Video Colorization Based on a Diffusion Model Implementation

Leandro Stival[1,2(✉)], Ricardo da Silva Torres[2,3], and Helio Pedrini[1]

[1] Institute of Computing, University of Campinas, Av. Albert Einstein 1251, Campinas, SP, Brazil
{leandro.stival,helio}@ic.unicamp.br, leandro.stival@wur.nl

[2] Artificial Intelligence Group, Wageningen University and Research, Wageningen, The Netherlands
ricardo.dasilvatorres@wur.nl, ricardo.torres@ntnu.no

[3] Norwegian University of Science and Technology, Larsgårdsvegen 2, 6009 Alesund, Norway

Abstract. Cutting-edge techniques are being employed by researchers to develop algorithms that have the capability to automatically add color to black-and-white videos. This advancement has the potential to revolutionize our experience of historical films and provide filmmakers and video producers with a powerful new tool. These algorithms employ sophisticated deep neural networks to analyze images, identifying patterns and offering a promising avenue for extracting meaning and insights from visual data in the field of computer vision. Although current studies primarily focus on image colorization, there is a noticeable gap when it comes to videos and movies in the realm of deep machine learning techniques. Our investigation aims to bridge this gap and demonstrate that the image colorization techniques used today can also be effectively applied to videos and match the current state of the art presented at NTIRE 2023 video colorization challenge. We explored the application of diffusion models, which have gained popularity due to their ability to generate images and text. Our implementation involves utilizing a diffusion model to introduce noise in the frames, while a U-Net with self-attention layers predicts the denoised frames, thereby predicting the color of the video frames. For training purposes, we utilized the DAVIS and LDV datasets. When comparing the colorized frames with the ground truth in the test set, we observed promising results under several quality metrics, such as PSNR, SSIM, FID, and CDC.

Keywords: Video colorization · Deep learning diffusion models · Evaluation metrics

K. Arai (Ed.): IntelliSys 2024, LNNS 1065, pp. 117–131, 2024.
https://doi.org/10.1007/978-3-031-66329-1_10

1 Introduction

The role of color in shaping our perception and comprehension of the visual world is undeniable, particularly in the context of video. However, numerous older videos exist solely in black and white, imparting a sense of antiquity and disconnection from the contemporary world. To address this issue, researchers have employed deep learning models to restore the missing color information in these videos. This field of research, known as Deep Learning Video Colorization (DLVC), aims to develop algorithms capable of automatically adding color to black-and-white videos.

Additionally, the New Trends in Image Restoration and Enhancement (NTIRE) [21] challenge, proposed by the Computer Vision Foundation (CVF), provides an opportunity to further developments in this research area by the visual computing community. The challenge invites proposals for solutions in video colorization.

The current literature on image generation reflects a significant interest in diffusion models [8] even as the other computer vision areas [52]. These models draw inspiration from thermodynamics and, akin to the early neurons in neural networks, have played a role in the advancement of machine learning techniques. Recognizing this evolution in information generation architectures, our paper aims to demonstrate the application of Deep Diffusion Probabilistic Mode (DDPM) in generating the missing color channels within video frames.

1.1 Colorization

In the existing literature on DLVC, three prevalent methods are commonly employed to introduce colors into video frames: scribble-based, example-based, and fully automatic approaches [41].

Scribble-Based. Among the mentioned techniques, the scribble-based technique stands out as the traditional approach for accomplishing the task, dating back to the era before deep learning gained prominence. This method primarily involves transferring color information between adjacent pixels based on the similarities in their luminance values. However, it overlooks significant factors, such as texture and contextual details of the objects.

Consequently, the utilization of scribble-based approaches has become less common due to the extensive human intervention they demand. Nevertheless, there exist notable studies that employ this technique for DLVC (Deep Learning-based Video Colorization) tasks [9,14,32,56,56]. However, achieving satisfactory outcomes using this approach requires substantial human interaction throughout the process.

Example-Based. An alternative approach, widely adopted following the rise of machine learning, is the example-based technique applied to line art [1,39,43], video restoration [4,19,30], video flow [47], object flow [18], region similarity [57], neighbor pixels [51], spatial-temporal dependence [6].

We can highlight notable achievements in the area, exemplified by works such as BisTNet [57], ColorVid [48], and Deoldify [38]. These studies stand out in terms of frame colorization and color propagation measures, representing the forefront of advancements in colorization techniques. Consequently, they were selected for comparison with our own results.

