electronics

MDPI

Article

Video Colorization Based on Variational Autoencoder

Guangzi Zhang *, Xiaolin Hong, Yan Liu, Yulin Qian and Xingquan Cai

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; hxl@mail.ncut.edu.cn (X.H.); liuyan@mail.ncut.edu.cn (Y.L.); yulinqian@mail.ncut.edu.cn (Y.Q.); caixingquan@ncut.edu.cn (X.C.)

* Correspondence: guangzi@ncut.edu.cn

Abstract: This paper introduces a variational autoencoder network designed for video colorization using reference images, addressing the challenge of colorizing black-and-white videos. Although recent techniques perform well in some scenarios, they often struggle with color inconsistencies and artifacts in videos that feature complex scenes and long durations. To tackle this, we propose a variational autoencoder framework that incorporates spatio-temporal information for efficient video colorization. To improve temporal consistency, we unify semantic correspondence with color propagation, allowing for simultaneous guidance in colorizing grayscale video frames. Additionally, the variational autoencoder learns spatio-temporal feature representations by mapping video frames into a latent space through an encoder network. The decoder network then transforms these latent features back into color images. Compared to traditional coloring methods, our approach accurately captures temporal relationships between video frames, providing precise colorization while ensuring video consistency. To further enhance video quality, we apply a specialized loss function that constrains the generated output, ensuring that the colorized video remains spatio-temporally consistent and natural. Experimental results demonstrate that our method significantly improves the video colorization process.

Keywords: video colorization; temporal consistency; variational autoencoder

check for updates

Citation: Zhang, G.; Hong, X.; Liu, Y.; Qian, Y.; Cai, X. Video Colorization Based on Variational Autoencoder. *Electronics* **2024**, *13*, 2412. https://doi.org/10.3390/electronics13122412

Academic Editor: Jitae Shin

Received: 16 May 2024
Revised: 14 June 2024
Accepted: 17 June 2024
Published: 20 June 2024

1. Introduction

Colorizing black-and-white videos is a challenging task that requires not only accurate color application but also maintaining temporal consistency across frames. This technique is valuable in various fields, such as film and television production, education, and cultural preservation. While significant progress has been made in image colorization, extending these techniques to videos remains complex. Researchers like Iizuka [1], Zhang [2], and Larsson [3] have made strides in integrating image colorization methods into video colorization, showing promising results on grayscale images. However, these methods struggle when applied directly to videos, often failing to maintain temporal consistency and resulting in flickering and artifacts.

To address this, Lai [4] and colleagues proposed end-to-end post-processing methods to enhance temporal consistency. While somewhat effective, these methods still struggle to ensure smooth continuity between color frames, often resulting in color fading and blurring. Additionally, the need to process each video frame twice significantly increases the processing time, reducing the efficiency of the colorization process.

Recent advancements in fully automated colorization techniques, developed by researchers like Zhao [5], Deshpande [6], and others, utilize large-scale datasets to learn color semantics. While these methods offer significant convenience, they face several challenges. These include color inaccuracies due to insufficient training data, slow processing speeds when handling large datasets, and poor generalization caused by model biases learned from the training data.

These challenges can be mitigated through reference picture-based methods. These approaches use a specified color reference image to guide the colorization of entire grayscale video frames, as exemplified by Zhang et al.'s [7] sample-based video coloring method. Although these methods show promising results, they often fall short of delivering fully satisfactory visual effects. Therefore, we propose optimizations and enhancements to this framework.

Firstly, within the Semantic Correspondence Network module, we propose utilizing RESNET-50 for feature extraction from images. This involves extracting features incrementally at each stage and then merging these features to produce the final feature map. These feature maps serve a dual purpose: they facilitate the calculation of image similarity and, more importantly, they help the model understand video content by identifying and differentiating between various objects. This, in turn, enables accurate coloring at precise locations.

Secondly, to compute the similarity between the reference picture and the grayscale frame, we employ the attention mechanism. This mechanism automatically learns the interrelations between different color channels, enhancing the model's ability to capture correlations among various image features. Moreover, the attention mechanism is highly adaptable, automatically adjusting attention weights based on the feature distribution of different reference pictures. This adaptability improves the model's robustness and generalization in similarity calculations across images.

Finally, within the colorization network module, we employ a variational autoencoder (VAE) to ensure both spatio-temporal continuity and visual consistency of the video. By inputting video sequences, we map them into latent space through the encoder network and subsequently generate color frames via the decoder network. This approach offers greater training stability compared to the Generative Adversarial Network (GAN) used by Zhang et al. [7]. This stability simplifies the experimental process, reducing the need for extensive debugging and optimization, and results in videos that are more natural and realistic, aligning better with human intuition and perception. Additionally, VAEs typically have faster model convergence, requiring fewer iterations during training to achieve high performance. For video colorization tasks, this advantage can reduce training costs and speed up model deployment.

In terms of VAE's reconstruction capability, it excels in accurately reconstructing input data by learning latent representations from training data. During the coloring process, VAE effectively captures semantic information from video frames and strives to preserve original details, resulting in more-realistic coloring outcomes. These attributes collectively position VAE favorably compared to GAN for video colorization.

Our approach not only efficiently addresses video colorization but also achieves notable improvements in generation quality and processing speed. Through rigorous experimentation and comparison with existing methods, we validate the effectiveness of our approach. Our results demonstrate that our method produces vibrant and clear videos while operating at an accelerated pace.

In summary, this paper contributes to the field in several key ways:

1. We leverage RESNET-50 for comprehensive image feature extraction, utilizing deep-level features in a layered merging approach.
2. The integration of an attention mechanism enhances the model's ability to calculate image similarity across different reference images, thereby improving generalization.
3. Our novel video coloring method based on Variational AutoEncoder (VAE) effectively utilizes spatiotemporal data to ensure coherence and authenticity in generated videos.

Through this research, we aim to introduce innovative methodologies that advance the state of the art in video coloring techniques.

2. Related Work

In this section, we introduce related work on grayscale image and video colorization. We categorize these efforts into two main areas: image colorization and video colorization. Each of these areas can be further subdivided based on the specific methods employed.

2.1. Image Colorization

Image colorization has recently become a prominent topic in picture-to-picture translation. Adding appropriate colors to black-and-white images can enhance the accuracy of related tasks such as image segmentation and recognition. Colorization techniques can be broadly classified into two categories: example-based methods and fully automated methods.

Example-based approaches (e.g., Irony [8], Gupta [9], Zhao [10]) rely on user-provided doodles or reference images to colorize grayscale pictures. In doodle-based methods, the colorization process is driven by an optimization framework that spreads the given doodle colors across the entire image. The quality of the coloring largely depends on the colors and locations chosen by the user. In reference-image-based methods, deep learning techniques establish semantic correspondence between the grayscale image and the reference image, transferring color information from matching regions of the reference to the target grayscale image.

Levin et al. [11] proposed an interactive colorization technique based on the premise that adjacent pixels with similar intensity should have similar colors. Many subsequent doodle-based methods have refined and improved this approach. A major advantage of doodle-based methods is that users can select colors themselves, allowing them to determine the final style of the colorized image. However, these methods can be tedious and time-consuming, often requiring numerous scribbles for reliable results.

In contrast, reference-image-based colorization transfers color information from the reference image to matching regions in the target grayscale image. This approach is more efficient than doodle-based techniques but is dependent on the chosen reference image. Therefore, the reference image should visually resemble the target image to achieve optimal results.

In recent years, advancements in computer vision and deep learning have significantly propelled image colorization. Cheng [12] was among the first to introduce deep neural networks for colorization, using them to automatically map pixel features in grayscale images to color values. Baldassarre [13] proposed a model that combines convolutional neural networks (CNNs) with a pre-trained Inception-ResNet-v2 network for feature extraction. Furthermore, Generative Adversarial Networks (GANs), introduced by Goodfellow [14] and colleagues, have had a profound impact on this field. GANs consist of two competing neural networks: a generator that creates images from input data, and a discriminator that assesses the authenticity of these generated images against real images. Variants of GAN networks, such as DualGAN [15], ChromaGAN [16] and cGAN [17], have since been adapted specifically for image colorization.

Overall, image colorization is a dynamic field of research incorporating advances in deep learning, digital image processing, and computer vision. As technology continues to evolve, even more effective image colorization methods are on the horizon.

2.2. Video Colorization

Compared to image colorization, relatively little research has been conducted on video colorization [18–20]. Current methods can be grouped into three main categories: extensions of image-based colorization, fully automatic video colorization, and example-based video colorization.

