IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 February 2024, accepted 20 May 2024, date of publication 27 May 2024, date of current version 14 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406249

RESEARCH ARTICLE

LatentColorization: Latent Diffusion-Based Speaker Video Colorization

RORY WARD[1,2], (Member, IEEE), DAN BIGIOI[2], (Graduate Student Member, IEEE), SHUBHAJIT BASAK[3], (Member, IEEE), JOHN G. BRESLIN[2], (Senior Member, IEEE), AND PETER CORCORAN[2], (Fellow, IEEE)

[1]SFI Centre for Research Training in Artificial Intelligence, Data Science Institute, University of Galway, Galway, H91 TK33 Ireland
[2]School of Engineering, University of Galway, Galway, H91 TK33 Ireland
[3]School of Computer Science, University of Galway, Galway, H91 TK33 Ireland

Corresponding author: Rory Ward (R.Ward15@nuigalway.ie)

This work was supported in part by the Science Foundation Ireland through the SFI Centre for Research Training in Artificial Intelligence under Grant 18/CRT/6223, in part by the Insight SFI Centre for Data Analytics under Grant 12/RC/2289_P2, and in part by the SFI Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant 18/CRT/6224.

ABSTRACT While current research predominantly focuses on image-based colorization, the domain of video-based colorization remains relatively unexplored. Many existing video colorization techniques operate frame-by-frame, often overlooking the critical aspect of temporal coherence between successive frames. This approach can result in inconsistencies across frames, leading to undesirable effects like flickering or abrupt color transitions between frames. To address these challenges, we combine the generative capabilities of a fine-tuned latent diffusion model with an autoregressive conditioning mechanism to ensure temporal consistency in automatic speaker video colorization. We demonstrate strong improvements on established quality metrics compared to existing methods, namely, PSNR, SSIM, FID, FVD, NIQE and BRISQUE. Specifically, we achieve an 18% improvement in performance when FVD is employed as the evaluation metric. Furthermore, we performed a subjective study, where users preferred LatentColorization to the existing state-of-the-art DeOldify 80% of the time. Our dataset combines conventional datasets and videos from television/movies. A short demonstration of our results can be seen in some example videos available at https://youtu.be/vDbzsZdFuxM.

INDEX TERMS Artificial intelligence, artificial neural networks, machine learning, computer vision, video colorization, latent diffusion, image colorization.

I. INTRODUCTION

With the rapid increase in the popularity of streaming video in recent years, today's media consumers have become accustomed to high-definition and vibrant video experiences, in color and on demand. However, there are also many substantial video archives with content that remains available in black and white only. Unlocking the potential of these archives, and infusing them with color, presents an exciting opportunity to engage with modern audiences, and breathe new life into classic movies and television episodes. By seamlessly blending cutting-edge technology with classic content, we not only enhance the visual appeal for contemporary viewers but also ensure that the historical significance of these timeless works are faithfully maintained.

The associate editor coordinating the review of this manuscript and approving it for publication was Zijian Zhang.

A. TRADITIONAL COLORIZATION

Colorizing black-and-white multimedia is a formidable challenge characterized by its inherent complexity. It presents a 'one-to-many' scenario, wherein multiple feasible colorization outcomes can be derived for a single black-and-white video, as illustrated by recent research [1].

Traditional approaches for video colorization are manual and labor-intensive, demanding the dedicated efforts of interdisciplinary teams comprised of skilled colorists and

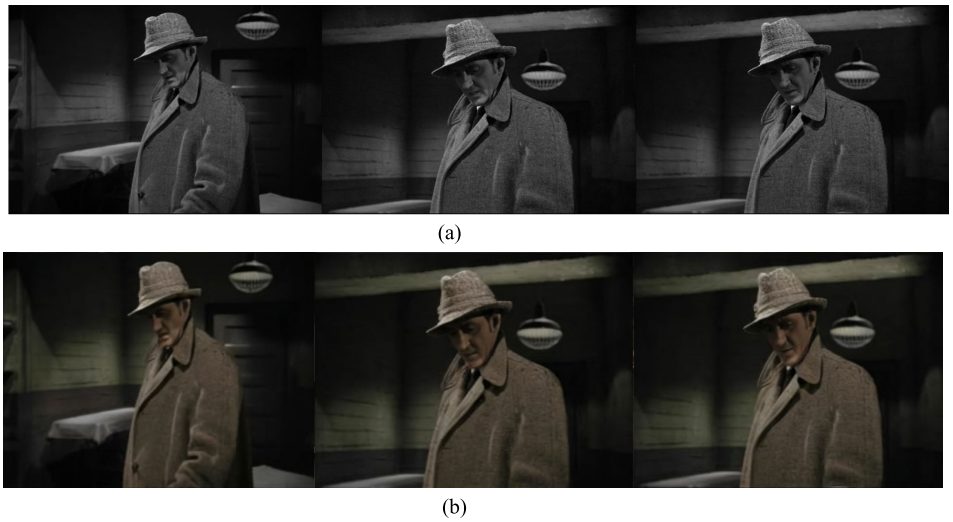


FIGURE 1. (a) "Sherlock Holmes and the Woman in Green" (1945) black-and-white frames. (b) "Sherlock Holmes and the Woman in Green" (1945) LatentColorization output frames.

rotoscoping animators, artists and historians. These teams invest extensive hours to ensure the production of a convincing and coherent end result. The intricacies of colorization are particularly difficult in the realm of videos, where the sheer volume of frames per second amplifies the complexity [2]. Therefore, automation of the video colorization process is highly desirable.

B. AUTOMATIC COLORIZATION

Automatic video colorization can be seen as a means to significantly reduce the cost traditionally associated with manually colorizing/restoring vintage movies, an expensive proposition that is often limited to organizations with substantial budgets. Since the labor costs associated with expert colorists are a significant barrier, manual colorization has also been largely limited to popular films or TV shows (e.g., Doctor Who), with numerous other works (social history movies, documentaries, films by lesser-known directors, etc.) omitted where the cost-benefit analysis could not justify their colorization.

As a consequence, various research efforts have tackled the need to automate aspects of the colorization process. These efforts span from earlier methods such as histogram matching [3], to more recent interactive approaches such as scribble-based systems [4] and exemplar-based approaches [5], as well as more recent developments in terms of deep learning-based colorization [6]. While the results still lag behind those that can be achieved of an experienced human colorizer, the automated approaches referred to above have made significant advancements in terms of their accuracy over prior systems.

In terms of the state-of-the-art, one current benchmark for automatic video colorization is held by Wan et al. [7]. However, it is important to note that their approach not only colorizes but also restores videos, making it a difficult benchmark for systems that are focused solely on colorization. DeOldify [6], provides colorized outputs without image restoration, and therefore can be more easily compared against colorization-only approaches such as the one presented in this paper.

Recent research [8] has shown the advantages of self-supervised learning methodologies for colorization, removing the resource-intensive need for creating and curating manually labelled datasets for training models. Constructing custom labelled datasets can be a resource-intensive and time-consuming endeavor, particularly when dealing with video content which has both static- and motion-related information.

C. RESEARCH CONTRIBUTION

Driven by the recent increase in the adoption of diffusion models [9], [10], [11], the field of generative modelling has produced a variety of contributions including Stable Diffusion [12], Imagen [13], and DALL•E 2 [14] which have gained attention in both research and the mainstream media.

Within the context of video colorization, the majority of techniques are based on GAN-based methods [15], [16], as well as the utilization of transformer-based approaches [17] such as those featured in [7], [18], and [19]. Notably, Saharia et al. [20] propose leveraging diffusion models for various image-to-image tasks, including colorization.

This paper introduces an innovative approach to video-based colorization, employing a latent-based denoising diffusion model. Our method demonstrates improvements over the state-of-the-art DeOldify [6] method, across a range of standard evaluation metrics, including Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), Fréchet Inception Distance (FID), Fréchet Video Distance (FVD), and Naturalness Image Quality Evaluator (NIQE). Furthermore, we provide comparative results for Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). It is also worth

noting that our method yields an average improvement of approximately 18% when FVD is employed as the evaluation metric. This result is also collaborated by our user study where LatentColorization is preferred 80% of the time to the previous state-of-the-art.

We introduce a novel system for achieving temporal consistency in video colorization through the application of a latent diffusion model. A sample visual, before and after, is given in Figures 1a and 1b.

To summarise, the unique contributions of our proposed work are as follows:

- We adapt fine-tuned latent diffusion models to the automatic speaker video colorization task.
- We ensure temporal consistency in automatic speaker video colorization using our autoregressive conditioning mechanism.