Compared to the scribble-based approach, establishing the mapping between the example and the target frame in DLVC using the example-based method is a more complex task. In most implementations, this technique is employed to learn the correlation between the colors of the input and the luminance of the target frame. Consequently, its capacity to generalize effectively is directly influenced by the diversity of videos encountered during training.

Fully Automatic. Among the various methods for DLVC, this particular technique stands out prominently. Unlike the other approaches, it does not rely on pre-colored examples during the inference process for frame coloring. As a result, this method has been extensively explored in current DLVC solutions and is regarded as some of the state of the art in the area using simple GAN [25], temporal flow [27], auto-regulation propagation [26], and key frame color propagation [29].

Another aspect of this technique is its training objective. While other methods utilize coloration training models that map colors from examples to target frames, this approach focuses on comprehending the objects present in the image, their illumination, and the temporal characteristics to preserve consistent coloring. Consequently, these factors contribute to making example-less DLVC more challenging compared to the other methods.

For our implementation, we will adopt the fully-based technique as it offers greater control over the colorization process, allowing us to influence the colors generated by the model. Consequently, it becomes crucial to develop a method that can effectively inform the diffusion model about the objects present in the frame, along with other contextual information, in order to facilitate accurate color generation.

1.2 Contributions

Our main focus is on devising an approach utilizing a diffusion model to generate color for monochromatic frames, aiming to generate results that closely resemble the originally colored frames.

We can summarize our contributions as follows: (i) implementing an algorithm capable of performing monochrome video frame colorization using the conditional diffusion technique, (ii) achieving the ability to generate high-quality results for videos that were not included in the training data, thus establishing a generalist approach, and (iii) contributing to the advancement of diffusion models in DLVC tasks by introducing innovative architectural enhancements.

2 Diffusion Models

Diffusion Models, including Generative Adversarial Networks (GANs) [13], Variational Autoencoders (VAEs) [24], and certain Autoregressive models [34], fall under the category of generative models. The utilization of generative models has witnessed significant growth in recent years, leading to advancements in both the quality of generated outputs and the methods employed for achieving them [7].

The prevailing implementation of diffusion models in computer vision, found in nearly every generative problem within the field, is known as the Deep Diffusion Probabilistic Model (DDPM) [40]. Drawing upon principles from thermodynamics, these methods have demonstrated notable achievements in computer vision tasks, positioning them as the current state of the art in the field [8].

The application of DDPM models can be conceptualized as a high-level Markov Chain process [12]. During the forward pass of the network, small increments of noise are introduced over a finite number of steps. On the other hand, the reverse process, which constitutes the learnable part of the network, aims to predict the noise that was added to the original sample [17]. Typically, autoencoders are employed in the reverse process to learn and predict the added noise.

To illustrate this process, suppose a sample x without any noise, denoted as x_0. During the diffusion process, noise is incrementally introduced at each step t. As a result, the sample becomes progressively noisier, resulting in a sequence of samples $x_1, x_2, x_3, \ldots, x_t$. Ultimately, at the end of the diffusion process, the final sample x_t is predominantly composed of pure noise.

Conversely, the reverse process starts with a random noise sample, and the model's objective is to predict the noise that needs to be subtracted in order to reconstruct the original sample x. Starting from the noisy sample x_t, the reverse process gradually eliminates the noise until reaching the noise-free original version x_0. At a high level, we can consider that the model learns to predict a slightly denoised version x_{t-1} based on the input of x_t.

The implementation of DDPM models is gaining ground in several areas of visual computing, such as super-resolution [22], image generation [33], manipulation [23], and multimodal implementations such as inpainting [2], and segmentation [35].

3 Related Work

The relevance of colorization in the actual literature and the possibility of applications of the diffusion models are discussed in depth in this section.

Significant progress has been made in the field of image colorization using diffusion processes. Several works have emerged, each showcasing slight variations in implementation with the aim of achieving increasingly natural and realistic results. These advancements contribute to pushing the boundaries of image colorization techniques and further enhancing the overall quality of the generated outputs.

One significant advancement in image colorization is the utilization of unsupervised methods for posterior samplings. This approach enables faster training and inference, enhancing the overall efficiency of the process. Additionally, these methods have shown potential for application in other related problems, such as inpainting and deblurring [22].