Extensions of image-based methods: These methods treat a video as a collection of frames, processing each frame individually using image-based techniques, and then applying post-processing to ensure temporal consistency across the frames. Bonneel et al. [21] proposed a gradient-domain technique that provides color information to infer temporal relationships, guiding the colorization of uncolored frames. Lai [4] introduced an end-to-end recurrent neural network to improve temporal consistency. More recently, Lei et al. [22] developed a novel approach to address the temporal inconsistencies common in video algorithms derived from image-based techniques, demonstrating strong performance in experiments.

Fully automatic video colorization: These methods primarily rely on neural network models trained on large datasets like ImageNet-10k [23] and DAVIS [24], learning both video colorization and temporal correspondence. Lei [25] introduced an automatic video colorization approach emphasizing regularity and diversity, while Kouzouglidis [26] used a 3D conditional generative adversarial network to achieve automatic video colorization. Building on end-to-end video colorization, Zhao [27] proposed a hybrid loop method using a hybrid adversarial network.

Example-based video colorization: This methodology harnesses hue data from selected reference images, integrating it with monochrome video sequences. The process commences by identifying reference images that embody the sought-after color schemes, scenery, or subjects. Subsequently, a sophisticated deep learning algorithm correlates the chromatic attributes of these references with the monochrome frames, encompassing operations such as feature identification and hue alignment. This systematic approach guarantees the precise replication of colors in the monochrome frames. Innovative methods, including style transfer, facilitate the direct application of the reference's color palette onto video frames. Concurrently, advanced deep learning architectures like Generative Adversarial Network (GAN) and Variational AutoEncoder (VAE), are adept at transferring the color schemes from reference images onto grayscale frames. These models are refined through extensive training regimes on diverse datasets, enabling them to master intricate color mappings, thereby enhancing the colorization process. Scholars such as Zhang [7], Wan [28], Chen [29], and Iizuka [30] have predominantly embraced this technique for imparting color to grayscale video content, and it is this very technique that forms the crux of the present paper's investigation.

3. Methodology

3.1. Overall Framework

In this paper, we use N to represent the overall model architecture (as shown in Figure 1), where R denotes the feature extraction network, A is the feature association network, and C represents the final coloring network.

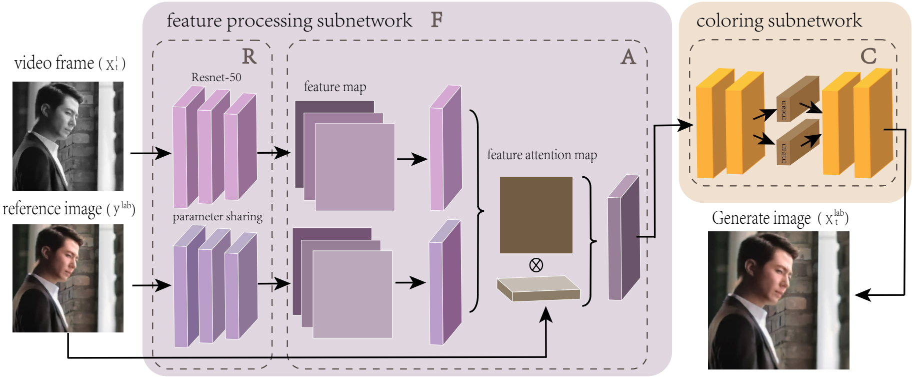

Figure 1. Overall network architecture diagram. Where F is feature processing subnetwork.

To start, assume the video consists of a series of frames, with the grayscale video frame at time t denoted as $X_t^l \in R^{H \times W \times 1}$ and the reference image as $y^{lab} \in R^{H \times W \times 3}$. Our experiments are conducted in the Lab color space, where *l* and *ab* represent the luminance and chromaticity of the color video frames, respectively. The ultimate objective is to generate a reasonable *ab*. To produce a coherent color video, we condition the colorization of frame X_t^l on the colorization result X_{t-1}^{lab} from the previous grayscale frame, as well as the reference image y^{lab}.

$$X_t^{lab} = N\left(X_t^1 \mid X_{t-1}^{lab}, y^{lab}\right) \tag{1}$$

3.2. Network Structure

Figure 1 shows the overall architecture diagram of our network. Each module is described in turn below.

3.2.1. Feature Processing Network F

We utilize RESNET-50, pre-trained for image classification, to extract information from the grayscale frame X_t^1 and the reference image y^{lab}, establishing a semantic correspondence between them. To accommodate different input dimensions, we removed the average pooling and fully connected layers at the top of RESNET-50 and added additional convolutional layers. This modification allows for flexible input processing. We then extract feature maps from multiple layers (as shown in Figure 2) and combine them to create multi-layer features $\Phi x, \Phi y \in R^{H \times W \times C}$ for the grayscale frame X_t^1 and the reference image y^{lab}, respectively.

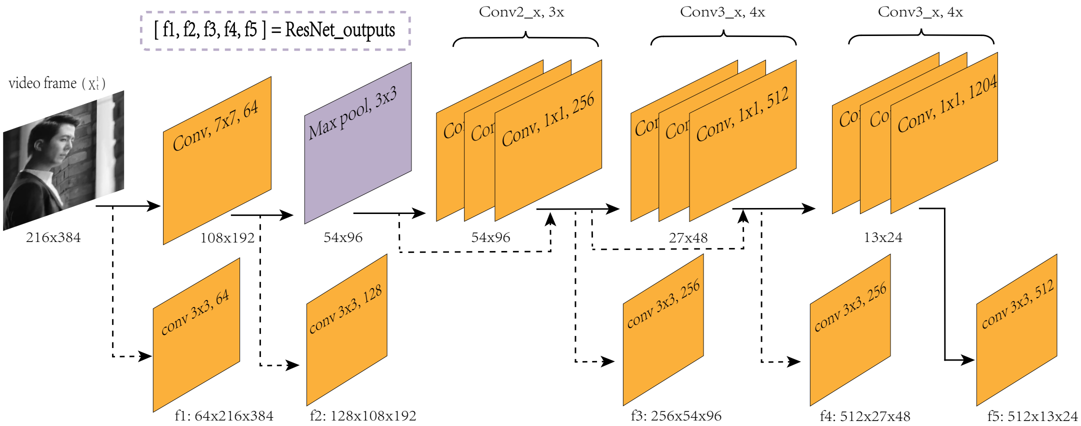

Figure 2. Image feature extraction diagram.

In the feature similarity section, we implement an attention mechanism to establish a dense correspondence between the grayscale frame X_t^l and the reference image y^{lab} (as shown in Figure 3). First, we compute the similarity matrix f between them and then convert the matrix into a corresponding similarity feature map. The similarity matrix is computed as follows:

$$\theta = \frac{self.theta\left(X_t^1\right)}{\| self.theta\left(X_t^1\right) \|_2 + \varepsilon} \tag{2}$$

$$\phi = \frac{self.theta\left(y^{lab}\right)}{\| self.theta\left(y^{lab}\right) \|_2 + \varepsilon} \tag{3}$$

$$f = \theta^T \phi \tag{4}$$

The input features are denoted as X_t^1, and the reference features as y^{lab}. First, the features are centered, and then $L2$ normalization is applied. These two steps yield the similarity matrix f.

Using the similarity matrix, we compute the weighted color W^{ab}. This weighted color approximates the pixels with the highest attention scores in the reference image, allowing

W^{ab} to serve as a reference for aligning colors and guiding the colorization process in the next step.

$$W_i^{ab} = \sum_j \text{softmax}\left(\frac{f(i,\ j)}{\tau}\right) y_i^{ab} \tag{5}$$

In summary, the feature processing network generates two outputs: the warped color W^{ab} and the feature attention map.

$$(W^{ab}, Atten_{map}) = F(X_t^1, y^{lab}) \tag{6}$$

In this formula, W^{ab} denotes the weighted color W^{ab} generated by the similarity matrix in the attention mapping module is used to guide the next step of the coloring process, map denotes the $Atten_{map}$, F denotes the feature processing sub-network, X_t^1 denotes the grayscale frame, and y^{lab}denotes the reference image.