The structure of this paper is as follows: In §2, we examine related work. §3 provides an in-depth description of our methodology. Then, §4 presents the results of our evaluations, which are further examined in §5. Conclusions are given in §6, and we outline our future research directions in §7.

II. RELATED WORK

A. CONVENTIONAL DEEP LEARNING APPROACHES

Generative adversarial networks, commonly referred to as GANs [15], have emerged as a common technology in the enhancement of existing video content, in domains including sign-language addition [21], low-light enhancement [22], and video colorization [5]. GAN-based methods have also been extensively used for image colorization [23], [24], [25], [26], [27], [28], [29], [30]. For example, Isola et al. proposed Pix2Pix [23], which has performed well on various benchmarks, including the FID-5K benchmark using the ImageNet Val dataset. In the context of video colorization, DeOldify [6] and, more recently, Generative Color Prior (GCP) [31] stand out as two of the more prominent GAN-based approaches.

DeOldify [6] is a self-attention-based GAN [32]. It incorporates NoGAN training [33] and adheres to a Two Time Scale Update Rule [34]. While DeOldify is capable of generating credible colorizations, it has a tendency to produce somewhat subdued or less vibrant colors, characteristic of GAN-based systems.

GCP [31] leverages color priors encapsulated in a pre-trained GAN for automatic colorization. Specifically, they ''retrieve'' matched features (similar to exemplars) via a GAN encoder and then incorporate these features into the colorization process with feature modulations.

Other works, such as [35], [36], and [37], have also made contributions to the field of video colorization. It is important to note that GANs, due to their reliance on multiple loss functions, are challenging to train, susceptible to mode collapse, and often encounter convergence issues [38], [39], [40]. Furthermore, only certain GAN-based automatic colorization systems consider temporal consistency, such as Zhao et al. [41]. This means that the systems that do not account for temporal consistency do not maintain coherence across successive frames, which is a crucial aspect of video colorization.

Video Colorization with Hybrid Generative Adversarial Network (VCGAN) [41] is an end-to-end recurrent colourization network that prioritises temporal consistency in automatic video colorization.

DeepRemaster, as introduced by Iizuka and Simo-Serra in their work [42], is a Convolutional Neural Network (CNN)-based colorization system. As well as colorization, it also performs super-resolution, noise reduction, and contrast enhancement. Its performance makes it a suitable benchmark for comparison in our work.

Transformers, known for their success in diverse machine learning domains, including Natural Language Processing (NLP) and Computer Vision (CV), have achieved state-of-the-art results in various low-resolution computer vision tasks, exemplified by their second-place ranking on the FID-5K benchmark using the ImageNet Val dataset. However, the computational complexity of their self-attention mechanism scales significantly with higher image resolutions, presenting a challenge for handling high-resolution images [19], [43]. While ongoing research efforts aim to mitigate this challenge, it remains an open area of investigation. Unlike GANs, transformers exhibit greater resilience to mode collapse, thanks to their distinctive attention mechanism.

Kumar et al. have introduced the Colorization Transformer (ColTran) [18], a transformer-based image colorization model that operates through a three-step process. Initially, it colorizes a low-resolution version of the image, as it leverages self-attention, which is computationally demanding for high-resolution photos. Subsequently, it upscales the image and then the colors, yielding high-resolution colorized images. ColTran excels in producing vibrant colorizations, yet it falls short of catering to the specific demands of video colorization, leading to inconsistencies in video colorizations.

B. DIFFUSION MODELS

Diffusion models, as initially introduced by Sohl-Dickstein et al. [9], operate by learning how to reconstruct data from noise. They encompass two distinctive stages:

Forward Diffusion Process: In this phase, Gaussian noise is incrementally incorporated into the data through a step-wise progression spanning multiple timesteps. This gradual introduction of noise gradually transforms the original information until the desired level of diffusion or alteration is attained.

Reverse Diffusion Process: Subsequently, a learning model is employed to reverse this diffusion process, effectively reconstructing the original data [44], as illustrated in Fig. 2. Unlike GANs, diffusion models are resilient to

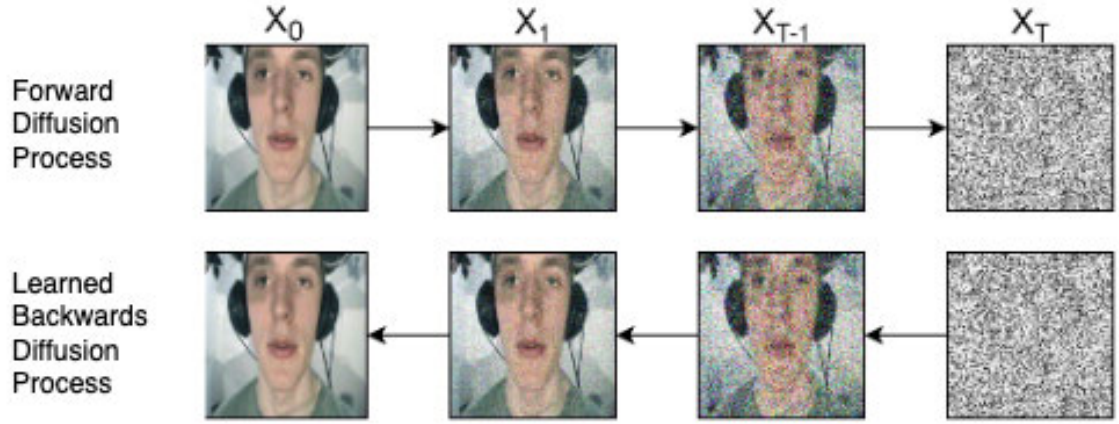


FIGURE 2. Diagram of the diffusion process: This diagram illustrates the operation of the diffusion model in both the forward and backward processes. In the forward process, it visually portrays the incremental addition of Gaussian noise to the input image x_0 until it becomes visually indistinguishable from Gaussian noise x_T (top). Subsequently, it showcases the learned backward diffusion process, where the model gradually removes the Gaussian noise from x_T to return to the original image x_0 (bottom).

mode collapse, and they have demonstrated success across various domains, including video generation [45], [46], audio generation [47], [48], and image generation [12], [14], [45].

An illustration of the application of diffusion models to still-image colorization can be found in Palette [49], a diffusion model tailored for a variety of image-to-image tasks. Palette attains the top position on the leader-board in the FID-5K benchmark using the ImageNet Val dataset.

Concurrently, Liu et al. [50] are engaged in research focused on the challenge of achieving temporally consistent video colorization, employing pre-trained diffusion models. A distinction lies in their approach as they utilize text-based conditioning for their system. In contrast, our methodology relies on exemplar frames as the conditioning input. This strategic choice was made based on our belief that using an image for conditioning provides a higher degree of expressive control compared to text-based approaches.

A challenge with diffusion models is their demanding computational requirements during both the training and testing phases. Nevertheless, ongoing research endeavors are actively addressing this issue [51], [52], [53]. Several approaches have emerged to mitigate this challenge:

Down-sampling and Super-resolution: Works such as Make-A-Video [54] tackle this issue by initially down-sampling the resolution of images in the diffusion process and subsequently restoring the resolution using a super-resolution algorithm.

Latent Diffusion: Another approach, exemplified by Latent Diffusion [12], modifies the diffusion process to operate in the latent space of a trained autoencoder, as opposed to the pixel space. This results in reductions in both inference and training times due to the reduced dimensionality of the data inputted into the diffusion process.

This paper presents the first work on the automatic speaker video colorization task using an image-to-image latent diffusion model adapted for video.

III. METHODOLOGY

A. DESIGN CONSIDERATIONS

One key consideration when designing an automatic speaker video colorization system is ensuring that the outputs are consistent throughout time. There are two ways to approach this: implicit temporal consistency and explicit temporal consistency.

- **Implicit Temporal Consistency:** In this approach, ensuring explicit temporal consistency is considered unnecessary. The belief is that with a sufficiently accurate system and reasonably similar input (e.g., consecutive frames in a video sequence), the colorized output should naturally exhibit similarity and relative consistency. As a result, temporal consistency is managed implicitly.
- **Explicit Temporal Consistency:** This project aligns with the second methodology, which emphasizes explicitly addressing temporal consistency. Rather than relying on the system to learn it implicitly, this approach involves conditioning for temporal consistency explicitly. The advantages of this approach include reduced training time, decreased data requirements, and a lower computational load. However, it necessitates more intricate system engineering to explicitly convey the requirements to the system.

