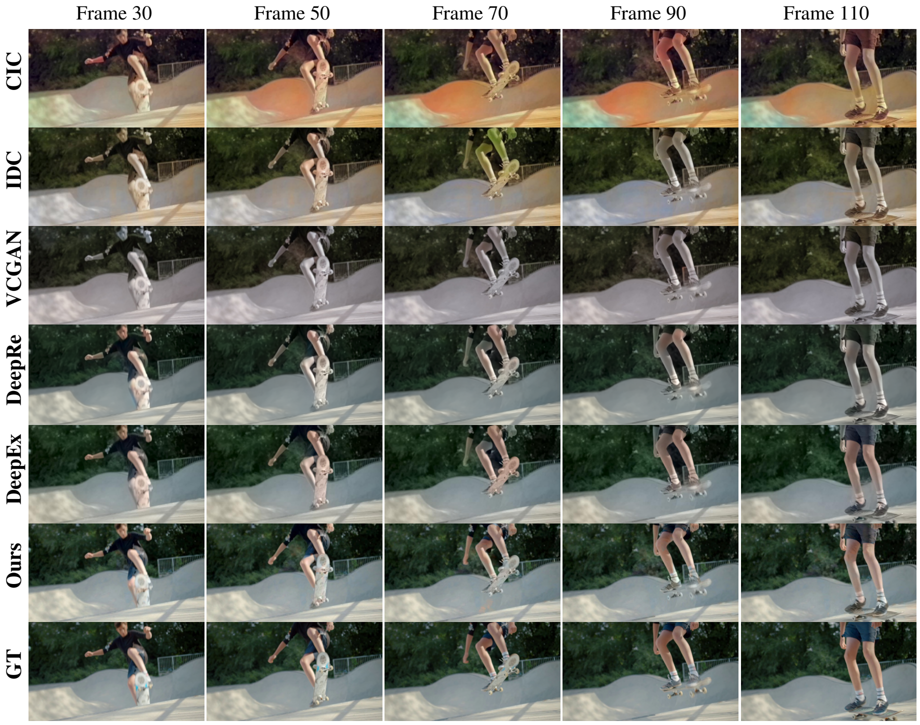

Fig. 4: Comparison with existing colorization methods on skate from videvo dataset[1]. They are corresponding colorization results generated by CIC[64], IDC[66], VCGAN[71], Deep Remaster[12],Deep Exemplar[62] and the proposed method.

This loss function quantifies the alignment between the generated colors in the grayscale regions and the grayscale levels of the corresponding pixel positions in the original image. By minimizing this loss through iterative optimization and denoising steps, we ensure that the generated image maintains a high degree of fidelity to the original grayscale image in the uncolored regions.

Similarly, we represent the intra-frame inpainting step with the following function:

$$x_i^{lab}, M_i = IntraFrameInpaintingStep(x_i, M_i) \tag{5}$$

where x_i denotes the frame to colorize and the binary mask M_i to be updated indicates whether each pixel in the frame has been colorized or not. This step outputs x_i^{lab} with complete color.

3.3 Iterative Color Refinement

The primary cause of flickering when using diffusion models for video colorization is the repetitive, random coloring of the same objects. Our iterative algorithm addresses this by anchoring the pixels that already acquire color, ensuring that each pixel can