Another implementation is the Generative Probabilistic Image Colorization (GPIM) [11], primarily employed for the colorization of line drawings. GPIM allows for the creation of multiple colorizations for the same input, adding flexibility and creativity to the process. Furthermore, this work introduces the use of positional embeddings, which facilitate the accurate filling of colors between the lines of the drawings, enhancing the quality and coherence of the colorization results.

The application of generative models using diffusion techniques offers various possibilities in the realm of video processing, including text-based editing. For instance, an example involves employing a pre-trained diffusion model in conjunction with keyframe generation to edit subsequent frames in a video sequence [5]. Other approaches focus on generating videos based on textual inputs [16,54,59]. In their methodology, the text descriptions are encoded into embeddings, and diffusion models are employed to generate videos that correspond to the provided textual descriptions.

Continuing with the application of DDPM in videos, another method involves video interpolation, where the model generates intermediate frames to preserve temporal consistency throughout the video [46].

While there has been a significant volume of work published on DDPM in the domain of video generation, we observe a disparity in terms of quantity and quality of results in the field of DLVC. This discrepancy highlights a research gap that exists in this area, which the present work seeks to bridge.

4 Methodology

In this section, we outline our methodology, starting with a description of the dataset utilized and its key characteristics. Then, we present the architecture of our model, followed by a detailed explanation of the training process.

4.1 Datasets

To train our model, we utilized two datasets, the DAVIS dataset [45] and the Large-scale Diverse Video (LDV) dataset [53], which DAVIS consists of 120 videos that were divided into training and validation sets, and LDV had 200 videos for training. Prior to feeding the images into the model, we resized them to dimensions of 224×224 pixels. Furthermore, we employed data augmentation techniques, such as random cropping and rotations. These augmentation operations were implemented to enhance the generalization capability of the dataset during training.

For our research, the dataset was divided into three distinct components. The first part consists of grayscale frames denoted as $S_g \in \mathcal{R}^{1\times H\times W}$. The second part consists of the original frames with color denoted as $R \in \mathcal{R}^{3\times H\times W}$. Finally, the output color frames generated by our model are represented as $S_c \in \mathcal{R}^{3\times H\times W}$. Here, H and W represent the height and width of the frames, respectively.

4.2 Model

Our architecture can be divided into four distinct deep-learning models, with diffusion serving as the primary model. In the following paragraphs, we will provide a breakdown of each of these models, elucidating their individual purposes and highlighting their interconnections within the overall architecture.

Encoder. The Encoder $\mathcal{E}$ is responsible for generating the latent space of the original frame R. The output of the Encoder is denoted as $\mathcal{L}at_R \in \mathcal{R}^{4\times 28\times 28}$, which serves as the prediction objective for the diffusion model. To ensure the quality of the generated latent space, we use a pre-trained model with weights [35] from ImageNet [37].

The aim of transforming video frames into latent spaces is twofold: simplifying the process and reducing the amount of data that the diffusion model has to handle. This trend towards diffusion processes in visual computing is illustrated by the use of this approach [35].

Visual Attention Conditioning. To guide the denoising process and facilitate the diffusion process, we incorporated the Visual Attention Conditioning (VAC) module. This module utilizes a pre-trained self-attention visual model, specifically the VIT_B_32 model provided by PyTorch [44], which employs Transformers for visual tasks [10].

In this process, the grayscale frame S_g is used to generate a latent representation denoted as $\mathcal{V}_{features} \in \mathcal{R}^{50\times 768}$, referred to as the *hotline*. This latent representation serves the purpose of guiding the diffusion model in determining how the colorized frame should be created. The *hotline* representation is directly incorporated into the convolution layers of our diffusion model, enabling effective colorization based on the provided information.

Diffusion Model Latent Colorization. The Diffusion Model Latent Colorization (DMLC) module is the core component of our implementation. It operates by taking a random noise distribution as input and utilizes the information from $\mathcal{V}features$ to generate a latent space representation denoted as $\mathcal{L}at_c$. This latent space represents a color version of $\mathcal{L}atR$, effectively removing noise from the original frame. To achieve this noise removal process, we employ an architecture inspired by the U-Net [36].