Once the decision to use explicit temporal consistency has been made, a specific method must be implemented. Three of the most commonly used methods are:

- **Optical Flow-Based:** Optical flow-based colorization methods operate by conditioning the system to maintain color consistency over time. However, it is worth noting that a limitation of this approach is the potentially high computational cost associated with calculating optical flow, making it less practical in certain applications [55].
- **Exemplar-Based:** Exemplar-based methods provide the system with a reference image to guide its colorization process. This typically entails human intervention or

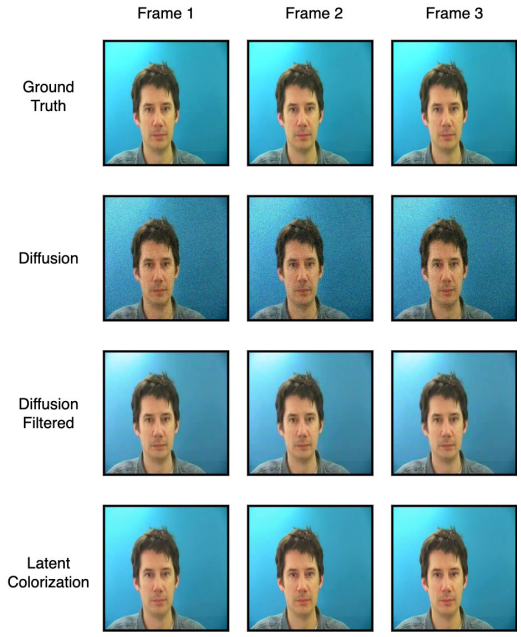


FIGURE 3. Comparison of 3 consecutive frames with different operations applied: First Row (Ground Truth): This row showcases the original, unaltered images, representing the ground truth reference. Second Row (Diffusion Model): In the second row, you can observe the colorization output generated by our original diffusion model. Third Row (Diffusion Model with Post-Processing): Here, the output of the diffusion model is presented with an additional post-processing procedure applied to enhance the results. Fourth Row (LatentColorization): The final row displays the results obtained from LatentColorization.

a database retrieval algorithm with a collection of reference images [56].

- **Hybrid-Based:** Some methods adopt a hybrid approach by combining different methodologies to harness the benefits of multiple systems simultaneously. This strategy, as seen in works like [5] and [57], seeks to leverage the strengths of various techniques to enhance overall performance.

LatentColorization uses an approach similar to exemplar-based, using an autoregressive conditioning mechanism which uses the previous frame as the exemplar. More details on this mechanism can be seen in III-D.

B. DATA PROCESSING

We use the following datasets as part of our experiments:

GRID Dataset: The GRID dataset [58] is a collection of video recordings featuring individuals speaking. It encompasses high-quality facial recordings of 1,000 sentences spoken by each of 34 talkers, with a distribution of 18 males and 16 females, resulting in a total of 34,000 sentences.

Lombard Grid Dataset: An extension of the GRID dataset, the Lombard Grid dataset [59], includes 54 talkers, each contributing 100 utterances. Among these 54 talkers, 30 are female, and 24 are male, expanding the dataset's diversity.

Sherlock Holmes Movies Dataset: This dataset is a collection of professionally colorized frames extracted from 'Sherlock Holmes and the Woman in Green,' 'Sherlock Holmes Dressed to Kill,' 'Sherlock Terror by Night,' and 'Sherlock Holmes and the Secret Weapon.'

These diverse datasets provide a foundation for our research in the field of speaker video colorization and temporally consistent diffusion models.

Our dataset consisted of 10,000 frames allocated for training the model, with an additional 700 frames reserved for testing purposes. Each frame was uniformly resized to 128×128 pixels.

To ensure the generalizability of our model, the training and testing frames were derived from distinct subjects, mitigating the risk of artificially inflated performance measures that would not extend to real-world scenarios.

By conducting tests on benchmark datasets, we could compare our approach against previous methods. Furthermore, testing on the Sherlock Holmes-related data provided a valuable means of comparing our results to expert human colorizations. Additionally, training on open-domain videos underscores the potential of these resources in advancing the field of automatic speaker video colorization.

C. SYSTEM OVERVIEW

1) IMAGE DIFFUSION BASED SET UP

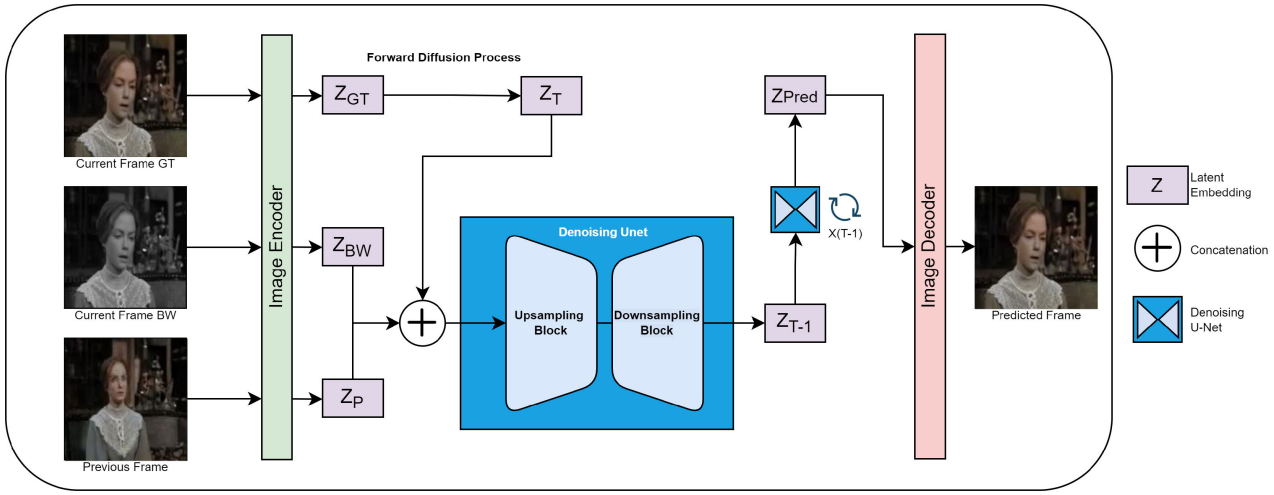
In our initial exploration, we considered adopting a setup akin to Palette [49], incorporating our temporal consistency mechanism and initial frame biasing, which will be elaborated on in Section III-D. However, we observed sub-optimal performance from this configuration, as the system's outputs exhibited undesired residual speckled noise, as illustrated in Fig. 3.

To address the speckled noise in the diffusion colorization outputs, we explored two approaches:

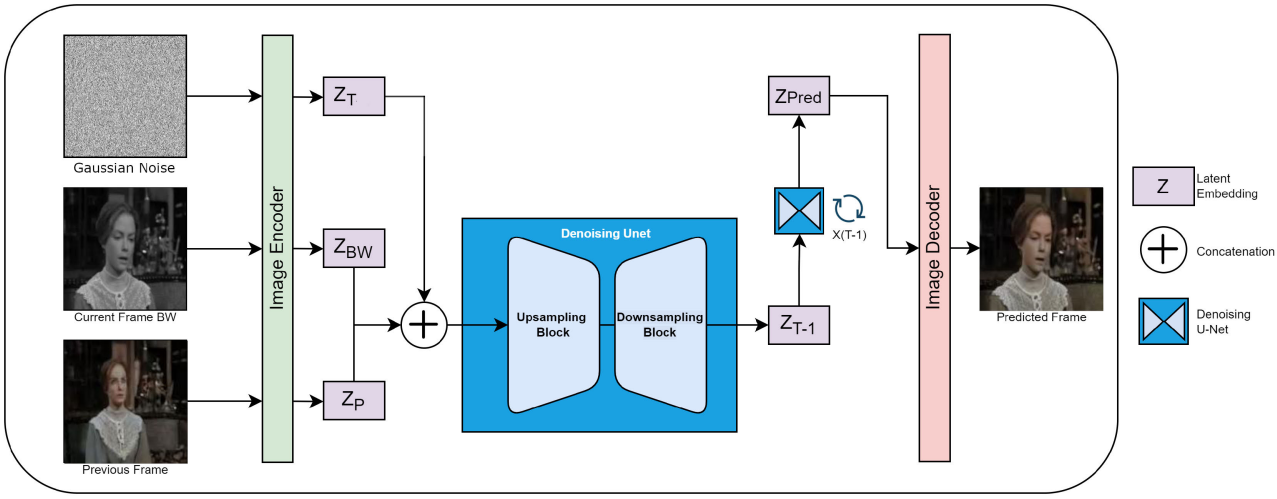
Non-Linear Means (nlmeans) Clustering: We initially applied the nlmeans clustering algorithm [60] to the images to mitigate the noise. However, this method relies on a hyper-parameter that dictates the filter's strength. A stronger filter results in smoother images but may inadvertently remove high-quality details, such as hair and facial features. Conversely, a weaker filter may leave more residual speckled noise unfiltered.