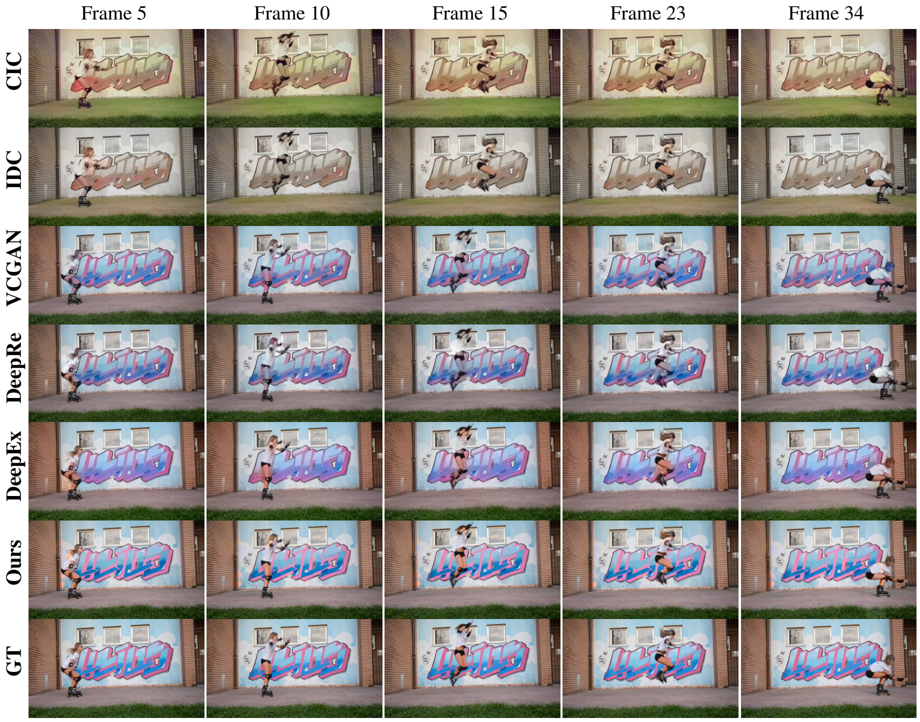

Fig. 5: Comparison with existing colorization methods on rollerblade from DAVIS dataset[35]. They are corresponding colorization results generated by CIC[64], IDC[66], VCGAN[71], Deep Remaster[12],Deep Exemplar[62] and our method.

be colorized only once. This prevents the color flickering that arises from repetitive colorization.

As illustrated in algorithm 1, we train OmniMotion and initialize the input grayscale video and the mask before iterating. Upon entering the iteration process, we leverage the tracking properties of OmniMotion[46] to propagate the reference frame's color to all other frames in the inter-frame propagation step. In the intra-frame inpainting step, we utilize Stable Diffusion[37] to colorize the next frame. The result is then used as color example for the next iteration, accepting queries from all pixels without color. Each colorization step updates the corresponding regions of the mask. This process effectively minimizes color errors and flickering, particularly in the task of colorizing long videos.

4 Experiment

In this section, we compare our methods with various existing video colorization approaches in section 4.1 and display necessary ablation results in section 4.2. Finally, in section 4.3 we present the colorization results on videos of different styles. Now, we

Table 1: Quantitative comparisons of the proposed method against video colorization methods on the DAVIS[35] and Videvo dataset[1]. Our method achieves the excellent performance in terms of PSNR, SSIM[47], LPIPS[65] and CDC[27].

Method	**DAVIS**				**Videvo**			
	PSNR ↑	SSIM ↑	LPIPS ↓	CDC ↓	PSNR ↑	SSIM ↑	LPIPS ↓	CDC ↓
CIC[64]	23.19	0.901	0.176	0.006180	22.51	0.899	0.185	0.003595
IDC[66]	24.88	0.949	0.116	0.005017	25.35	0.952	0.095	0.002578
InsColor[41]	24.51	0.951	0.082	0.009574	24.80	0.953	0.079	0.008019
VCGAN[71]	23.77	0.920	0.142	0.006114	25.11	0.926	0.085	0.002998
ColorDiffuser[25]	23.73	0.939	0.213	0.004107	25.27	0.951	0.205	0.003591
FAVC[21]	24.38	0.906	0.191	0.004221	24.81	0.905	0.194	0.001880
TCVC[27]	25.50	0.955	0.175	**0.003947**	25.43	0.956	0.068	**0.001649**
DeepRemaster[12]	27.03	0.964	**0.057**	0.005098	32.25	0.964	**0.054**	0.003607
Ours-SDv1.5	27.32	0.968	0.077	0.004143	31.11	0.952	0.065	0.002176
Ours-SDv2.0	**27.87**	**0.970**	0.070	0.004026	**32.89**	**0.967**	0.058	0.002020

start by describing the datasets used in our experiments, performance evaluation metrics and implementation details.