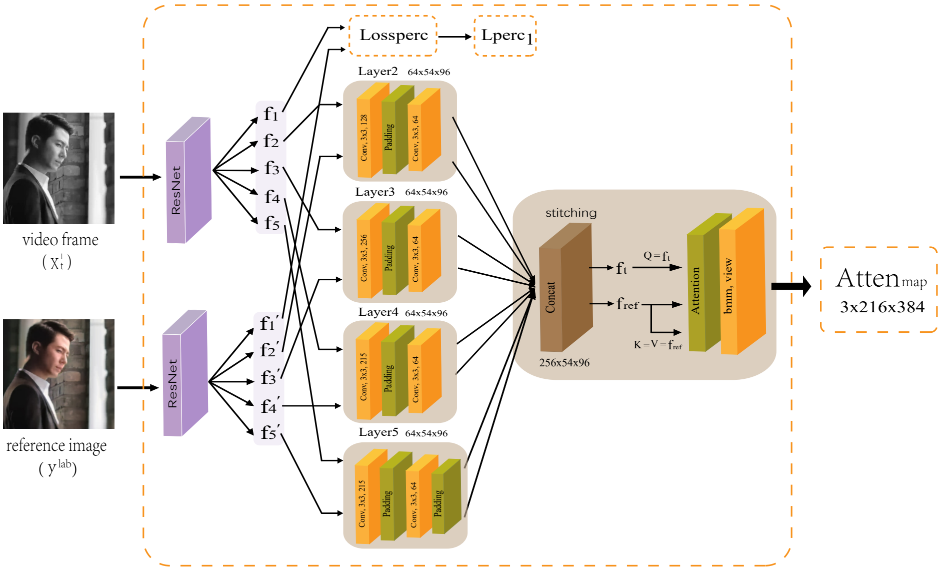

Figure 3. Similar feature calculation diagram.

3.2.2. Color Network C

Our main VAE network structure to achieve the coloring of grayscale frames, the coloring network (shown in Figure 4) C has four inputs, which are the distorted color map Wab, feature attention map ($Atten_{map}$,), reference picture y^{lab}, truth image X^{lab} and the previous moment's coloring frame X_{t-1}^{lab}. Eventually the coloring network C generates the predicted color picture X_t^{ab} of the current frame with the given luminance channel X_t^1 to obtain the final colored video frame X_t^{lab}.

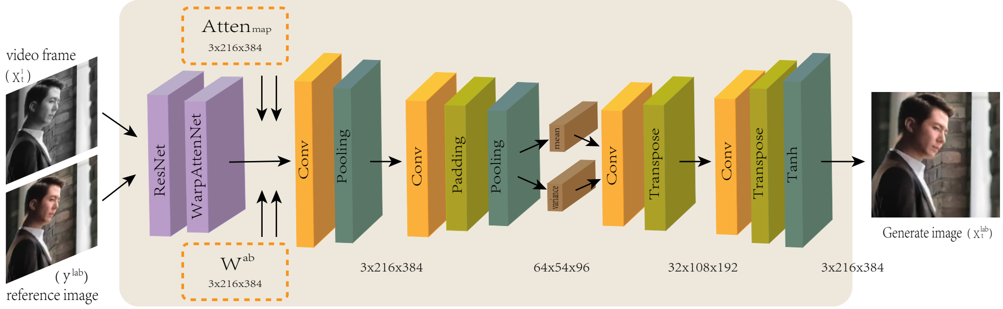

Figure 4. Colored network diagram.

$$X_t^{lab} = C(X_t^1, W^{ab}, Atten_{map}, y^{lab}) \tag{7}$$

3.3. Loss Function

The objective of our model is to produce coherent color videos without artifacts, while ensuring that the style of the generated video aligns with that of the reference image. Therefore, we employ specific loss functions to achieve this. In the coloring network C, the mean and variance of the latent space are first generated using the encoder of the VAE. Next, the mean and variance are sampled using a multilayer perceptron to obtain the latent variable $z = (mean, variance)$. Finally, the decoder maps these latent variables to the reconstructed output, $x = D(z)$. To meet this objective, we apply the following loss functions to the network.

3.3.1. Perceptual Loss

We use perceptual loss to measure the difference between generated video frames and real images and use L2 norm to constrain.

$$L_{prec} = \| \phi^L_{X_t^{lab}} - \phi^L_{X^{lab}} \|_2^2 \tag{8}$$

φL represents the feature map extracted from the last layer of the RESNET-50 network, and we set L to 5. X_t^{1ab}denotes the grayscale frame and X^{1ab} represents the truth image.

3.3.2. KL Divergence Loss

KL divergence loss calculates the difference between the mean and variance of the latent space and the prior distribution.

$$L_{\text{KL}} = \sum_{i=1} \left(1 + \log(variance_i) - mean_i^2 - variance_i^2\right) \tag{9}$$

In this formula, information about the mean and variance is included in the calculation and is responsible for calculating the KL divergence between two Gaussian distributions.

3.3.3. Context Loss

Context loss can be used as a loss function for image generation and image editing tasks, aiming to measure the semantic similarity between two images. First we can calculate the distance $dL(i,j)$ between each pair of feature points $(\varphi_x^L(i), \varphi_y^L(j))$ and then normalize it. Based on the above calculations, a similarity matrix $A(i,j)$ can be constructed to represent the similarity between each pixel in the two feature maps. To compute the loss based on the similarity matrix, we use select the most similar pixel of each pixel in the other feature map and compute the average value of the similarity as the loss value.

$$L_{context} = \sum_l W_L \left[-\log\left(\frac{1}{N_L} \sum_i \max f_j^L(i,j) \right) \right] \tag{10}$$

Here we use multiple feature maps: L equals 2 to 4. N_L represents the number of features of layer L, while W_L coefficients are set for higher-level features.

3.3.4. Smoothness Loss

We take the smoothness loss to promote coherence between neighboring frames, assuming that the pixels of neighboring frames should be similar if they have similar chromaticity in the real image. We expect the neighboring pixels of x_t to exhibit similarity if they share similar chrominance values in the ground truth image x_t. The smoothness loss can be positioned as the difference between the color of the current pixel and the weighted color of its 10 contiguous regions.

$$L_{smooth} = \frac{1}{N} \sum_{c \in (a,b)} \sum_{i} \left(\tilde{x}_t^c(i) - \sum_{j \in N(i)} w_{i,j} \tilde{x}_t^c(j) \right) \tag{11}$$

where $w_{i,j}$ is the WLS weight measuring neighborhood correlation.

Combining all the above losses, the overall goal we want to optimize is

$$Loss \ = \ \lambda_{prec} \mathrm{L}_{prec} + \lambda_{\mathrm{KL}} \mathrm{L}_{\mathrm{KL}} + \lambda_{context} \mathrm{L}_{context} + \lambda_{smooth} \mathrm{L}_{smooth} \tag{12}$$

We set L_{prec} = 0.001, L_{kL} = 1.0, $L_{content}$ = 0.2, L_{smooth} = 5.0. Through meticulous adjustment of these critical hyperparameters, we can achieve a finer equilibrium in the model's performance across perceptual similarity, distributional properties, semantic coherence, and smoothness. This refined balance leads to more gratifying generation outcomes overall.

4. Experiment

In this section, we first perform ablation experiments to investigate the effectiveness of the loss function and the attention mechanism, and then compare our method with the improved Zhang [7]-based method and that of Zhao et al. [31].

4.1. Efficiency and Datasets

In this small section, we will present the efficiency of each method and the dataset used for our model and related narratives, respectively.

4.1.1. Efficiency

First and foremost, all three of our models were meticulously trained utilizing the computational prowess of an RTX 2080Ti (11 GB) GPU coupled with a robust 12 VCPU INTEL(R) Xeon(R) Platinum 8255C CPU. The versions of Pytorch and Cuda are 1.5.1 and 10.1.

In order to compare the efficiency of each model, we used frames per second (FPS) and average processing time (calculations are all in milliseconds), and the results of averaging the three models after multiple trainings are shown in Table 1.

Table 1. Comparison of efficiency.

	FPS (Frames per Second)	**Average Process Time (Milliseconds)**
Deep	24.7638	399.29 ms
SVC	10.7024	937.60 ms
Ours	25.3438	389.03 ms

As illustrated in Table 1, our approach is superior to deep and SVC in two key aspects.

4.1.2. Datasets

In the feature processing stage, we employed the ImageNet-10k dataset for feature extraction pre-training. This dataset encompasses over 10,000 categories and approximately 15 million images, making it one of the most extensive publicly available image classification datasets. Leveraging this dataset enables our model to acquire highly diverse and robust visual representation features, serving as a solid foundation for subsequent coloring tasks.

During the colorization phase, we have selected 15 DAVIS videos at random to comprise the test set, while the remaining DAVIS videos constitute the training set. This partition ensures the independence of the test set, preventing the model from accessing test data during the training phase. Additionally, this random sampling method helps ensure the objectivity of the test results and provides an accurate reflection of the model's performance. Apart from the DAVIS dataset, we have proactively gathered 100 high-definition videos from online resources, encompassing a diverse array of scenarios including urban environments, landscapes, and human activities. This initiative aims to bolster the model's adaptability to a broad spectrum of real-world situations. We fine-tune the model on the DAVIS dataset and self-collected heterogeneous datasets, which not only enhances the generalization ability of the model, but also improves the colorization performance. We employ a variety of evaluation metrics including, but not limited to, Structural Similarity (SSIM) Peak Signal-to-Noise Ratio (PSNR), and Fréchet Inception Distance (FID) to comprehensively evaluate the model's colorization effectiveness.