Overlaying Colorized Output with Black-and-White Inputs: As an alternative, we experimented with overlaying the colorized output with the original black-and-white inputs. This approach yielded superior results compared to the nlmeans filter, and it required less parameter tuning for filter strength. We opted to proceed with this approach, referred to as 'Diffusion Filtered'.

Despite our efforts to optimize noise reduction while preserving critical details, the final output quality still fell short of our improved approach, LatentColorization, which we will detail in the following section. Consequently, our final experiments did not incorporate the Palette-based approach [49].



(a)



(b)

FIGURE 4. **(a) The system architecture during training is depicted in the diagram, illustrating the key elements of the network and their interactions: Image Encoder: This component is responsible for encoding the input frames into embedding representations. It generates the ground truth embedding Z_{GT}, the embedding of the current black-and-white frame Z_{BW}, and the embedding of the previous color frame Z_P. Denoising Unet: This is a critical part of the architecture, responsible for denoising and refining the embeddings generated by the Image Encoder that have passed through the forward diffusion process. Conditioning Mechanism: The conditioning mechanism is integral to the network, providing contextual information and conditioning signals to guide the colorization process. It takes into account various embeddings, including Z_{BW}, Z_P, and Z_T, which represent the black and white input frame, the output of the model at the previous timestep, and the noisy frame to be denoised. Image Decoder: This component is responsible for decoding the predicted frames from their embedding representations. The architecture's design and interactions are essential for the model's training process, ensuring that it learns to generate accurate and temporally consistent colorizations over multiple timesteps. (b) During inference, the system architecture remains largely consistent with the training phase, with one significant difference: Gaussian Noise in Place of Ground Truth Frame: Instead of the ground truth frame, the system introduces Gaussian noise as input during the testing phase. This alteration simulates real-world scenarios where the model must colorize frames without the ground truth. The rest of the architecture, including the Image Encoder, Denoising Unet, Conditioning Mechanism, Image Decoder, and their interactions, remains unchanged. This design allows the model to assess its performance under conditions that more closely resemble practical, ground truth-free scenarios.**

2) LATENT DIFFUSION BASED SET UP

Inspired by Latent Diffusion [12], we devised LatentColorization.

LatentColorization comprises three core components: an autoencoder, a latent diffusion model, and a conditioning mechanism, as visually represented in Fig. 4a.

The latent diffusion model follows a two-step process, commencing with the forward diffusion phase (formulated in Eqn. 1). During this phase, Gaussian noise is systematically introduced to the data, incrementally transforming it until it becomes indistinguishable from Gaussian noise. During the second phase, the learned backward diffusion process is applied. This is where a neural network is trained to learn the original data distribution, and to draw samples from it by reconstructing the data from Gaussian noise. We represent formulations of this process with conditioning in Eqn.3 and without conditioning, as in Eqn.2.

The forward diffusion process, as defined by [9], can be represented by the following formula:

$$q(x_t|x_{t-1}) = N(x_t; \mu_t = \sqrt{1-\beta}x_{t-1}, \Sigma_t = \beta_t I) \quad (1)$$

In this formulation, the probability distribution $q(\cdot)$ of the image at each timestep x_t, given the previous timestep x_{t-1}, is characterized as a normal distribution N. This distribution is centred around a mean equal to the previous timestep x_{t-1}, with noise incorporated. The magnitude of this noise is determined by the noise scheduler β at time t and is further modulated by the identity matrix I. The noise scheduler β typically follows a linear pattern, as exemplified in [44], or a cosine pattern, as demonstrated in [61].

The backward diffusion process, in accordance with [10], can be defined as follows:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

In this definition, the probability distribution $p_\theta(\cdot)$ of the slightly denoised image x_{t-1}, given the noisier image x_t, is characterized as a normal distribution $N()$. This distribution has a mean denoted as μ and a variance represented by Σ, both of which are learned and parameterized by the neural network indicated by θ.

The diffusion process can be conditioned using the following equation:

$$p_\theta(x_{0:T}|y) = p_\theta(x_t)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y) \quad (3)$$

In this equation, the probability density function p_θ is akin to the unconditioned diffusion process, but conditioning is introduced at each timestep of the diffusion process, denoted as $p_\theta(x_{t-1}|x_t, y)$. In our specific scenario, the conditions encompass the previous frame, the grayscale frame, and the current frame during training, as illustrated in Fig.4a. During inference, the conditions consist of the previous frame, the grayscale frame, and noise, as indicated in Fig.4b.

For a visual representation of our network architecture during training and inference, as well as a breakdown of where each equation is utilized, please refer to Fig. 4a and Fig. 4b. Additionally, for a more in-depth explanation of these equations and their derivation, you can explore the references provided in [9], [10], and [62].

In the training process, the current frame ground truth, the current frame in black and white, and the previous frame are fed into the image encoder. These images are compressed into their respective embeddings, namely Z_{GT}, Z_{BW}, and Z_P. The chosen autoencoder for this purpose is a Vector Quantized Variational AutoEncoder (VQ-VAE), as detailed in [63].

During the forward diffusion process, the current frame's ground truth embedding Z_{GT} has noise applied to it based on the noise timestep, resulting in Z_T. Simultaneously, the ground truth black and white embedding Z_{BW} and the previous frame embedding Z_P are concatenated. The noised embedding Z_T is then denoised using the Unet and conditioned on Z_{BW} and Z_P.

During the backward diffusion process, the neural network learns to predict the noise that was added during the forward diffusion process at time step T. Denoising the noise embedding Z_T using the predicted noise results in Z_{T-1}. We use a simple mean square error loss between the predicted noise, vs the actual noise added to the embedding in order to train the network.

By employing the previous frame as conditioning, temporal consistency between frames is ensured throughout the video sequence, resulting in coherent colorization.

During inference, the same process as the training scheme is followed, with the exception that the model is fed pure Gaussian noise representing frame Z_T. The denoising is then repeated T times, after which the denoised embedding is passed through to the image decoder in order to produce the predicted frame. This process is depicted explicitly in Fig.4b.

The system can be used in two different ways. First, it can be employed in an entirely end-to-end manner, where no additional guidance from the user is needed. In this setup, the system serves as an image colorization tool for the first frame. Then, this initial colorized frame is used in an autoregressive fashion to guide the colorization of subsequent frames in the video clip. Second, the system can be used interactively, allowing the user to manually colorize the initial frame. This manual colorization becomes the condition for initiating the colorization of the following frames. This second approach provides control over the colorization process but requires the user to provide the initial colorization.

D. TEMPORAL CONSISTENCY

Temporal consistency was maintained through an autoregressive conditioning mechanism, where the current video frame was conditioned on the previous frame and the grayscale version of the current frame. This approach ensured that colorization remained consistent across the video frames. For a detailed illustration, refer to Fig. 4a and Eqn 4. This mechanism is similar to the approaches used in other studies such as [64] and [65], where models were conditioned with information from the previous frame to guarantee temporal consistency in the context of video generation. Essentially, maintaining consistent colors throughout a video sequence becomes more achievable when the model can ''remember'' the colors from the previous frame.

$$C_t = f(C_{t-1..n}, G_t) \forall t \in T \quad (4)$$

Following convention [66], [67], we use the CIE LAB color space. We describe a color image as a combination of the L (Luminance), A (Alpha) and B (Beta) channels. Each value of CIE LAB is a real number; therefore, we denote it as $\mathbb{R}$. We describe a grayscale image as only the L (Luminance) channel. In Eqn 4, we denote the color image as $C \subseteq \mathbb{R}^{L,a,b}$, the grayscale image as $G \subseteq \mathbb{R}^L$, and $f()$ represents the colorization function performed by the neural network. Here, n signifies the length of the conditioning window frame, T is the total length of the video, and t indicates a specific moment within the video sequence. This equation describes how the

colorization process is conditioned on both the previous frame and the grayscale version of the current frame, ensuring temporal consistency across the video frames.