Similar to many recent implementations, our diffusion model generates information in the form of a latent space rather than directly from pixels [3]. This approach offers several advantages, including the ability to leverage existing pre-trained models and reduce the computational resources required for the diffusion process. To achieve this, the network responsible for creating the latent space

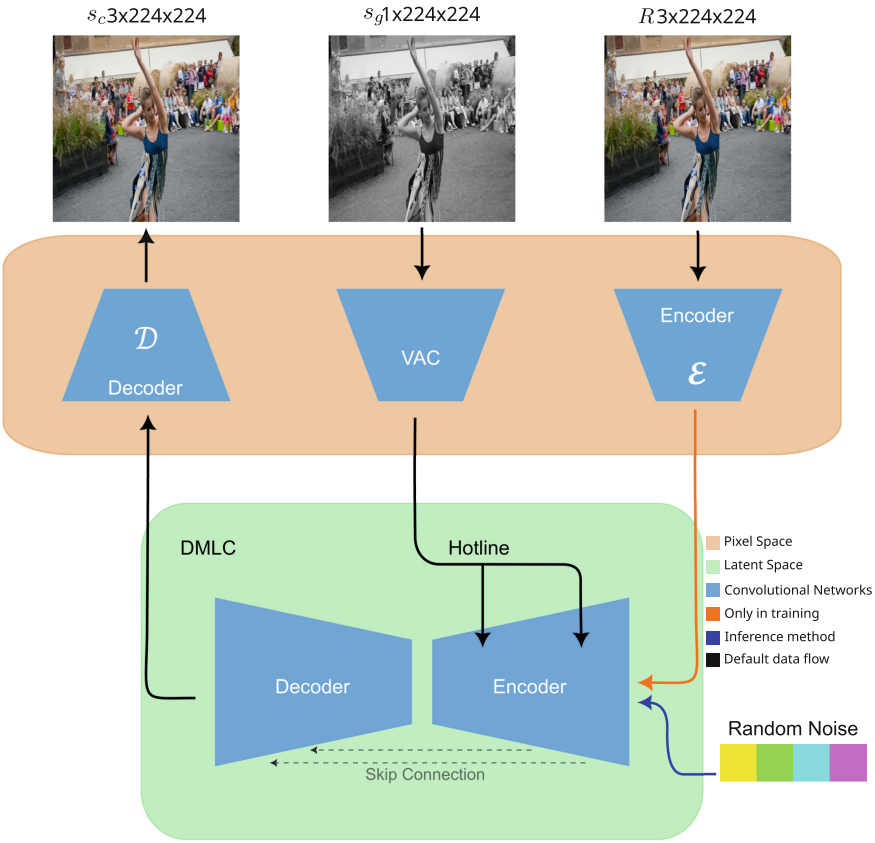


Fig. 1. Topology of our network, illustrating the data flow during both training and inference. The grayscale frame S_g is transformed into a colored version S_c using the output $\mathcal{L}at_R$ of the DMLC and decoded by $\mathcal{D}$. During training, the data flow is represented by the orange line, while during inference, the blue line demonstrates the process.

and decoding the diffusion output was pre-trained and is only used for inference during the diffusion process.

Decoder. The Decoder $\mathcal{D}$ is used as the final component of our architecture, responsible for converting the latent space obtained from the DMLC back into an image representation denoted as S_c. Similar to the Encoder $\mathcal{E}$, the implementation and weights of the Decoder $\mathcal{D}$ were pre-trained.

The interconnection between the models in our architecture is illustrated in Fig. 1, which depicts the flow of information through the different components, showcasing the relationships and dependencies among the models.

4.3 Training

To enhance the performance of each component in the architecture, we specifically train the DMLC model, enabling an isolated evaluation of the diffusion's effectiveness in the colorization task.

Our training batch consisted of 100 frames per iteration, and we conducted a total of 300 epochs. We employed the AdamW optimization method [28], a popular choice for diffusion and colorization problems, with a learning rate of $2e^{-5}$. The learning rate decayed by a factor of 0.1 every 50 epochs. For the DMLC model, we used the Mean Squared Error (MSE) as the loss function, this configuration had used in both datasets.

All experiments were conducted on a Windows 11 computer with the following specifications: AMD Ryzen-5600g (12 cores) running at 4.20 GHz, 32 GB RAM operating at 3.2 GHz, and a GPU setup consisting of an NVIDIA GeForce GTX 1080 Ti with 11 GB GDDR5 memory [31].

4.4 Evaluation

Our implementation requires only the monochromatic frame S_g as input for the inference process, resulting in the generation of the colored frame S_c.

5 Evaluation Metrics

In our work, we conducted a quantitative evaluation of color videos by comparing the colored frames with their corresponding original colored frames. This approach allows us to measure the difference between the output of our architecture and the ground truth.