The integration of both publicly available and self-collected datasets for training offers a holistic approach to enhance the model's performance in real-world scenarios. Primarily, the expansive ImageNet-10k dataset is leveraged for robust feature learning. Subsequently, the DAVIS dataset, along with the self-collected data, is utilized for training and fine-tuning, ensuring adaptability to diverse visual contexts. Lastly, the DAVIS dataset serves as the benchmark for evaluating the model's performance. This sequential approach to dataset utilization proves to be an effective strategy for comprehensive model training and refinement.

4.2. Ablation Experiment

4.2.1. Loss Function Analysis

We conducted an ablation study to evaluate the effectiveness of each loss function individually, as shown in Figure 5. When L_{prec} is removed, the coloring is still based on the reference image, but the resulting video contains artifacts due to the lack of a loss function to ensure semantic similarity between input and output. Without $L_{context}$, the output video does not resemble the reference image. If L_{smooth} is absent, the color information from the reference image fails to propagate consistently across the video frames. In the absence of L_{KL}, the generated video may appear faded. When all four loss functions are included, our complete model is able to produce vivid, coherent, and artifact-free color videos.

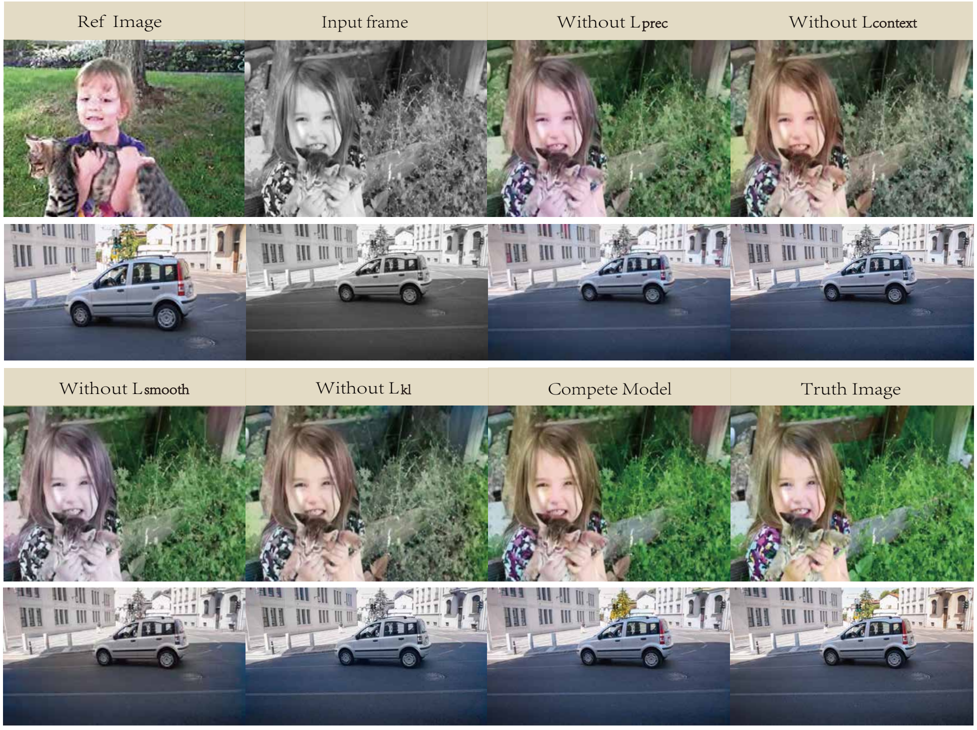

Figure 5. Comparison of loss function ablation.

We also calculated the appropriate metrics to demonstrate the ablation experiments, as shown in Table 2.

Table 2. Comparison of ablation metrics.

	SSIM	PSNR	FID	LPIPS
Without L_{prec}	0.957	29.33	71.82	0.18
Without L_{KL}	0.942	26.83	100.79	0.37
Without $L_{context}$	0.955	28.14	69.66	0.15
Without L_{smooth}	0.948	28.06	80.75	0.23
Complete model	0.977	31.80	46.04	0.10

The synthesized results indicate that our holistic model achieves excellence when evaluated across a spectrum of metrics. The distinct loss functions each play a crucial role in different dimensions of image synthesis, underscoring the notion that a well-crafted integration can substantially elevate the model's efficacy. These insights are instrumental in guiding the enhancement of our ablation study model.

4.2.2. Subnetwork Module Analysis

To substantiate the individual efficacy of the model components, an ablation analysis has been executed. The VGG-19 network is used for feature extraction, we omit the attention mechanism when computing similar feature maps, and the coloring network is implemented using a GAN. Figure 6 illustrates our coloring network based on RESNET-50, an attention mechanism, and VAE. Comparing these results with those from the VGG-19-based network demonstrates the superiority and effectiveness of our chosen architecture. Firstly, the VGG-19 network struggles to extract image features accurately, resulting in some features of the output video appearing blurry. Secondly, incorporating the attention mechanism makes color correspondence more precise. Lastly, the VAE network delivers more accurate coloring, providing realistic and consistent colors.

In the realm of reference imagery, we have isolated and tailored the inaugural frame of the source video to function as a reference, facilitating a more precise comparative analysis. The input frames undergo a digitization process, transitioning the original video into a grayscale sequence, which is then subjected to our network's modeling protocol. The evaluation is segmented into three distinct phases: initially, the deployment of the VGG-19 model, followed by an examination in the absence of a channel attention mechanism, and ultimately, the application of a GAN for colorization. Our comprehensive model, an ensemble of RESNET-50, an attention mechanism, and a VAE, is juxtaposed with the outcomes of these three distinct stages to assess the colorization proficiency.

Similarly, we similarly calculate metrics to make comparisons. The results are shown in Table 3.

Table 3. Comparison of module metrics.

	SSIM	PSNR	FID	LPIPS
VGG-19	0.976	24.02	33.76	0.72
No $Atten_{map}$	0.967	23.56	70.33	0.70
GAN	0.959	25.94	68.60	0.71
Complete model	0.977	31.80	46.04	0.10

At the outset, the model showcases an impressive performance when assessed using the metrics of Structural Similarity (SSIM) and Peak Signal-to-Noise Ratio (PSNR), a testament to the incorporation of the VGG-19 module. Nonetheless, it falls somewhat short in the realms of Fréchet Inception Distance (FID) and the nuanced Learned Perceptual Image Patch Similarity (LPIPS) measures. This disparity suggests that the VGG-19, while adept at producing images of visual likeness, might make slight sacrifices in the finer aspects of quality and authenticity. In contrast, the utilization of the RESNET-50 model could offer a strategic advantage in bolstering the model's perceptual acuity, thereby enhancing and refining the caliber of the synthesized imagery.

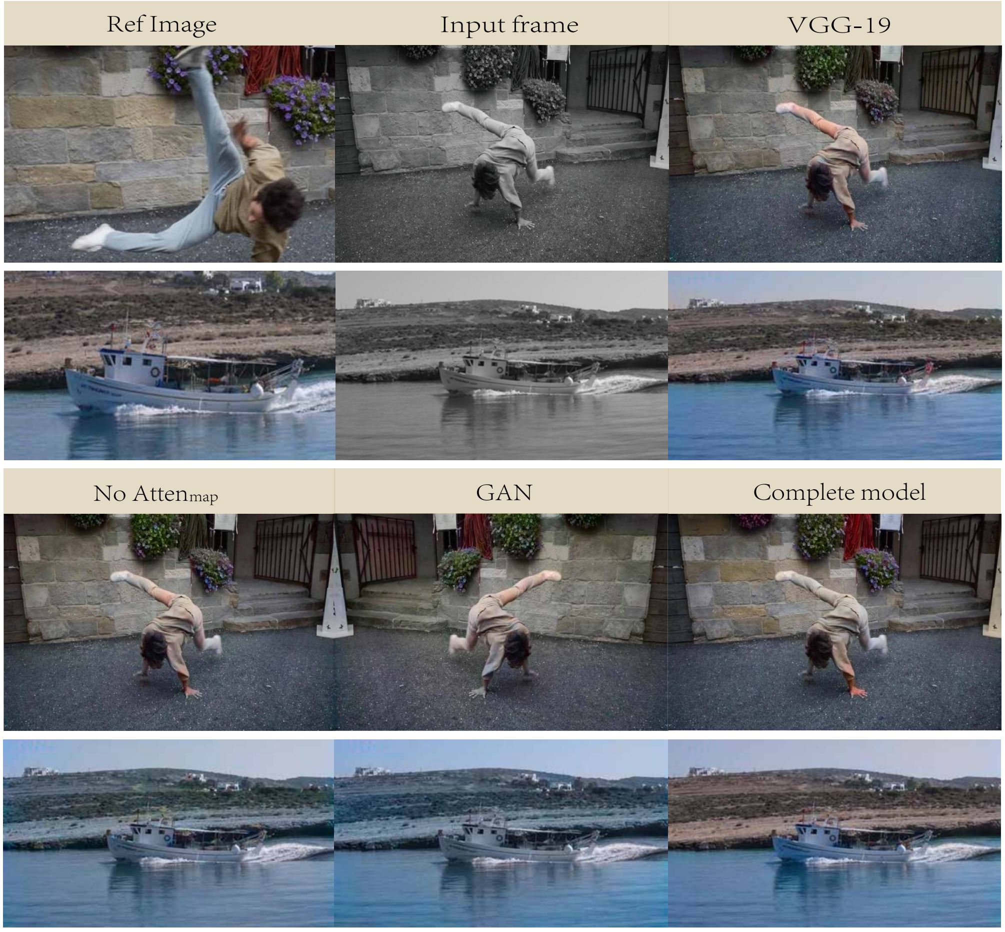

Figure 6. Comparison of subnetwork module ablation.