Datasets. To evaluate the actual performance of our proposed colorization method, we compared it with existing methods on the same test set. We selected the DAVIS video dataset[35] and the Videvo dataset [1] as our test sets. The DAVIS testset consists of 20 video clips of various scenes, each clip containing approximately 30 to 100 frames. The Videvo testset [1] contains 20 videos, each with about 200 frames. In total, we evaluated our models and baselines on 40 test videos.

Evaluation metrics. To evaluate the quality of the colorized videos comprehensively, we use the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM)[47], and the learned perceptual image patch similarity (LPIPS) [65] matrices. To assess the temporal consistency of colored videos, we use the color distribution consistency index (CDC)[27] .

Implementation details. We train all models using the PyTorch framework on a machine with the GeForce RTX 3090 GPUs. We adopt the CIE LAB color space for each frame in our experiments. In the inter-frame propagation step of our method, we remove the color MLP of Omnimotion for video representation. In the intra-frame inpainting step, the diffusion model uses SDv1.5 or SDv2.0. While our method does not require training or fine-tuning, during the optimization of intra-frame inpainting step, the learning rate η is set as 0.02, the number of denoising steps is set as 50 and each step is optimized once.

4.1 Comparisons with Existing Methods

We compare our proposed method with existing state-of-the-art approaches including both the automatic colorization ones[64,66,71,21,27] and several exemplar-based video colorization ones[12,62,25]. Considering that exemplar-based methods require a reference exemplar as guidance, we uniformly choose the first frame with the ground truth color as the color reference for fair comparisons.

Quantitative comparison. table 1 exhibits the quantitative evaluation results of various methods on the DAVIS[35] and Videvo dataset [1], among which the video colorization