We employed several methods to compare the pixel information of each frame, including the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Metric (SSIM), and Fréchet Inception Distance (FID) and Color Distribution Consistency (CDC). In the following sections, we will provide more details on how each of these metrics is calculated and how their results can be interpreted.

5.1 Peak Signal-to-Noise Ratio

The Peak Signal-to-Noise Ratio (PSNR) directly operates on the pixel intensities by comparing the maximum possible value with the Mean Squared Error (MSE). Equation 1 demonstrates how the calculation can be performed for a single pair of frames.

$$\mathrm{PSNR} = 10 \log_{10} \frac{(L_{\max})^2}{\mathrm{MSE}} \tag{1}$$

where, L_{max} is the maximum intensity value. A higher PSNR value indicates a closer similarity between the ground truth and the created image. It is important to note that while PSNR provides a quantitative measure of image quality, it does not directly correspond to human visual perception. This is because PSNR

does not take into account the structural characteristics of objects in the images during the comparison process [58].

Despite its limitations in capturing human visual perception, PSNR remains one of the commonly used metrics for measuring and comparing image quality. Its simplicity and intuitive scoring make it widely adopted. When applied to videos, PSNR can capture both spatial and temporal issues by comparing frames individually, although it may not account for inconsistencies in frame sequences [49]. In the domain of DLVC, many works still rely on frame-by-frame comparison using PSNR as the evaluation metric [20,25,39,55].

5.2 Structural Similarity Index Metric

Unlike PSNR, the Structural Similarity Index Metric (SSIM) takes into account additional information about the images, such as contrast and structural patterns. This approach provides a more human-like way of observing and analyzing the inputs [50].

The calculation of SSIM involves comparing local image patches and considering their luminance, contrast, and structural similarities. Equation 2 presents a simplified version of the SSIM calculation.

$$\mathrm{SSIM}(a,b) = \frac{(2\mu_a\mu_b + C_1)(2\sigma_{ab} + C_2)}{(\mu_a^2 + \mu_b^2 + C_1)(\sigma_a^2 + \sigma_b^2 + C_2)} \tag{2}$$

where, a and b represent the original and processed frames, respectively. μ_a and μ_b denote the mean intensities of the frames, while σ_a and σ_b represent their respective standard deviations. The parameter C reflects the contrast between the images being compared, and it is calculated as $C_1 = (k_1 \cdot L)^2$ and $C_2 = (k_2 \cdot L)^2$, where $L = 2^x - 1$. The values of k_1 and k_2 are set to 0.001 and 0.003, respectively, and x represents the number of bits per pixel.

5.3 Fréchet Inception Distance

The Fréchet Inception Distance (FID) is a metric commonly used to assess the quality of generated images, initially introduced for evaluating generative machine learning models [15]. FID compares images by analyzing the feature space extracted by a deep learning model, typically utilizing Inception V3 as the reference model [42].

Utilizing the feature space instead of pixel-level comparison has shown to be an effective approach for measuring image similarity. The obtained distance value through FID represents the dissimilarity between the two samples. In DLVC methods, FID is commonly applied in supervised or semi-supervised approaches, where the ground truth color frame is available.

5.4 Color Distribution Consistency

While FID serves as a valuable metric for assessing colorization quality, an additional metric is essential to evaluate the consistency of color propagation across

video frames. To address this requirement, we opted for the Color Distribution Consistency (CDC) metric.

Quality assessment within the CDC metric is conducted using the Jensen-Shannon (JS) factor, which evaluates consecutive frames to gauge the similarity of color distribution between them. The resulting value is normalized, ranging from 0 to 1, similar to the FID metric. The calculation of CDC can be expressed as depicted in Eq. 3:

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

where N represents the number of frames in the video. $P_c(I^i)$ denotes the normalized probability distribution over the histogram of the image I^i across the color channels (r, g, b). The parameter t is the temporal distance between frames being compared. Thus, the values of t are responsible for defining the window size between the frames being evaluated.

To comprehensively evaluate the model's capability to consistently propagate color across various temporal distances, we employed the standard configuration with three different intervals for t ($t = 1$, $t = 2$, and $t = 3$). This approach allows us to assess the model's performance in propagating color between nearby (short-term) and more distant (long-term) frames effectively. The process is expressed in Eq. 4:

$$\mathrm{CDC} = \frac{1}{3}(\mathrm{CDC}_1 + \mathrm{CDC}_2 + \mathrm{CDC}_4) \tag{4}$$

Hence, our choice of evaluating our model and its various facets using these metrics serves the purpose of highlighting improvements over the current state-of-the-art methods. This comprehensive evaluation approach allows us to demonstrate the advancements and effectiveness of our proposed model.