Additionally, the model's omission of an attention mechanism led to underwhelming results when evaluated through a comprehensive set of metrics. This suggests that neglecting the integration of an attention mechanism could limit the model's proficiency in critical areas, including feature extraction, which in turn could adversely affect the fidelity of the rendered images. The inclusion of such a mechanism is thus seen as a beneficial strategy for bolstering the model's overall efficacy.

Moreover, GAN-based networks, while demonstrating robust performance with respect to the Peak Signal-to-Noise Ratio (PSNR), may not fare as well in other evaluative categories. The deployment of GAN, although it can amplify the perceived quality of the generated imagery, might do so at the cost of structural congruence and a sense of authenticity. Conversely, VAE tend to produce images that are more faithful in terms of visual similarity and realism, indicating a distinct advantage in these specific dimensions.

Finally, the comprehensive model demonstrates outstanding performance across all metrics. This underscores the effectiveness of leveraging RESNET-50, attention mechanisms, and VAE in tandem, allowing for the full realization of their respective strengths and culminating in the attainment of optimal image generation results.

In summary, the findings illustrate the distinct contributions of various techniques within the ablation learning model. Effective utilization of RESNET-50, attentional mechanisms, and VAE not only enhances model performance but also offers valuable insights for refining and optimizing the ablation learning approach further.

4.3. Comparative Experiment

In our comparative experiments Table 4, our approach is assessed through both quantitative metrics and qualitative observations, juxtaposed against the latest deep learning-driven video colorization methodologies. We have selected the Deep and SVC models to serve as our points of reference. In the realm of image feature extraction, we have chosen the RESNET-50 architecture, which has demonstrated superior precision over our baseline models on the esteemed ImageNet test suite. The refined feature extraction capabilities of this network contribute to the creation of more impactful and contextually rich colorization outcomes. Additionally, our approach is fortified by the incorporation of an attention mechanism alongside a Variational Auto-Encoder (VAE), which synergistically amplify the overall visual and perceptual quality of the colorization process.

Table 4. Comparison table between RESNET-50 and VGG-19.

	Top-1	**Top-5**
RESNET-50	80.1%	93.0%
VGG-19	75.5%	92.4%

4.3.1. Qualitative Analysis

The specific experiment is shown in Figure 7. The first and second rows represent the input grayscale frames and the reference image, respectively, while the third through fifth rows depict the results of Deep, SVC, and our experimental method. According to the results, SVC tends to exhibit color overflow in the test samples, and Deep lacks vibrant colors, with some features appearing blurred. In comparison, our method demonstrates more vivid colors and fewer artifacts, delivering superior experimental results.

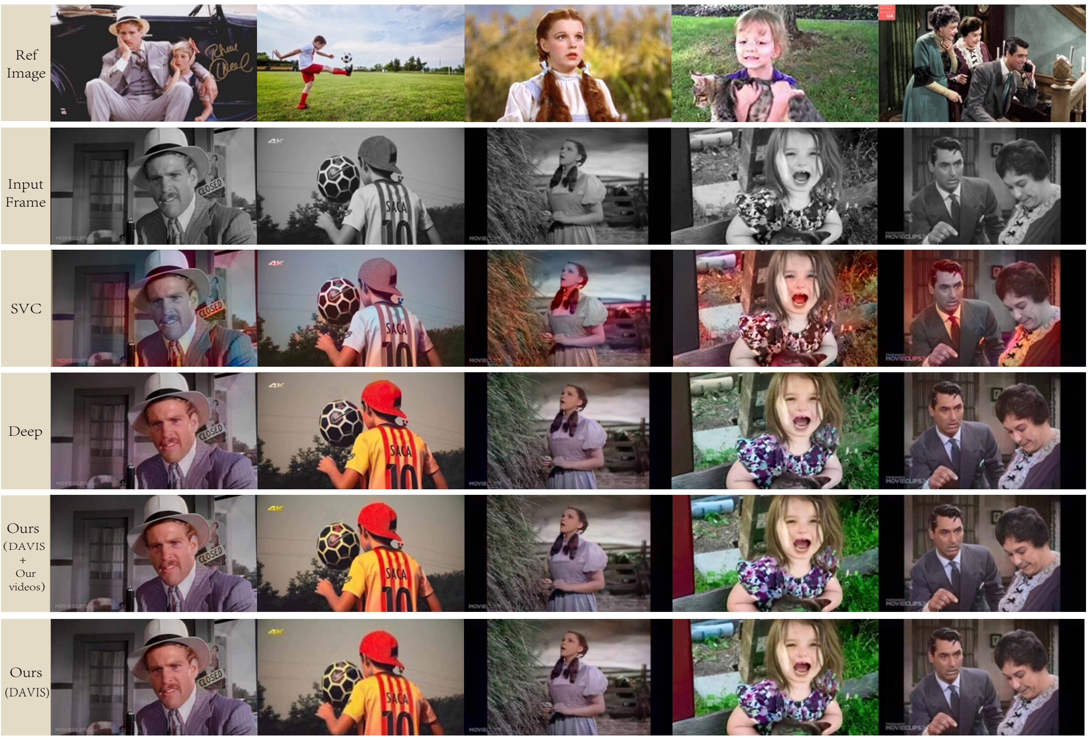

Figure 7. Comparison of experimental results.

This part demonstrates the effect of artifact processing, as shown in Figure 8.