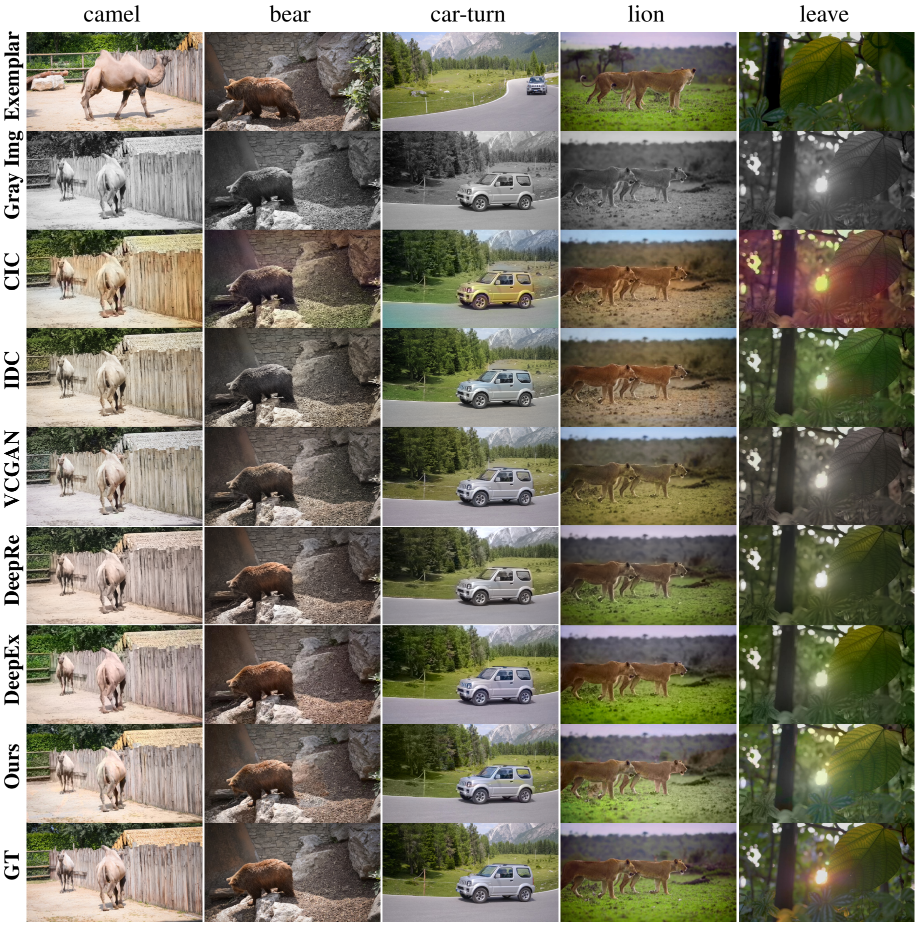

Fig. 6: Comparison with existing colorization methods. The first two rows are input target-exemplar image pairs. The next rows are corresponding colorization results generated by CIC[64], IDC[66], VCGAN[71], Deep Remaster[12],Deep Exemplar[62] and the proposed method.

approach proposed in this paper generates fair color results. In particular, the performance in PSNR is better than the exemplar-based method DeepRemaster[12] by 0.84 dB. Compared with the generative model including VCGAN[71] and ColorDiffuser[25] used for video colorization, our method also shows strong superiority in various indexes. table 1 fully demonstrates that our algorithm proposed can generate color videos with good temporal consistency and accuracy. Since Ours-SDv2 has better comprehensive index performance, Ours-SDv2 is used to represent our proposed method in subsequent comparative experiments.

Qualitative comparison. Our method is primarily compared with several notable approaches: the single image colorization methods CIC[64], IDC[66], the automatic video

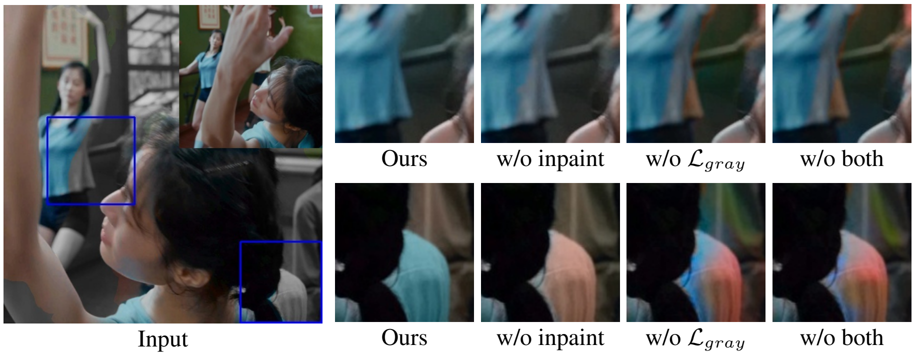

Fig. 7: Effectiveness of the proposed method. The left image illustrates the partially colored input for the intra-frame inpainting step, with the exemplar positioned in the top-left corner. The two rows of images on the right show the different colorization details of the highlighted region in the left image. From left to right, they represent our method, the method without color condition introduction, the method without grayscale constraint, and the method without both.

colorization method VCGAN[71], and the exemplar-based colorization methods Deep Remaster[12] and Deep Exemplar[62]. The exemplar-based methods take in account color image hints for colorization.

In order to more intuitively show the consistency and coloring effect of videos generated by different methods, we selected relatively typical cases from both datasets as illustrate in the figs. 3 to 5. Firstly, image colorization methods[64,66] exhibit significant color inconsistency issues, primarily due to their failure to consider temporal correlation. On the Videvo dataset [1], the results of VCGAN[71] appear overly gray and lack vibrancy and in the fig. 5, VCGAN[71] causes background colors to bleed into the foreground, resulting in poorly harmonized colorization. While Deep Remaster[12] and Deep Exemplar[62] generally produce videos with rich and harmonious colors, they still fail to achieve consistent colorization throughout the video. This is particularly evident in emphasized regions, such as the boy's legs in the skateboarding example, where color inconsistency is noticeable. In contrast, our method demonstrates superior performance in these examples. As shown in the magnified region, the color of the boy's legs is largely consistent, and there is no color cast in the red region in the lower right corner.