Throughout the video sequence, we maintain temporal consistency by providing the colorizer with the previous frame as a reference. However, a challenge arises at the beginning of the sequence, denoted as t_0, where there is no previous frame available for conditioning. To address this, we introduce an initial colorized frame at t_0. This initial frame is advantageous because it introduces an element of user preference, which can be highly practical. It effectively reduces the video colorization task to that of coloring a single image, which then serves as the starting point for colorizing the entire video with a bias towards the initial frame.

This approach offers flexibility and aligns with human-centric AI concepts for video colorization. We refer to this approach as ''initial frame biasing''. Additionally, it provides a clear method for evaluating the system, as ground truth is available for the initial frame, making traditional reference-based metrics such as PSNR, SSIM, FID, and FVD effective for assessment. It also allows for a user study where one can compare performance against the ground truth.

E. HYPERPAREMETER AND TRAINING SET UP

The hyperparameters used in the experiment are detailed in Table 1. The experiment employed the ADAM optimizer [68], with most of the values being adopted from the specifications of Stable Diffusion [12]. Any additional hyperparameters were determined through a process of empirical testing.

An image size of 128 × 128 pixels required a 4x decrease in processing time as opposed to 256 × 256 pixels. Training at 256 × 256 pixels takes 165 minutes per epoch on an NVIDIA RTX 2080, whereas training at 128 × 128 pixels takes 38 minutes per epoch. Using 200 Diffusion steps for training and 50 for testing resulted in good performance. Input channels must be nine to account for the conditioning, three channels for color previous frame, three channels for the image and three channels for the black-and-white current frame. Having a batch size of 256 and a learning rate of $1.25e^{-7}$ resulted in convergence and reasonably fast training times.

TABLE 1. The hyperparameter setup provides the values used for both training and testing.

	Train	Test
Image Size	$128x128$	$128x128$
Total Frames	10000	700
Diffusion Steps	200	50
Noise Schedule	$Linear$	$Linear$
Linear Start	$1.5e-03$	$1.5e-03$
Linear End	0.0195	0.0195
Input Channels	9	9
Inner Channels	64	64
Channels Multiple	$1, 2, 3, 4$	$1, 2, 3, 4$
Res Blocks	2	2
Head Channels	32	32
Drop Out	0	0
Batch Size	256	8
Epochs	350	-
Learning Rate	$1.25e^{-7}$	-

IV. EVALUATION

The performance evaluation of the colorization process combines both qualitative and quantitative assessments to gauge its success. Following similar colorization studies [7] our work is compared on standard metrics. The key metrics used for this evaluation are as follows:

Peak Signal to Noise Ratio (PSNR): This metric measures the quality of colorized images by comparing them to the corresponding ground truth images. It quantifies the difference between the pixel values of the colorized and ground truth images. Higher PSNR values indicate better performance.

Structural Similarity Index (SSIM): SSIM evaluates the structural similarity between colorized images and ground truth images. It considers not only pixel values but also the structure and patterns in the images. Higher SSIM values indicate greater similarity to the ground truth.

Fréchet Inception Distance (FID): FID assesses the distance between the distribution of features extracted from colorized images and real images. Lower FID values indicate closer similarity to real images.

Fréchet Video Distance (FVD): FVD is a video-specific metric that measures the difference between generated and real videos by comparing the mean and covariance of their features. Lower FVD values represent better video colorization quality.

Naturalness Image Quality Evaluator (NIQE): NIQE is a referenceless metric that quantifies the naturalness of colorized images using statistical measures. Lower NIQE values indicate more natural-looking images.

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE): BRISQUE is another referenceless metric that evaluates the quality of colorized images. It learns the characteristics of natural images and quantifies the deviation from these characteristics. Lower BRISQUE values represent better image quality.

Mean Opinion Score (MOS): MOS is a weighted average of survey participants' perceived quality of an image or video. Higher MOS score represents a higher opinion of the subjective quality of the media.

A combination of these quantitative metrics and visual inspection, see Fig. 7, allows for a comprehensive assessment of the colorization process, enabling objective and subjective evaluation of its performance.

Evaluating colorization is a very subjective task, and therefore, as well as the metrics used, a survey was conducted to obtain a subjective measure of our performance. This survey was conducted in a similar manner to the survey conducted by Wu et al. [31].

TABLE 2. The quantitative comparisons provide a detailed evaluation of different colorization methods across various datasets. These methods include DeOldify, DeepRemaster, ColTran, GCP, VCGAN, Human Colorized, LatentColorization without Temporal Consistency and LatentColorization. The evaluation criteria encompass several metrics, including PSNR, SSIM, FID, FVD, NIQE, and BRISQUE. By comparing these metrics on individual datasets and a combined dataset (consisting of GRID, Lombard Grid, and Sherlock Holmes Movies), the study aims to assess and compare the performance of these colorization methods. This information allows for an evaluation of how LatentColorization compares to other state-of-the-art methods in various scenarios.

Dataset	Method	PSNR ↑	SSIM ↑	FID ↓	FVD ↓	NIQE ↓	BRISQUE ↓
Grid [58]	DeOldify [6]	28.07	0.79	52.67	520.75	44.04	**32.47**
	DeepRemaster [42]	27.7	0.77	108.68	927.91	51.19	41.44
	ColTran [18]	28.08	0.76	91.76	759.32	49.69	34.55
	GCP [31]	27.74	0.75	109.75	1555.53	48.44	33.76
	VCGAN [41]	27.86	0.83	67.79	951.28	44.24	37.16
	LatentColorization w/o TC	29.63	0.89	**20.92**	350.35	46.4	33.73
	LatentColorization	**30.88**	**0.9**	22.26	**241.94**	**41.46**	34.68
Lombard Grid [59]	DeOldify [6]	30.69	0.93	17.63	396.2	46.43	33.1
	DeepRemaster [42]	30.09	0.95	32.9	1382.56	52.36	35.97
	ColTran [18]	29.96	0.89	37.7	1583.94	51.25	**29.71**
	GCP [31]	29.86	0.91	85.09	432.31	48.73	33.65
	VCGAN [41]	30.2	**0.96**	72.17	2146.79	50.72	31.01
	LatentColorization w/o TC	30.35	0.92	18.41	490.89	**44.53**	33.84
	LatentColorization	**30.71**	0.93	**17.01**	**375.34**	45.79	34.61
Sherlock Holmes Movies	DeOldify [6]	-	-	-	-	**42.07**	41.15
	DeepRemaster [42]	-	-	-	-	62.36	42.98
	ColTran [18]	-	-	-	-	47.15	**37.52**
	GCP [31]	-	-	-	-	49.87	41.95
	VCGAN [41]	-	-	-	-	49.84	39.86
	Human colorized	-	-	-	-	48.43	39.78
	LatentColorization w/o TC	-	-	-	-	47.13	38.49
	LatentColorization	-	-	-	-	46.24	41.11
Overall	DeOldify [6]	29.19	0.86	40.47	520.85	45.22	35.14
	DeepRemaster [42]	28.90	0.86	70.79	1155.24	55.60	40.13
	ColTran [18]	29.02	0.83	64.73	1171.63	49.36	**33.93**
	GCP [31]	29.80	0.83	97.42	993.92	49.01	36.45
	VCGAN [41]	29.03	0.9	69.98	1549.04	48.27	36.01
	Human colorized	-	-	-	-	48.43	39.78
	LatentColorization w/o TC	29.99	0.91	19.67	420.62	46.02	35.35
	LatentColorization	**30.80**	**0.92**	**19.64**	**308.64**	**44.50**	36.80

A. QUALITATIVE ANALYSIS

The qualitative results in Fig. 7 visually compare the colorization performance of different methods, including DeOldify [6], ColTran [18], DeepRemaster [42], GCP [31], VCGAN [41], LatentColorization without temporal consistency enabled, LatentColorization, and the ground truth. These comparisons are based on image sequences from the GRID [58] and Lombard Grid [59] datasets. Additional qualitative results can also be seen in our appendices. This visual assessment allows for a direct comparison of how well LatentColorization performs in relation to other state-of-the-art methods. Based on the qualitative analysis of the results in Fig. 7, the following conclusions can be drawn:

DeOldify [6] produces consistent colorizations, but they tend to appear dull and have a halo effect around the subject. ColTran generates colorful images, but it suffers from inconsistencies throughout the sequence. DeepRemaster [42] provides produces dull, conservative colorizations. GCP [31] produces colorful, consistent colorizations, but they are not faithful to the ground truth. VCGAN [41] seems to mostly apply a blueish filter to the frames. LatentColorization w/o TC produces colorization similar to the ground truth. It is difficult to visually distinguish between LatentColorization w/o TC, LatentColorization and the ground truth itself. LatentColorization impressively colorizes the sequence, maintaining faithfulness to the original, vibrancy in color, and overall consistency. Overall, LatentColorization appears to outperform the other methods in terms of fidelity to the original, colorfulness, and consistency.