6 Experimental Results

After the training phase, we proceeded to evaluate our model on a separate dataset containing samples that were not part of the training set these being the DAVIS test set and the LDV validation set, as well as the NTIRE. In this section, we present the quantitative results obtained using each metric, as well as the qualitative results showcasing examples of the colorization tables generated by our architecture.

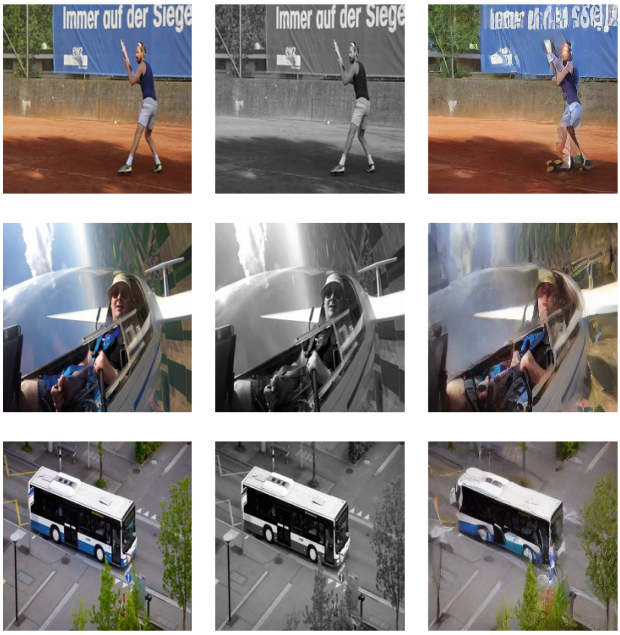
6.1 Quantitative Evaluation

The quantitative evaluation of our architecture was performed using three metrics: PSNR, SSIM, FID, and CDC. The results obtained are shown in Table 1.

All methods reported in Table 1 utilized the same dataset for evaluating their results. However, there was variation in the metrics used by different authors to demonstrate the quality of their models. In our evaluation, we opted to utilize the values provided by the respective authors and directly compare them with our own results.

Table 1. Comparison of results achieved by our architecture on the DAVIS dataset with other approaches

Comparison	DAVIS Dataset			
	PSNR ↑	SSIM ↑	CDC ↓	FID ↓
Lei et al. [26]	30.35	–	–	$6.22e^{-4}$
Chen et al. [6]	–	–	$4.02e^{-3}$	$5.87e^{-4}$
Huang et al. [18]	30.61	–	–	$6.87e^{-4}$
Ours	27.95	0.27	$3.19e^{-3}$	$5.02e^{-4}$



Original frame R Gray version S_g Colorized version S_c

Fig. 2. Results of our model's inference on the DAVIS dataset, illustrating the quality of frame colorization. While some instances exhibit color leaking, the results demonstrate a strong resemblance to the original colored images.

6.2 Qualitative Evaluation

Visual inspection is a commonly used qualitative method to evaluate DLVC (Deep Learning-based Video Colorization) models. It involves comparing the results of colorized images with their original colored counterparts.

In Fig. 2, we present the outcomes generated by our model, which include the original colored frame denoted as R, its monochromatic version labeled as S_g, and our implementation result represented as S_c.

7 Conclusions

In summary, the findings of our work demonstrate that the utilization of probability-based models for video colorization is a promising approach, yielding satisfactory results in terms of color quality for video frames. The models exhibited color fidelity when compared to the original samples, indicating that they can serve as a viable solution for video colorization tasks, including areas such as the restoration of old films and the recreation of historical content.

From the architectural innovations and comprehensive result evaluations, our approach provides a notable advancement in the utilization of diffusion models for DLVC problems. The outcomes of this study are anticipated to exert a positive influence on the development of more efficient and impactful models for addressing DLVC challenges. In conclusion, our research demonstrates that the conditional diffusion technique holds great promise as an approach for colorizing monochrome videos. The state of the art shows results comparable to ours, underscoring the potential of our architecture. It is worth noting that we achieved these results using limited hardware resources during model training.

Acknowledgments. We would like to thank CAPES, CNPq (grant #304836/2022-2), and FAPESP (grants #2022/12294-8 and #2023/11556-1) for their financial support. This work was also partially funded by the NorDark project, supported by NordForsk (grant #105116).