To provide a more comprehensive comparison with other video colorization methods regarding their visual effects on standard video sequences, we also conducted extensive tests on the DAVIS[35] dataset and Videvo dataset[1]. We selected several representative and challenging samples, which are illustrated in the fig. 6. For instance, the challenge in the camel and car turning examples lies in whether the newly appeared scene on the right side can be colorized plausibly. Our method produced satisfactory results. The experimental result demonstrates that both VCGAN[71] and Deep Remaster[12] exhibit some color mismatches. For example, in the rollerblading girl sample, the girl's hair is incorrectly tinted blue. In contrast, both our method and Deep Exemplar [62] generate relatively stable and consistent colorization results.

Table 2: Quantitative evaluations of each pipeline on the DAVIS dataset[35]

denoising steps	optimization times	inpaint	$\mathcal{L}_{gray}$	PSNR↑	SSIM↑	LPIPS↓
50	1	×	×	24.78	0.892	0.155
50	1	✓	×	26.24	0.904	0.127
50	1	×	✓	25.37	0.922	0.111
50	3	✓	✓	27.17	0.941	0.103
100	1	✓	✓	27.99	0.965	0.075
50	1	✓	✓	27.87	0.970	0.070

4.2 Ablation Study

In this section we conduct necessary ablation experiments to illustrate the effectiveness of our method. We remove certain adjustments from the intra-frame inpainting step and test on the DAVIS dataset[35]. table 2 presents various quantitative evaluation metrics for different pipelines. Additionally, we test these models on clips from the movie Youth. fig. 7 shows the colorization details of these pipelines.

Grayscale latent optimization. We introduced a grayscale latent optimization during the SD inference stage to maintain the grayscale value consistency. The experimental results in the fig. 7 demonstrate that images lacking grayscale information exhibit significant color deviations. For instance, the image without grayscale constraints shows the girl's blue short with incorrect colors, with noticeable red appearing at the edges.

Color conditions. In our method, the inter-frame propagation step generates partially colored video frames based on the reference color image. To enable the diffusion model in the intra-frame inpainting step to recognize the colors of the already colored regions from the inter-frame propagation step, we need to incorporate these colored parts of the video frames as conditions into the Stable Diffusion model through eq. (2). Experimental results shown in the fig. 7 demonstrate that the absence of color condition incorporation leads to disharmonious edge areas in the generated images. For instance, there is a noticeable boundary between gray and blue on the girl's blue top, and the lack of blue color hints on the girl's clothes results in the clothes being rendered red. With the incorporation of color hints, the overall effect becomes significantly more natural, with smoother color transitions.

Optimization. We also conducted experiments on the number of denoising steps and the number of optimizations per step. The experimental results in the table 2 demonstrate that increasing the number of optimizations per denoising step will cause various evaluation indicators to deteriorate and increasing the number of denoising steps can slightly improve the colorization effect.

4.3 Evaluations on Real-World Grayscale Videos

Since our approach does not require training a generalized model on any trainset, we tested our proposed method on other grayscale videos. We applied it to various videos, including some from cartoons and others from real-world scenarios or movies. The results, as illustrated in the fig. 8, demonstrate the effectiveness of our method across different types of videos.