B. QUANTITATIVE ANALYSIS

Quantitative evaluation is an essential aspect of assessing the quality and performance of colorization methods. It helps provide an objective measure of how well these methods perform. By evaluating colorizations both frame by frame and as a video sequence, you can gain insights into the strengths and weaknesses of each approach and determine how well they maintain consistency and quality throughout the sequence. This quantitative assessment complements the qualitative analysis and provides a more comprehensive understanding of the colorization results.

Table 2 provides a quantitative evaluation of the colorization methods, considering various image metrics. It is a useful way to compare the performance of DeOldify [6], DeepRemaster [42], ColTran [18], GCP [31], VCGAN [41], LatentColorization without temporal consistency mechanism, LatentColorization, and human colorization. By assessing metrics such as PSNR, SSIM, FID, FVD, NIQE, and BRISQUE, you can analyze the quality, similarity, and naturalness of the colorized images. This comparison enables

a more data-driven and objective assessment of how well each method performs.

The results presented in Table 2 indicate that LatentColorization performs well across all of the referenced and non-referenced metrics, surpassing the state-of-the-art DeOldify [6] by an average of ≅18% in terms of FVD. This performance showcases the effectiveness of LatentColorization in achieving high-quality and consistent video colorization results.

Comparing LatentColorization against human-level colorization is an important evaluation. Using non-reference image quality assessment metrics like NIQE and BRISQUE to assess the relative performance when no ground truth is available is a valuable approach. These metrics provide insights into how closely the colorization generated by LatentColorization aligns with human-expert colorization in terms of image quality.

The results in Table 2 show that LatentColorization outperforms human colorization according to NIQE and BRISQUE, which indicates that the colorizations produced by LatentColorization are of high quality when assessed using these non-reference metrics.

The other methods also perform well on BRISQUE and NIQE scores relative to the Human Colorized version of the video. Colorization is a subjective matter, and therefore, these metrics must be paired with a user survey to evaluate the systems' performances.

C. SURVEY

A survey was conducted to get a more subjective view of the performance of LatentColorization. This study aimed to evaluate the difference in performance between our proposed approach, LatentColorization, and its closest competitor in our experiments, DeOldify [6]. Thirty-two participants were shown three sets of three videos and were asked a question on each set. Each dataset had an associated video set. The survey questions can be seen in our appendices.

For the Grid [58] dataset, the participants were shown three versions of the same video taken from the dataset side-by-side. One video version had been colorized by LatentColorization, the other by DeOldify [6], and the third was the ground truth. The ground truth video was labelled as such, whereas the LatentColorization and DeOldify [6] versions of the video were anonymous. To distinguish the LatentColorization version of the video from the DeOldify version [6] they were labelled with 1 and 2. After the participants had watched the videos, they were asked which video they thought was closer to the ground truth. The purpose of this question (Question 1) was to differentiate in a head-to-head competition in which the colorization system was able to produce outputs which were similar to the ground truth colors of the video.

For the Lombard Grid [59] dataset, the participants were shown three versions of an example video taken from the dataset shown side-by-side. Again, one version was colorized by LatentCololorization, the other by DeOldfiy [6], and the third was the ground truth. In contrast to the previous question, the ground truth video was anonymous this time, and the three videos were titled 1, 2 and 3. After the participants watched the video, they were asked to rank the three videos in terms of which one looked the most realistic. Therefore, this question (Question 2) acted as a visual turning test where humans were tested to see if they could tell the difference between a colorization and a ground truth video. The idea behind this is that the better the performance of the colorization system, the more difficult it should be to distinguish between the colorization system and the ground truth.

For the Sherlock Holmes dataset, the participants were shown three versions of an example video from the dataset side-by-side. One version had been colorized by LatentColorization, the other by DeOldify [6], and the third was the human-colorized version. This time, the human-colorized version of the video was labelled, and the LatentColorization and DeOldify [6] versions were left anonymous. After the participants had watched the clips, they were asked which of the automatically colorized versions of the clip was closer to the human-colorized version. The purpose of this question (Question 3) then was to determine the relative performance of LatentColorization, DeOldify [6] with respect to human expert colorizations.

We then collated the survey results and analysed them. The results can be seen visually in Fig. 5. The X-axis represents the Mean Opinion Score (MOS) for each question's methods. The Y axis indicates the relevant question. The color-coded bars represent each of the methods. The mean opinion score was calculated for each method for each question. For Question 1 and Question 2, the mean opinion score is simply the tally of each of the votes as it compares two methods. For Question 3, the mean opinion score is the sum of the ratings for each method divided by the number of methods.

Interpreting the graph, we can see that overall LatentColorization was preferred to DeOldify [6]. For Question 1, DeOldify [6] received seven votes, and LatentColorization received 25 votes, indicating a preference for LatentColorization on this question. For Question 2, the ground truth received the highest MOS score of 28.00, followed by LatentColorization at 20.00 and DeOldiy [6] at 13.67. Summarising this result, the ground truth was preferred most of the time, followed by LatentColorization and finally DeOldify [6]. For Question 3, LatentColorization was chosen 26 times out of 31, indicating a strong preference for LatentColorization.

D. ABLATION STUDY

An ablation study was undertaken to evaluate the impact of the temporal consistency mechanism on the LatentColorization system. The results for both LatentColorization and LatentColorization without temporal consistency mechanism are recorded in Table 2. LatentColorization refers to the version of LatentColorization with the temporal

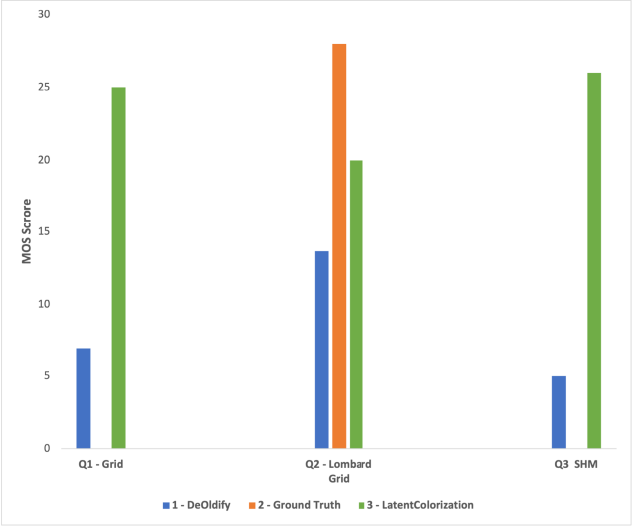


FIGURE 5. The graph of the results of the survey. Each group represents a particular question. The X-axis represents the Mean Opinion Score (MOS) for each question's methods. The Y axis indicates the relevant question. The color-coded bars represent each of the methods.

consistency mechanism enabled, and LatentColorization w/o TC refers to the version of LatentColorization where the temporal consistency metric has been disabled. The results of LatentColorization and LatentColorization without temporal consistency appear similar. The main difference is that the FVD values for LatentColorization are roughly 10% lower than LatentColorization without temporal consistency's FVD values. As a result of this observation, it can be deduced that the temporal consistency mechanism is indeed improving the video quality of the output and, therefore, ensuring temporal consistency.

E. FAILURE CASES

There were also instances where the system failed to colorize faithfully to the ground truth. This particularly occurred for out-of-distribution data where the videos were from a different domain than speaker videos, see Fig. 6. LatentColorization fails to apply realistic colors to the bedroom scene. It initially manages to separate the walls from the bed as it colorizes the walls blue and the bed orange, see Frame 1. As time passes, see Frame 2 and Frame 3; LatentColorization tends towards a dull grey color. This indicates that LatentColorization is sensitive to the domain that the video is from, and when it does not recognize the contents of a video, it resorts to drab, dull colors.

V. DISCUSSION

In this section, we discuss our model's results compared to other approaches from the field.