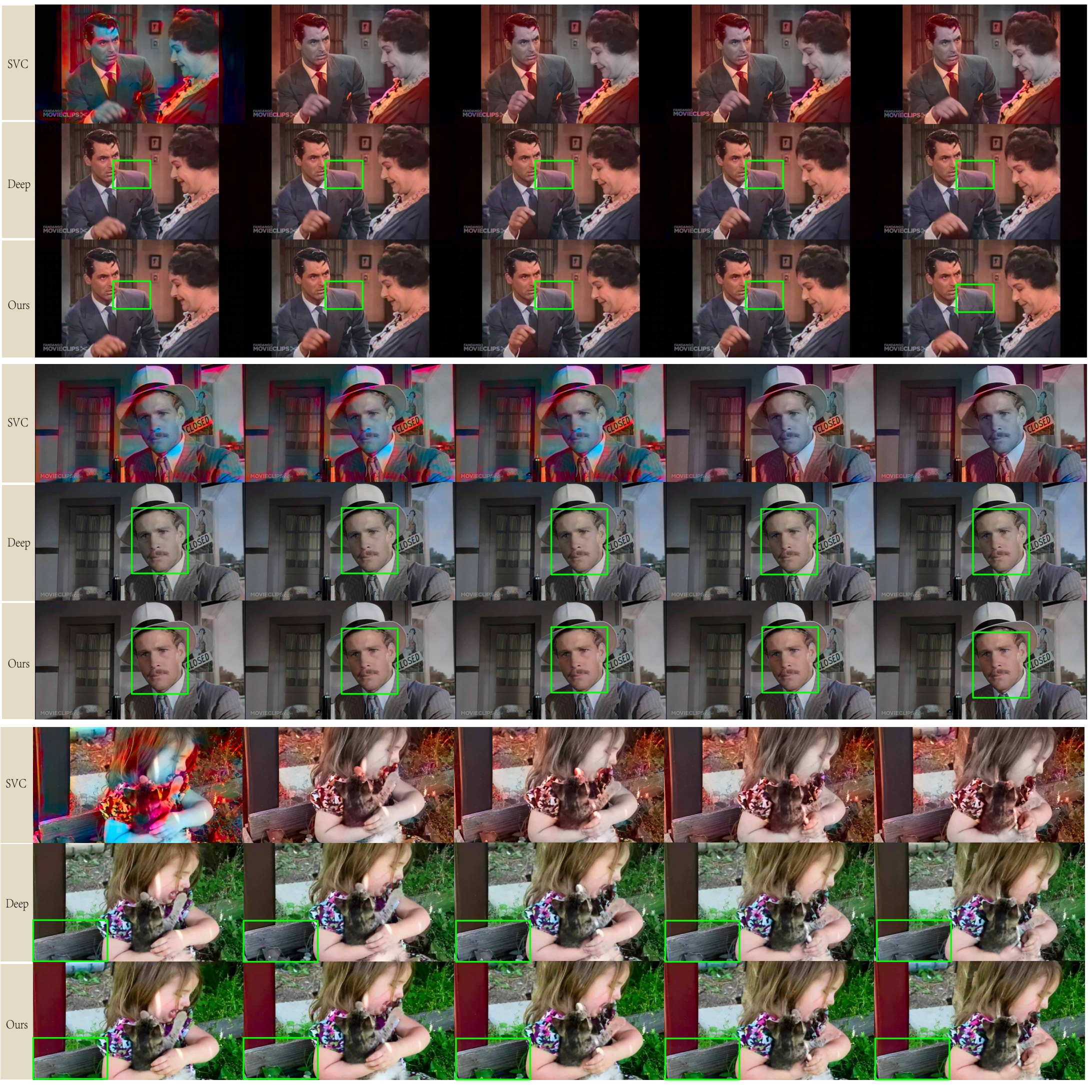

Figure 8. Comparison of artifact processing.

Finally the image generated by our method is compared with the real image and the result is shown in Figure 9.

4.3.2. Quantitative Analysis

In this section, we use two widely adopted indicators, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [32], Fréchet Inception Distance (FID) and the Learned Perceptual Image Patch Similarity(LPIPS) [33], to comprehensively assess the experimental effectiveness. PSNR measures the difference between the reconstructed image and the original image at the pixel level. It calculates the peak signal ratio between the reconstructed and original images. The formula for PSNR is as follows:

$$PSNR = 10 \times \log_{10} \frac{\left(2^k - 1\right)^2}{MSE} \tag{13}$$

$$MSE = \frac{1}{M \times N} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \left(f(x,y) - \tilde{f}(x,y)\right)^2 \tag{14}$$

where k denotes the number of binary bits corresponding to the image (typically 8) and MSE is the mean square error (which denotes the square of the pixel difference between the reconstructed image and the original image). The higher the value of PSNR, the smaller the difference between the reconstructed image and the original image, and the better the quality of the generated image result.

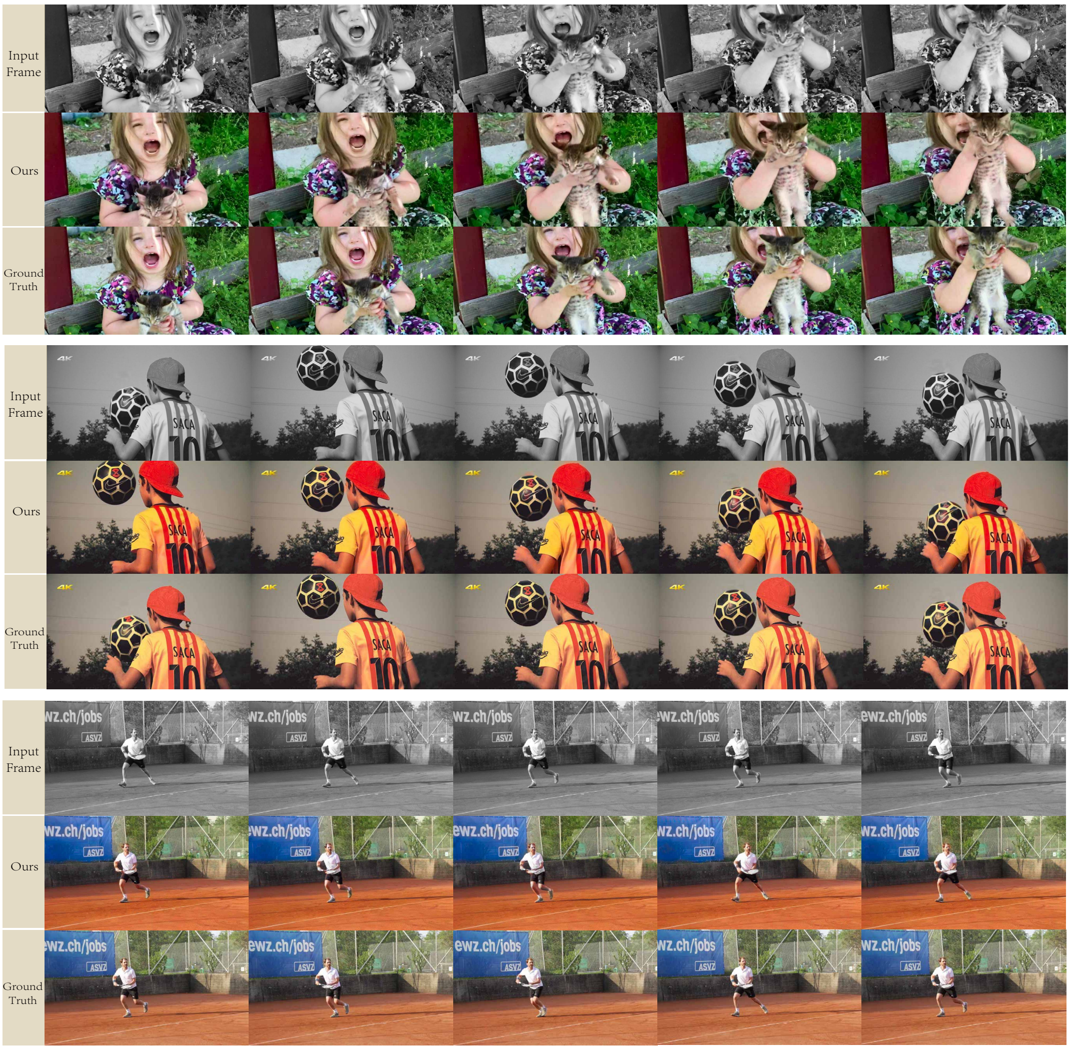

Figure 9. Comparison of real effect.

SSIM is a metric used to measure the structural similarity between images, which takes into account the information of brightness, contrast and junction at the same time, so it is more comprehensive compared to PSNR. Its calculation formula is as follows:

$$SSIM(x,\ y)\ = \frac{(2\mu_x\mu_y + C1)(2\sigma\_xy + C2)}{\left(\mu_x^2 + \mu_y^2 + C1\right)\left(\sigma_x^2 + \sigma_y^2 + C2\right)} \tag{15}$$

where x and y denote the local windows of the original and reconstructed images, respectively, μ denotes the mean of the pixel values, σ denotes the standard deviation of the pixel values, σ_{xy} denotes the covariance between the two images, and C1 and C2 are constants used for stabilization calculations. The range of the SSIM values is from -1 to 1, and the closer it is to 1 means that the structural similarity between the reconstructed and the original images is higher, and the quality of the reconstruction is better.

FID (Fréchet Inception Distance) is a metric used to evaluate the quality of generated generated images. It combines the realism and diversity of the generated images and is calculated by comparing the difference between the feature distribution of the generated image and the feature distribution of the real image. It combines the realism and diversity of the generated images and is calculated by comparing the difference between the feature distribution of the generated image and that of the real image.

Its calculation formula is as follows:

$$FID \;=\; \|\mu_{x^{lab}} - \mu_{x_t^{lab}}\|^2 + Tr\left(\sum_{x^{lab}} + \sum_{x_t^{lab}} - 2\left(\sum_{x^{lab}}\sum_{x_t^{lab}}\right)^{\frac{1}{2}}\right) \tag{16}$$

where μ_x^{lab} denotes the feature mean of the real image, $\mu_{x_t}^{lab}$ represents the feature mean of the generated image, x^{lab} represents the covariance matrix of the real image, x_t^{lab} represents the skewness matrix of the generated image and Tr denotes the trace of the matrix (i.e., the sum of the diagonal elements). The smaller FID is, the higher the quality of the generated image and the closer it is to the real image set.