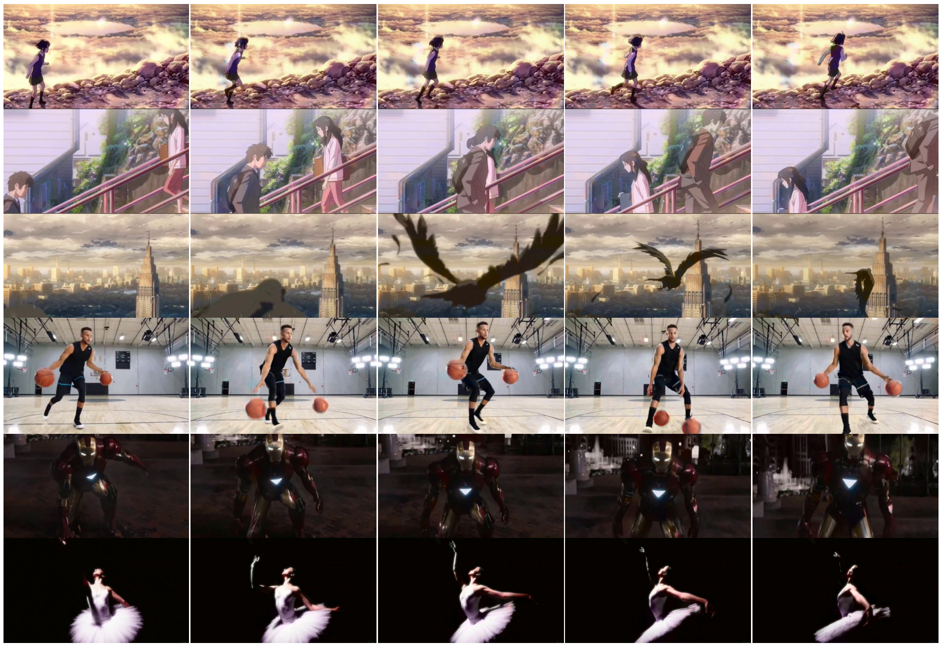

Fig. 8: Effectiveness of the proposed method.The first column represents the input exemplars, while the other columns display the outputs generated by our method.

5 Conclusion

This paper explores a novel video colorization approach utilizing Stable Diffusion with strong image priors. Our method achieves consistent colorization of complete videos through iterative processing through two step. Specifically, in the inter-frame propagation step, we utilize OmniMotion to ensure the propagation of color information from the reference frame to the grayscale frames. In the intra-frame inpainting step, we leverage the existing large generative model, Stable Diffusion, to optimize latent variables and introduce grayscale constraints, thereby inpainting in the color information within the image and generating a new reference frame with complete color. Experimental results demonstrate the effectiveness of our approach, showing excellent colorization performance.

However, there is still significant room for improvement in our method. Due to the iterative generation in two steps, the expenditure of prediction time is relatively high. Additionally, the generated results sometimes still exhibit inconsistencies due to the lack of error correction and refinement mechanisms. In the future, we plan to introduce more reasonable constraints to allow the two steps to mutually correct and adapt, thereby producing more coherent and harmonious colorized images.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 62372237. We would like to express our sincere gratitude to the Foundation for their financial support.

References

1. Videvo. `https://www.videvo.net`, accessed: Sep. 30, 2019 [Online]
2. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Conference on Computer Vision and Pattern Recognition (2023)
3. Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S., Pfister, H.: Blind video temporal consistency. Transactions on Graphics (2015)
4. Cao, Y., Meng, X., Mok, P., Lee, T.Y., Liu, X., Li, P.: AnimeDiffusion: Anime diffusion colorization. Transactions on Visualization and Computer Graphics (2024)
5. Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder–decoder approaches. Association for Computational Linguistics (2014)
6. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Conference and Workshop on Neural Information Processing Systems (2021)
7. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using Real NVP. In: International Conference on Learning Representations (2016)
8. He, M., Chen, D., Liao, J., Sander, P.V., Yuan, L.: Deep exemplar-based colorization. Transactions on Graphics (2018)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Conference and Workshop on Neural Information Processing Systems (2020)
10. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2021)
11. Huang, Z., Zhao, N., Liao, J.: UniColor: A unified framework for multi-modal colorization with transformer. Transactions on Graphics (2022)
12. Iizuka, S., Simo-Serra, E.: DeepRemaster: Temporal source-reference attention networks for comprehensive video enhancement. Transactions on Graphics (2019)
13. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. Transactions on Graphics (2016)
14. Ji, X., Jiang, B., Luo, D., Tao, G., Chu, W., Xie, Z., Wang, C., Tai, Y.: ColorFormer: Image colorization via color memory assisted hybrid-attention transformer. In: European Conference on Computer Vision (2022)
15. Kang, X., Yang, T., Ouyang, W., Ren, P., Li, L., Xie, X.: DDColor: Towards photo-realistic image colorization via dual decoders. In: International Conference on Computer Vision (2023)
16. Kasten, Y., Ofri, D., Wang, O., Dekel, T.: Layered neural atlases for consistent video editing. Transactions on Graphics (2021)
17. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Conference and Workshop on Neural Information Processing Systems (2022)
18. Kim, G., Kang, K., Kim, S., Lee, H., Kim, S., Kim, J., Baek, S.H., Cho, S.: BigColor: Colorization using a generative color prior for natural images. In: European Conference on Computer Vision (2022)
19. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer. In: International Conference on Learning Representations (2020)
20. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: European Conference on Computer Vision (2018)
21. Lei, C., Chen, Q.: Fully automatic video colorization with self-regularization and diversity. In: Conference on Computer Vision and Pattern Recognition (2019)