ColTran [18] Vs LatentColorization: The comparison between LatentColorization and non-autoregressive models like ColTran [18] provides insights into the importance of the autoregressive nature of the system in the context

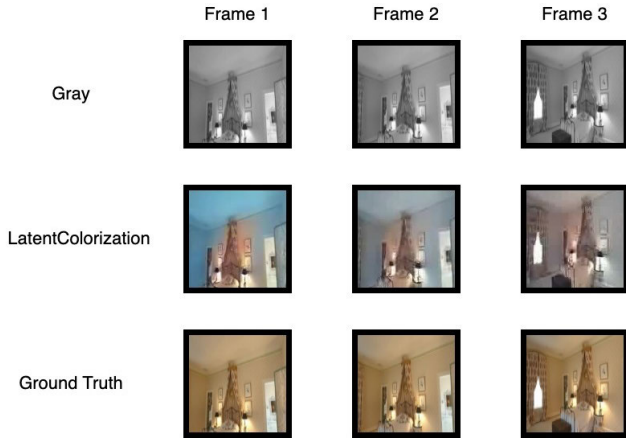


FIGURE 6. The comparison of three frames from the system taken from out-of-distribution data. The top row is the black-and-white version of the video, the middle frame is the output of LatentColorization, and the bottom row is the ground truth. It can be seen that LatentColorization has failed to colorize faithfully to the ground truth.

of video colorization. Fig. 7 demonstrates the difference in consistency between the two approaches. The frames colorized by LatentColorization appear more consistent throughout the video sequence, while those generated by ColTran [18] exhibit more variation. This suggests that the autoregressive nature of LatentColorization, where each frame is conditioned on the previous ones, plays a role in maintaining temporal consistency and ensuring that the colorization is coherent across the entire video. In contrast, approaches like ColTran [18] which do not have a temporal consistency mechanism may struggle to achieve the same level of consistency in colorized sequences.

DeOldify [6] Vs LatentColorization: The qualitative assessment of the colorizations in Fig. 7 highlights the differences in colorfulness among LatentColorization, and DeOldify [6]. LatentColorization produces colorful results. In contrast, DeOldify [6] appears grey, suggesting that it may suffer from a lack of color diversity. This observation is consistent with the idea that GANs, which DeOldify [6] is based on, can be susceptible to mode collapse, where they produce limited and less diverse color variations. This observation also correlated with the survey results where LatentColorization was preferred to DeOldify [6] 80% of the time.

DeepRemaster [42] Vs LatentColorization: DeepRemaster [42] has struggled with the colorization of this material and has resorted to very bland, dull colors, unlike LatentColorization.

GCP [31] Vs LatentColorization: it can be seen that LatentColorization is closer to the ground truth than GCP [31]. GCP has produced colorful output, but it is different in color from the ground truth. It has not succumbed to the mode collapse of its GAN-based architecture, especially on the Lombard Grid [59] dataset. This could potentially be a result of its retrieval mechanism.

VCGAN [41] Vs LatentColorization: it can be seen that LatentColorization is closer to the ground truth than

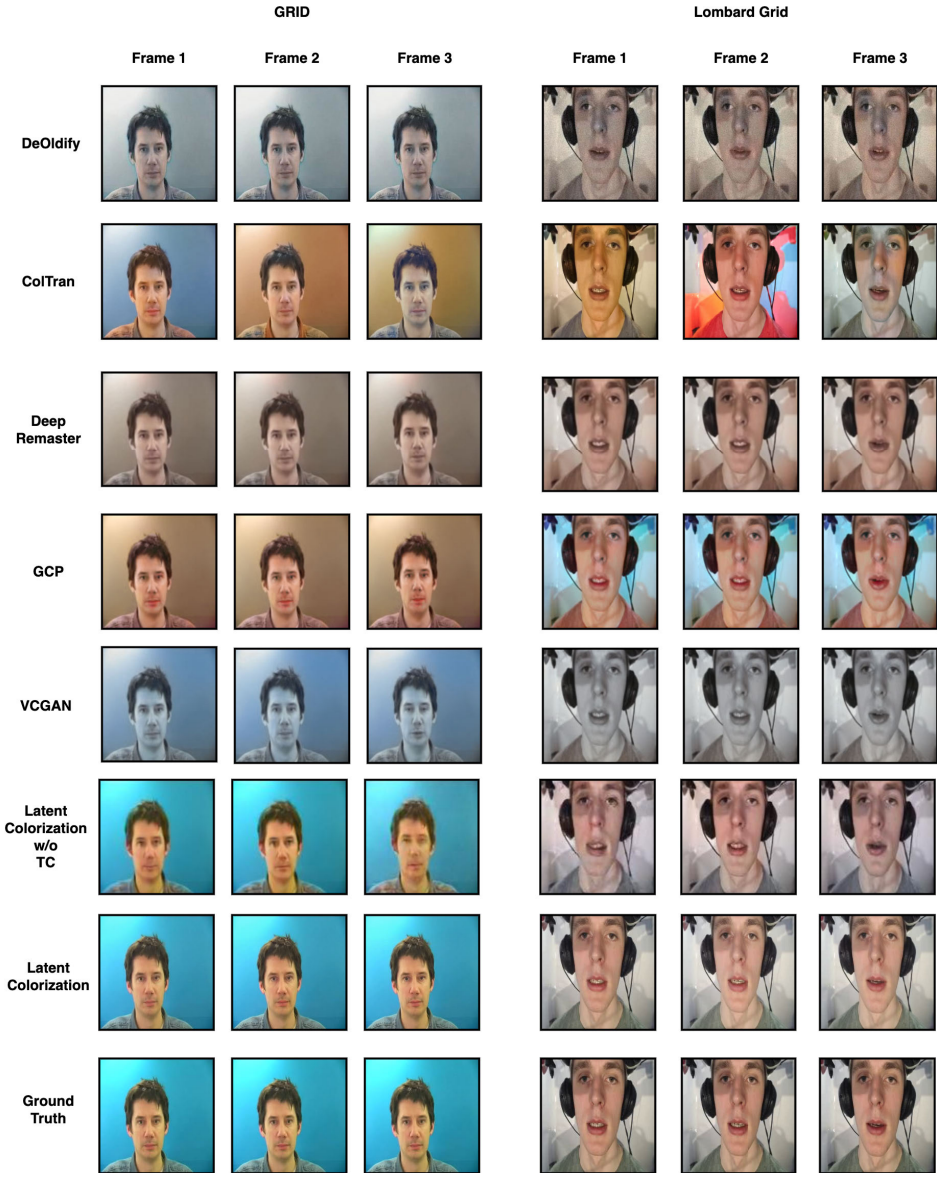


FIGURE 7. The qualitative comparison of colorization results from various systems, including DeOldify [6], ColTran [18], DeepRemaster [42], GCP [31], VCGAN [41], LatentColorization without the temporal consistency mechanism enabled (LatentColorization w/o TC), LatentColorization and the ground truth, for both the GRID [58] dataset (left) and the Lombard Grid [59] dataset (right) is shown. In the GRID [58] dataset, DeOldify's [6] colorization, depicted in the first row, exhibits desaturated colors and a halo effect around the subject. ColTran [18], in the second row, produces more colorful results but lacks consistency throughout the sequence. DeepRemaster [42] produces dull, conservative colorizations. GCP [31] produces colorful, consistent colorizations, but they are not faithful to the ground truth. VCGAN [41] produces drab, monotone colorizations. LatentColorization w/o TC produces colorization similar to the ground truth. It is difficult to visually distinguish between LatentColorization w/o TC, LatentColorization and the ground truth itself. The ground truth, represents the original color frames. Similar observations can be made for the Lombard Grid [59] dataset. These visual comparisons demonstrate that LatentColorization consistently delivers colorization results that closely match the original colors, making it a promising technique for automatic video colorization tasks.

VCGAN [41]. VCGAN has produced a blue filter type effect on the frames.

LatentColorization Vs LatentColorization without temporal consistency: has been investigated in the ablation study. Essentially, it is difficult to visually differentiate between the two, and the main difference can be seen quantitatively in their relative FVD scores.

The quantitative evaluation, as shown in Table 2, indicates that LatentColorization achieved scores on the NIQE and BRISQUE metrics that are close to human-level colorization.