LPIPS (Learned Perceptual Image Patch Similarity) is a metric for comparing the similarity between two images, which simulates human perception of images. Unlike traditional image similarity metrics, LPIPS takes into account the perceptual properties of images and is more in line with human perception of image quality and content. We define a LPIPS (Learned Perceptual Image Patch Similarity) computational function whose mathematical formula can be expressed as

$$LPIPS = \frac{1}{N}\sum_{i=1}^{N}\left|\phi_i\left(x^{lab}\right) - \phi_i\left(x_t^{lab}\right)\right| \tag{17}$$

where x^{lab} denotes the real image x_t^{lab} denotes the generated image N denotes the number of features. Similarly, the smaller the value of LPIPS, the better. We use pre-trained VGG-16 to extract their feature representations, then calculate the absolute difference between two feature representations, and finally average the absolute difference between all features. This averaged absolute difference value represents the perceptual similarity between the images x^{lab} and x_t^{lab}, that is, the extent to which they are visually different. We calculated the average PSNR, SSIM, FID, and LPIPS for the three experimental scenarios on the DAVIS dataset and the joint dataset (DAVIS dataset and our self-collected dataset), and the data tables are shown in Table 5.

Table 5. Indicator calculation table.

	SSIM	**PSNR(dB)**	**FID**	**LPIPS**
Deep	0.971	29.3	60.55	0.15
SVC	0.953	19.9	130.25	0.23
Ours (DAVIS)	0.976	30.0	46.53	0.14
Ours (DAVIS + our videos)	0.977	31.8	46.04	0.10

In order to show the comparison more clearly, we plotted the corresponding curves to visualize the experimental results. First, the PSNR curve is shown in Figure 10, illustrating the trend of PSNR values as the number of frames increases. Comparing the results, we observe that SVC consistently maintains lower PSNR values. While the Deep method initially performs better with fewer frames, our method consistently surpasses Deep as the frame count increases, ultimately achieving higher PSNR values and producing excellent experimental results.

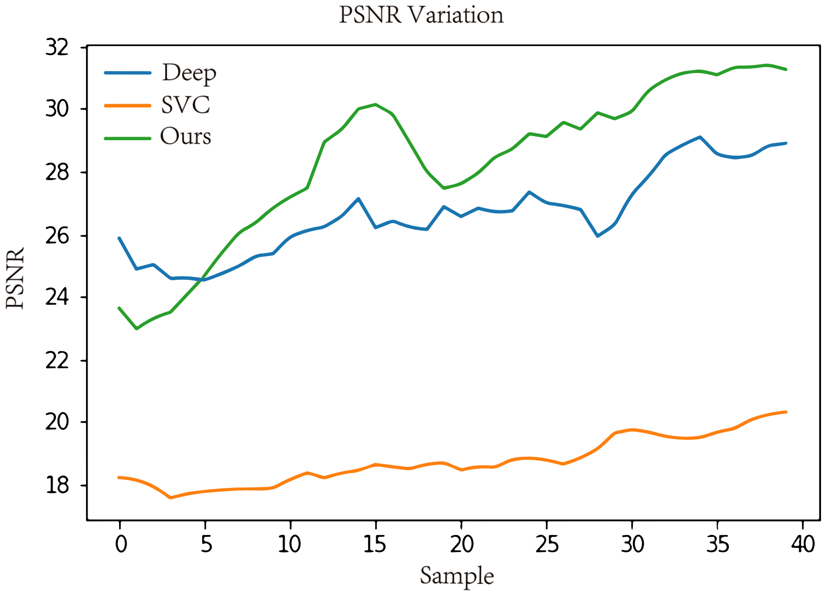

Figure 10. PSNR comparison chart.

The second curve is the SSIM, as shown in Figure 11. The SSIM values for all three methods exhibit similar trends, but our approach achieves the best results in most cases.

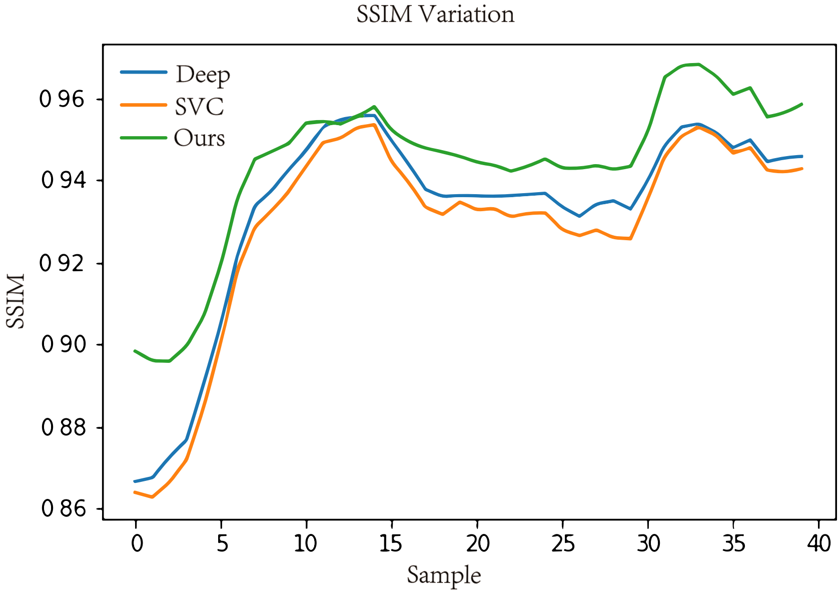

Figure 11. SSIM comparison chart.

The third curve is the FID, As illustrated in Figure 12. Initially, the FID value was higher than both of them. However, following training, our model effectively leveraged historical information, resulting in smoother coloring and consistently maintaining a very low FID value.

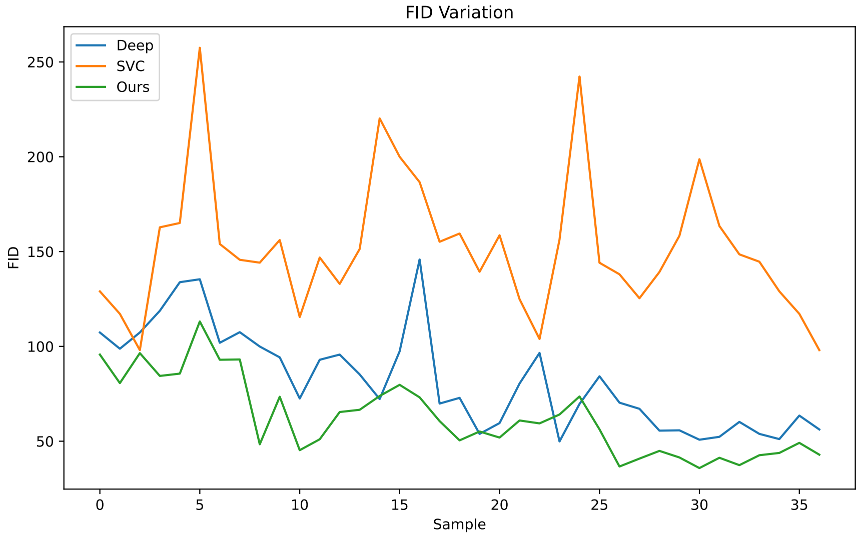

Figure 12. FID comparison chart.

The last curve represents LPIPS, As shown in Figure 13. Before training commenced, the LPIPS value was smaller compared to the other two methods. As training progressed, all three methods exhibited a similar trend. However, after a certain number of iterations, our approach began to stabilize.

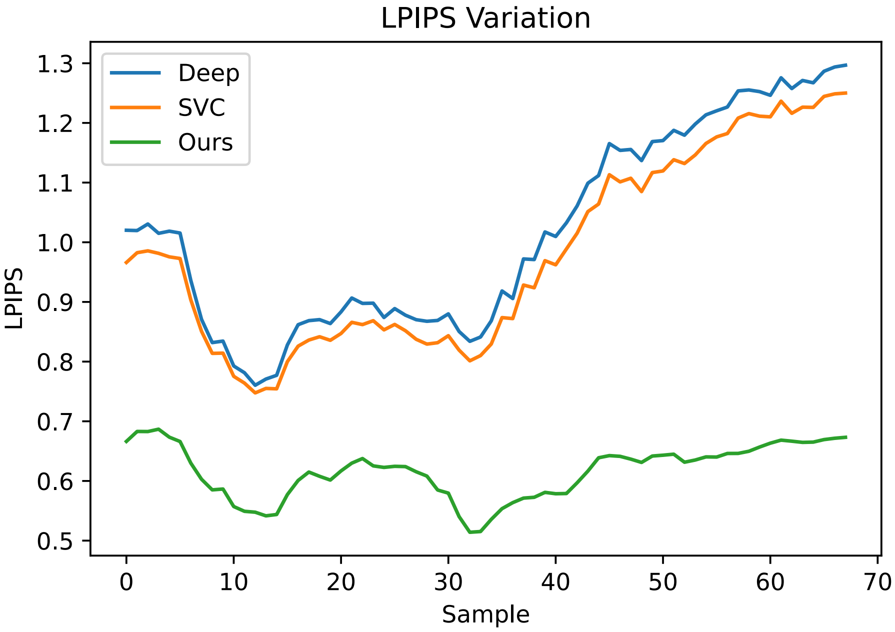

Figure 13. LPIPS comparison chart.