22. Lei, J., Daniilidis, K.: CaDeX: Learning canonical deformation coordinate space for dynamic surface representation via neural homeomorphism. In: Conference on Computer Vision and Pattern Recognition (2022)
23. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. In: Special Interest Group for Computer Graphics (2004)
24. Liang, Z., Li, Z., Zhou, S., Li, C., Loy, C.C.: Control Color: Multimodal diffusion-based interactive image colorization. arXiv preprint arXiv:2402.10855 (2024)
25. Liu, H., Xie, M., Xing, J., Li, C., Wong, T.T.: Video colorization with pre-trained text-to-image diffusion models. arXiv preprint arXiv:2306.01732 (2023)
26. Liu, S., Zhang, X.: Automatic grayscale image colorization using histogram regression. Pattern Recognition Letters (2012)
27. Liu, Y., Zhao, H., Chan, K.C., Wang, X., Loy, C.C., Qiao, Y., Dong, C.: Temporally consistent video colorization with deep feature propagation and self-regularization learning. Computational Visual Media (2024)
28. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using denoising diffusion probabilistic models. In: Conference on Computer Vision and Pattern Recognition (2022)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (2020)
30. Mo, S., Mu, F., Lin, K.H., Liu, Y., Guan, B., Li, Y., Zhou, B.: FreeControl: Training-free spatial control of any text-to-image diffusion model with any condition. In: Conference on Computer Vision and Pattern Recognition (2024)
31. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning (2021)
32. Ouyang, H., Wang, Q., Xiao, Y., Bai, Q., Zhang, J., Zheng, K., Zhou, X., Chen, Q., Shen, Y.: CoDeF: Content deformation fields for temporally consistent video processing. In: Conference on Computer Vision and Pattern Recognition (2024)
33. Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024)
34. Paschalidou, D., Katharopoulos, A., Geiger, A., Fidler, S.: Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In: Conference on Computer Vision and Pattern Recognition (2021)
35. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Conference on Computer Vision and Pattern Recognition (2016)
36. Pierre, F., Aujol, J.F.: Recent approaches for image colorization. In: Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision (2023)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Conference on Computer Vision and Pattern Recognition (2022)
38. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Conference on Computer Vision and Pattern Recognition (2023)
39. Shi, Y., Xue, C., Liew, J.H., Pan, J., Yan, H., Zhang, W., Tan, V.Y., Bai, S.: DragDiffusion: Harnessing diffusion models for interactive point-based image editing. In: Conference on Computer Vision and Pattern Recognition (2024)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)