References

1. Akimoto, N., Hayakawa, A., Shin, A., Narihira, T.: Reference-Based Video Colorization with Spatiotemporal Correspondence. CoRR, bs/2011.12528:1–14 (2020)
2. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18208–18218 (2022)
3. Blattmann, A., Rombach, R., Oktay, K., Ommer, B.: Retrieval-Augmented Diffusion Models. https://arxiv.org/abs/2204.11824 (2022)
4. Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S., Pfister, H.: Blind video temporal consistency. ACM Trans. Graph. **34**(6), 1–9 (2015)
5. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2Video: Video editing using image diffusion. arXiv preprint. arXiv:2303.12688 (2023)
6. Chen, S., Li, X., Zhang, X., Wang, M., Zhang, Y., Han, J., Zhang, Y.: Exemplar-based video colorization with long-term spatiotemporal dependency. arXiv preprint. arXiv:2303.15081 (2023)
7. Croitoru, F.-A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: a survey. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
8. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Adv. Neural. Inf. Process. Syst. **34**, 8780–8794 (2021)
9. Doğan, P., Aydın, T.O., Stefanoski, N., Smolic, A.: Key-frame based spatiotemporal scribble propagation. In: Eurographics Workshop on Intelligent Cinematography and Editing, pp. 13–20 (2015)

10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations, vol. 1 (2021)
11. Furusawa, C., Kitaoka, S., Li, M., Odagiri, Y.: Generative probabilistic image colorization. arXiv preprint. arXiv:2109.14518 (2021)
12. Gagniuc, P.A.: Markov Chains: From Theory to Implementation and Experimentation. Wiley (2017)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Bing, X., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
14. Heu, J.-H., Hyun, D.-Y., Kim, C.-H., Lee, S.-U.: Image and video colorization based on prioritized source propagation. In: 16th IEEE International Conference on Image Processing, pp. 465–468. IEEE (2009)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Adv. Neural. Inf. Process. Syst. **30**, 1–12 (2017)
16. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J.: Imagen video: high definition video generation with diffusion models. arXiv preprint. arXiv:2210.02303 (2022)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
18. Huang, R., Li, S., Dai, W., Li, C., Zou, J., Xiong, H.: Improving optical flow inference for video colorization. In: IEEE International Symposium on Circuits and Systems, pp. 3185–3189. IEEE (2022)
19. Iizuka, S., Simo-Serra, E.: DeepRemaster: temporal source-reference attention networks for comprehensive video enhancement. ACM Trans. Graph. **38**(6), 1–13 (2019)
20. Jampour, M., Zare, M., Javidi, M.: Advanced Multi-GANs towards near to real image and video colorization. J. Ambient Intell. Hum. Comput. 1–18 (2022)
21. Kang, X., Lin, X., Zhang, K., Hui, Z., Xiang, W., He, J.-Y., Li, X., Ren, P., Xie, X., Timofte, R., Yang, Y., Pan, J., Peng, Z., Zhang, O., Dong, J., Tang, J., Li, J., Lin, C., Li, O., Liang, O., Gang, R., Liu, X., Feng, S., Liu, S., Wang, H., Feng, C., Bai, F., Zhang, Y., Shao, G., Wang, X., Lei, L., Chen, S., Zhang, Y., Xu, H., Liu, Z., Zhang, Z., Luo, Y., Zuo, Z.: NTIRE 2023 Video colorization challenge. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1570–1581 (2023)
22. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising Diffusion Restoration Models. arXiv preprint arXiv:2201.11793 (2022)
23. Kim, G., Kwon, T., Ye, J.C.: DiffusionCLIP: text-guided diffusion models for robust image manipulation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2426–2435 (2022)
24. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. arXiv preprint. arXiv:1312.6114 (2013)
25. Kouzouglidis, P., Sfikas, G., Nikou, C.: Automatic video colorization using 3D conditional generative adversarial networks. In: International Symposium on Visual Computing, pp. 209–218. Springer (2019)
26. Lei, C., Chen, O.: Fully automatic video colorization with self-regularization and diversity. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3753–3761 (2019)