By comparing the experimental outcomes, we showcase the effectiveness and superiority of our approach. Using the four evaluation metrics, PSNR, SSIM, FID, and LPIPS, we can objectively quantify the differences between our method and others. In qualitative terms, our approach generates natural coloring results, while quantitatively it excels in all four metrics.

5. Conclusions

The video coloring method based on variational autoencoder holds immense potential for diverse applications. Firstly, it can significantly impact the field of historical image restoration by seamlessly combining historical scenes and color information, thereby transforming black and white movies and documentaries into vibrant, realistic portrayals of the past. This process enriches historical scenes, enabling audiences to connect with the depth and allure of history. Furthermore, this method is invaluable in movie and TV production for colorizing video footage of special effects scenes, thereby enhancing the overall environmental realism. Through meticulous coloring of special effects and scenes, the visual impact of films is elevated, captivating audiences and drawing them into immersive experiences. Lastly, this approach also has substantial implications for the digital industry, particularly in improving the quality and realism of game images. By applying colorization to game scenes and characters, visual performance is enhanced, ultimately elevating the gaming experience and transporting players into a more authentic virtual realm.

The details of the test results can be found in the experimental section in Section 4. We conducted a large number of experiments to validate and compare the results with traditional coloring methods. Through qualitative and quantitative analyses, we found that our proposed method has made significant progress in terms of accuracy and efficiency. Although there are still some challenges in dealing with complex scenes and multi-domain coloring, satisfactory results were achieved in general. These test results demonstrate the practicality of the method and lay the foundation for its application in real-world situations.

In this study, we propose a variational autoencoder for video coloring, which not only solves the difficulties faced in traditional video coloring but also makes progress in terms of accuracy and efficiency. By combining the variational self-encoder, we successfully improve the video coloring effect and also increase the processing efficiency. This opens up new opportunities for deep learning in video processing.

Despite the breakthrough of our approach, there are still some limitations that need to be further overcome. In particular, there are still some challenges in processing complex scenes and multi-domain colorization. Future research can focus on how to better handle these challenges to achieve more comprehensive video coloring effects. This will bring more possibilities to the field of video processing and promote the development of deep learning techniques in practical applications.

Author Contributions: G.Z. was responsible for the conceptualization, methodology, and the writing—reviewing and editing, as well as contributing to grant acquisition. X.H. was responsible for software development, data management, and the writing of the first draft, as well as visualization. Y.L. performed the mapping work. Y.Q. was responsible for format coding and revising the paper. X.C. performed project management. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Funding Project of Humanities and Social Sciences Foundation of the Ministry of Education in China, grant number 22YJAZH002.

Data Availability Statement: ImageNet-10k: Available at https://image-net.org/challenges/LSVRC/2010/ DAVIS: Available at https://davischallenge.org/davis2017/code.html.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Let There Be Color! *ACM Trans. Graph.* **2016**, *35*, 1–11. [CrossRef]
2. Zhang, R.; Isola, P.; Efros, A.A. Colorful Image Colorization. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 649–666. [CrossRef]
3. Larsson, G.; Maire, M.; Shakhnarovich, G. Learning Representations for Automatic Colorization. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 577–593. [CrossRef]
4. Lai, W.S.; Huang, J.B.; Wang, O.; Shechtman, E.; Yumer, E.; Yang, M.H. Learning Blind Video Temporal Consistency. *arXiv* **2018**, arXiv:1808.00449v1.
5. Zhao, J.; Liu, L.; Snoek, C.G.M.; Han, J.; Shao, L. Pixel-Level Semantics Guided Image Colorization. *arXiv* **2018**, arXiv:1808.00672.

6. Deshpande, A.; Rock, J.; Forsyth, D. Learning Large-Scale Automatic Image Colorization. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015. [CrossRef]
7. Zhang, B.; He, M.; Liao, J.; Sander, P.V.; Yuan, L.; Bermak, A.; Chen, D. Deep Exemplar-Based Video Colorization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019. [CrossRef]
8. Irony, R.; Cohen-Or, D.; Lischinski, D. Colorization by Example. In Proceedings of the Eurographics Symposium on Rendering Techniques, Konstanz, Germany, 29 June–1 July 2005.
9. Gupta, R.K.; Chia, A.Y.S.; Rajan, D.; Ng, E.S.; Huang, Z. Image Colorization Using Similar Images. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012. [CrossRef]
10. Zhao, H.; Wu, W.; Liu, Y.; He, D. Color2Embed: Fast Exemplar-Based Image Colorization Using Color Embeddings. *arXiv* **2021**, arXiv:2106.08017.
11. Levin, A.; Lischinski, D.; Weiss, Y. Colorization Using Optimization. *ACM Trans. Graph.* **2004**, *23*, 689–694. [CrossRef]
12. Cheng, Z.; Yang, Q.; Sheng, B. Deep Colorization. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015. [CrossRef]
13. Baldassarre, F.; Morín, D.G.; Rodés-Guirao, L. Deep Koalarization: Image Colorization using CNNs and Inception-ResNet. *arXiv* **2017**, arXiv:1712.03400.
14. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. GAN (Generative Adversarial Nets). *J. Jpn. Soc. Fuzzy Theory Intell. Inform.* **2017**, *29*, 177 . [CrossRef] [PubMed]
15. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [CrossRef]
16. Vitoria, P.; Raad, L.; Ballester, C. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020. [CrossRef]
17. Treneska, S.; Zdravevski, E.; Pires, I.M.; Lameski, P.; Gievska, S. GAN-Based Image Colorization for Self-Supervised Visual Feature Learning. *Sensors* **2022**, *22*, 1599. [CrossRef] [PubMed]
18. Zhang, L.; Liu, Y.; Wang, Z.; Yang, X. Temporally Consistent Video Colorization with Deep Feature Propagation and Self-regularization Learning. *arXiv* **2023**, arXiv:2304.08947.
19. Chen, H.; Yu, Q.; Wu, J.; Zhang, L. BiSTNet: Semantic Image Prior Guided Bidirectional Temporal Feature Fusion for Deep Exemplar-based Video Colorization. *arXiv* **2022**, arXiv:2212.02268.
20. Li, X.; Sun, L.; Jiang, J.; Gao, X. DeepExemplar: Deep Exemplar-based Video Colorization. *arXiv* **2022**, arXiv:2203.15797.
21. Bonneel, N.; Tompkin, J.; Sunkavalli, K.; Sun, D.; Paris, S.; Pfister, H. Blind Video Temporal Consistency. *ACM Trans. Graph.* **2015**, *34*, 6. [CrossRef]
22. Lei, C.; Xing, Y.; Chen, Q. Blind Video Temporal Consistency via Deep Video Prior. In Proceedings of the Neural Information Processing Systems, Virtual, 6–12 December 2020.
23. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
24. Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Gool, L.V.; Gross, M.; Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 724–732. [CrossRef]
25. Lei, C.; Chen, Q. Fully Automatic Video Colorization With Self-Regularization and Diversity. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019. [CrossRef]
26. Kouzouglidis, P.; Sfikas, G.; Nikou, C. Automatic Video Colorization Using 3D Conditional Generative Adversarial Networks. *arXiv* **2019**, arXiv:1905.03023v1.
27. Zhao, Y.; Po, L.M.; Yu, W.Y.; Ur Rehman, Y.A.; Liu, M.; Zhang, Y.; Ou, W. VCGAN: Video Colorization with Hybrid Generative Adversarial Network. *IEEE Trans. Multimed.* **2023**, *25*, 3017–3032. [CrossRef]
28. Wan, Z.; Zhang, B.; Chen, D.; Liao, J. Bringing old films back to life. *arXiv* **2022**, arXiv:2203.17276.
29. Chen, S.; Li, X.; Zhang, X.; Wang, M.; Zhang, Y.; Han, J.; Zhang, Y. Exemplar-Based Video Colorization with Long-Term Spatiotemporal Dependency. *arXiv* **2023**, arXiv:2303.15081.
30. Iizuka, S.; Simo-Serra, E. DeepRemaster. *ACM Trans. Graph.* **2019**, *38*, 1–13. [CrossRef]
31. Zhao, Y.; Po, L.M.; Liu, K.; Wang, X.; Yu, W.Y.; Xian, P.; Zhang, Y.; Liu, M. SVCNet: Scribble-Based Video Colorization Network with Temporal Aggregation. *arXiv* **2023**, arXiv:2303.11591.
32. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
33. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.