41. Su, J.W., Chu, H.K., Huang, J.B.: Instance-aware image colorization. In: Conference on Computer Vision and Pattern Recognition (2020)
42. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: Conference on Computer Vision and Pattern Recognition (2023)
43. Thasarathan, H., Nazeri, K., Ebrahimi, M.: Automatic temporally coherent video colorization. In: Conference on Computer and Robot Vision (2019)
44. Vitoria, P., Raad, L., Ballester, C.: ChromaGAN: Adversarial picture colorization with semantic class distribution. In: Winter Conference on Applications of Computer Vision (2020)
45. Wang, H., Chai, X., Wang, Y., Zhang, Y., Xie, R., Song, L.: Multimodal semantic-aware automatic colorization with diffusion prior. arXiv preprint arXiv:2404.16678 (2024)
46. Wang, Q., Chang, Y.Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. In: International Conference on Computer Vision (2023)
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. Transactions on Image Processing (2004)
48. Ward, R., Bigioi, D., Basak, S., Breslin, J.G., Corcoran, P.: LatentColorization: Latent diffusion-based speaker video colorization. Access (2024)
49. Weng, S., Sun, J., Li, Y., Li, S., Shi, B.: CT2: Colorization transformer via color tokens. In: European Conference on Computer Vision (2022)
50. Weng, S., Zhang, P., Li, Y., Li, S., Shi, B., et al.: L-CAD: Language-based colorization with any-level descriptions using diffusion priors. Conference and Workshop on Neural Information Processing Systems (2024)
51. Wu, Y., Chen, Z., Liu, S., Ren, Z., Wang, S.: CASA: Category-agnostic skeletal animal reconstruction. Conference and Workshop on Neural Information Processing Systems (2022)
52. Xu, Z., Wang, T., Fang, F., Sheng, Y., Zhang, G.: Stylization-based architecture for fast deep exemplar colorization. In: Conference on Computer Vision and Pattern Recognition (2020)
53. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Chang, H., Ramanan, D., Freeman, W.T., Liu, C.: LASR: Learning articulated shape reconstruction from a monocular video. In: Conference on Computer Vision and Pattern Recognition (2021)
54. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Liu, C., Ramanan, D.: ViSER: Video-specific surface embeddings for articulated 3D shape reconstruction. Conference and Workshop on Neural Information Processing Systems (2021)
55. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: FRESCO: Spatial-temporal correspondence for zero-Shot video translation. In: Conference on Computer Vision and Pattern Recognition (2024)
56. Yang, S., Mou, C., Yu, J., Wang, Y., Meng, X., Zhang, J.: Neural video fields editing. arXiv preprint arXiv:2312.08882 (2023)
57. Yang, Y., Dong, J., Tang, J., Pan, J.: ColorMNet: A memory-based deep spatial-temporal feature propagation network for video colorization. arXiv preprint arXiv:2404.06251 (2024)
58. Yang, Y., Pan, J., Peng, Z., Du, X., Tao, Z., Tang, J.: BiSTNet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. Transactions on Pattern Analysis and Machine Intelligence (2024)
59. Yatziv, L., Sapiro, G.: Fast image and video colorization using chrominance blending. Transactions on Image Processing (2006)
60. Ye, V., Li, Z., Tucker, R., Kanazawa, A., Snavely, N.: Deformable sprites for unsupervised video decomposition. In: Conference on Computer Vision and Pattern Recognition (2022)
61. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: Conference on Computer Vision and Pattern Recognition (2023)
62. Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., Chen, D.: Deep exemplar-based video colorization. In: Conference on Computer Vision and Pattern Recognition (2019)
63. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: International Conference on Computer Vision (2023)

64. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European Conference on Computer Vision (2016)
65. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Conference on Computer Vision and Pattern Recognition (2018)
66. Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. Transactions on Graphics (2017)
67. Zhang, Z., Liu, H., Chen, J., Xu, X.: GoodDrag: Towards good practices for drag editing with diffusion models. arXiv preprint arXiv:2404.07206 (2024)
68. Zhao, H., Wu, W., Liu, Y., He, D.: Color2Embed: Fast exemplar-based image colorization using color embeddings. arXiv preprint arXiv:2106.08017 (2021)
69. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-ControlNet: All-in-one control to text-to-image diffusion models. Conference and Workshop on Neural Information Processing Systems (2024)
70. Zhao, X., Guan, J., Fan, C., Xu, D., Lin, Y., Pan, H., Feng, P.: FastDrag: Manipulate anything in one step. arXiv preprint arXiv:2405.15769 (2024)
71. Zhao, Y., Po, L.M., Yu, W.Y., Rehman, Y.A.U., Liu, M., Zhang, Y., Ou, W.: VCGAN: Video colorization with hybrid generative adversarial network. Transactions on Multimedia (2022)