27. Liu, Y., Zhao, H., Chan, K.C.K., Wang, X., Loy, C.C., Qiao, Y., Dong, C.: Temporally consistent video colorization with deep feature propagation and self-regularization learning. arXiv preprint. arXiv:2110.04562, pp. 1–17 (2021)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint. arXiv:1711.05101 (2017)
29. Mahajan, A., Patel, N., Kotak, A., Palkar, B.: An end-to-end approach for automatic and consistent colorization of Gray-Scale videos using deep-learning techniques. In: International Conference on Machine Intelligence and Data Science Applications, pp. 539–551. Springer (2021)
30. Meyer, S., Cornillère, V., Djelouah, A., Schroers, C., Gross, M.: Deep video color propagation. arXiv preprint. arXiv:1808.03232, pp. 1–15 (2018)
31. NVIDIA, Vingelmann, P., Fitzek, F.H.P.: CUDA, Release: 10.2.89, 2020. https://developer.nvidia.com/cuda-toolkit
32. Paul, S., Bhattacharya, S., Gupta, S.: Spatiotemporal colorization of video using 3D steerable pyramids. IEEE Trans. Circuits Syst. Video Technol. **27**(8), 1605–1619 (2017)
33. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: toward a meaningful and decodable representation. In: IEEE Conference on Computer Vision and Pattern Recognition (2022)
34. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831. PMLR (2021)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
36. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
37. Russakovsky, O., Deng, J., Hao, S., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
38. Salmona, A., Bouza, L., Delon, J.: DeOldify: a review and implementation of an automatic colorization method. Image Process. OnLine **12**, 347–368 (2022)
39. Shi, M., Zhang, J.-Q., Chen, S.-Y., Gao, L., Lai, Y.-K., Zhang, F.-L.: Deep line art video colorization with a few references. arXiv preprint. arXiv:2003.10685, pp. 1–10 (2020)
40. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
41. Stival, L., Pedrini, H.: Survey on video colorization: concepts, methods and applications. J. Signal Process. Syst. 1–24 (2023)
42. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
43. Thasarathan, H., Nazeri, K., Ebrahimi, M.: Automatic temporally coherent video colorization. In: 16th Conference on Computer and Robot Vision, pp. 189–194. IEEE (2019)
44. TorchVision: PyTorch's Computer Vision Library. https://github.com/pytorch/vision (2016)

45. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., Chen, L.-C.: FEELVOS: fast end-to-end embedding learning for video object segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9481–9490 (2019)
46. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Masked conditional video diffusion for prediction, generation, and interpolation. arXiv preprint. arXiv:2205.09853 (2022)
47. Vondrick, C., Shrivastava, A., Fathi, A., Guadarrama, S., Murphy, K.: Tracking emerges by colorizing videos. In: European Conference on Computer Vision, pp. 391–408 (2018)
48. Wan, Z., Zhang, B., Chen, D., Zhang, P., Chen, D., Liao, J., Wen, F.: Bringing old photos back to life. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2747–2757 (2020)
49. Wang, Y.: Survey of objective video quality measurements. EMC Corporation Hopkinton **1748**, 39 (2006)
50. Wang, Z., Lu, L., Bovik, A.C.: Video quality assessment based on structural distortion measurement. Signal Process. Image Commun. **19**(2):121–132 (2004)
51. Welsh, T., Ashikhmin, M., Mueller, K.: Transferring color to Greyscale images. In: 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 277–280 (2002)
52. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: text, images and variations all in one diffusion model. In: IEEE/CVF International Conference on Computer Vision, pp. 7754–7765 (2023)
53. Yang, R., Timofte, R.: NTIRE 2021 Challenge on quality enhancement of compressed video: dataset and study. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2021)
54. Yang, R., Srivastava, P., Mandt, S.: Diffusion probabilistic modeling for video generation. arXiv preprint. arXiv:2203.09481 (2022)
55. Yang, Y., Liu, Y., Yuan, H., Chu, Y.: Deep colorization: a channel attention-based CNN for video colorization. In: 5th International conference on image and graphics processing, pp. 275–280. Beijing, China (2022)
56. Yatziv, L., Sapiro, G.: Fast image and video colorization using Chrominance blending. IEEE Trans. Image Process. **15**(5), 1120–1129 (2006)
57. Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., Chen, D.: Deep exemplar-based video colorization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8052–8061. Long Beach, CA, USA (2019)
58. Zhang, L., Zhang, L., Mou, X., Zhang, D.: A comprehensive evaluation of full reference image quality assessment algorithms. In: 19th IEEE International Conference on Image Processing, pp. 1477–1480. IEEE (2012)
59. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: MagicVideo: efficient video generation with latent diffusion models. *arXiv preprint* arXiv:2211.11018 (2022)