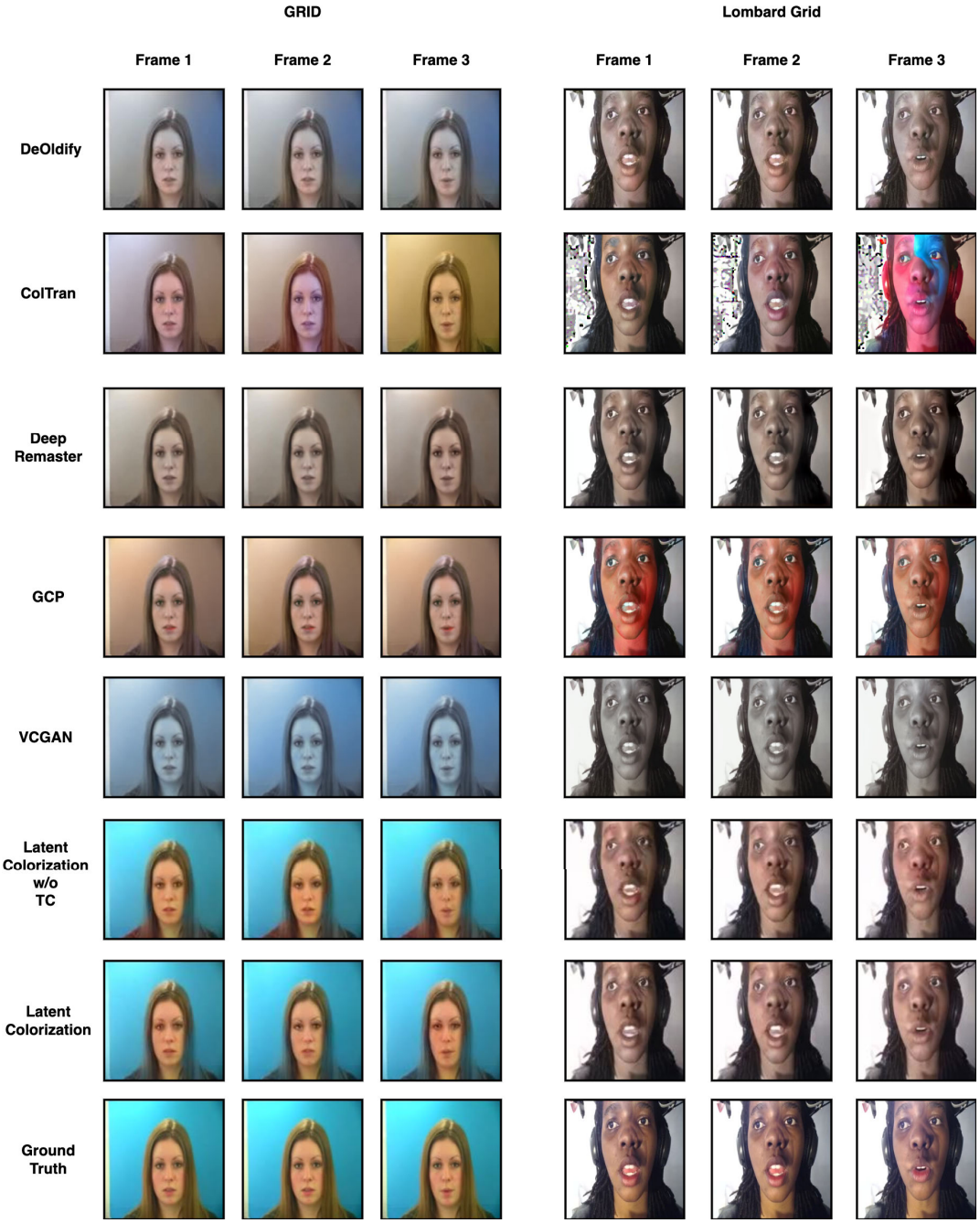


FIGURE 8. The qualitative comparison of colorization results from various systems, including DeOldify [6], ColTran [18], DeepRemaster [42], GCP [31], VCGAN [41], LatentColorization without the temporal consistency mechanism enabled (LatentColorization w/o TC), LatentColorization and the ground truth, for both the GRID [58] dataset (left) and the Lombard Grid [59] dataset (right) reveals differences in their performance.

In summary, these results suggest that LatentColorization, in this experiment, is comparable to human-level colorization in terms of the assessed quality metrics. This highlights the effectiveness of the LatentColorization method in generating high-quality colorized videos. This evaluation also correlates with our survey, where LatentColorization received a higher preference from the subjects than DeOldify [6]. The survey also shows a tendency of the users to prefer the ground truth videos over both LatentColorization and DeOldify [6].

VI. CONCLUSION

In conclusion, our work demonstrates the effectiveness of diffusion-based models, particularly the LatentColorization method, in achieving results comparable to the state of the art across multiple datasets. Notably, the system performs comparably to human-level colorization on the 'Sherlock Holmes Movie' dataset, indicating its practical significance and the potential for application-specific video colorization. The use of a latent diffusion model and the incorporation

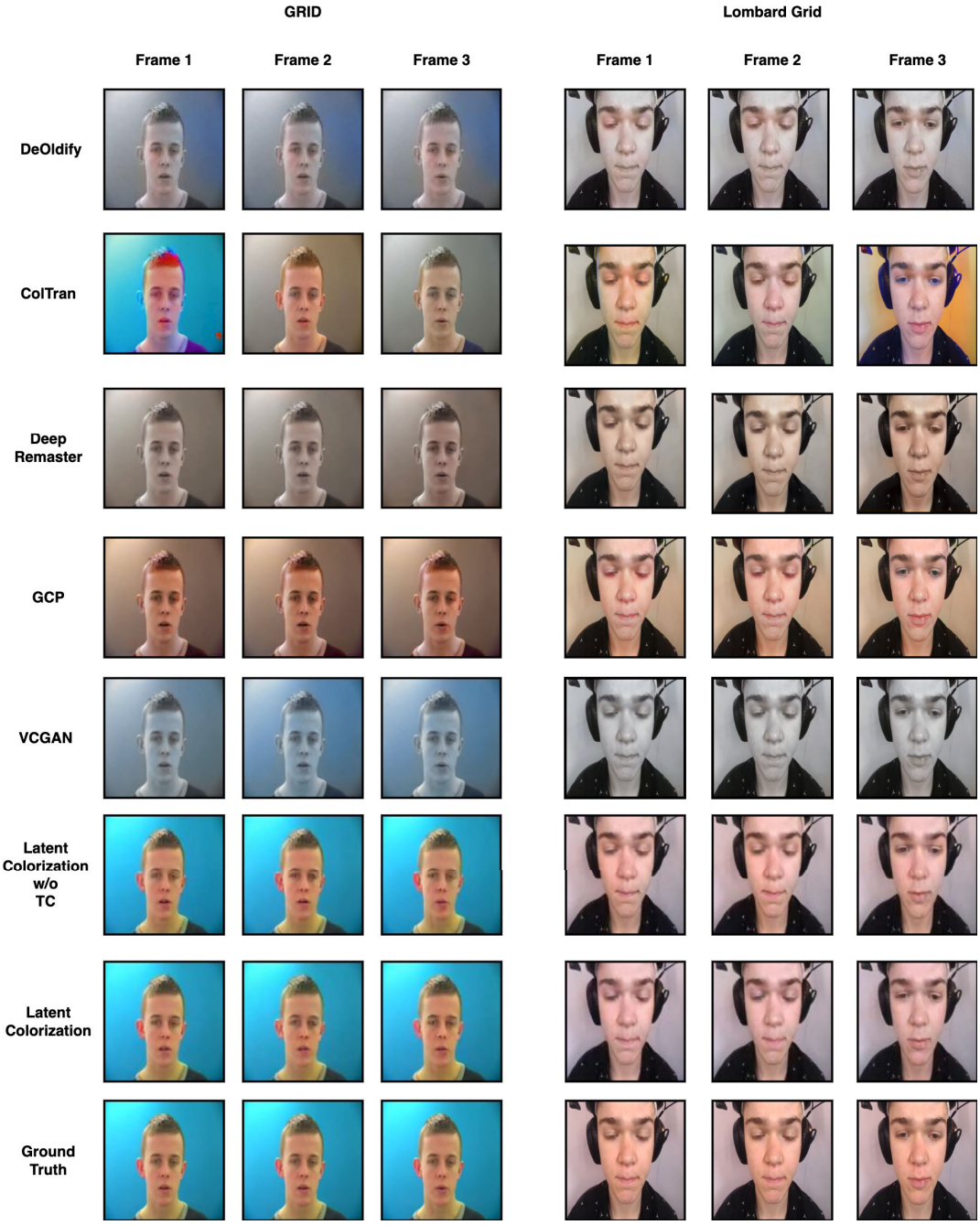


FIGURE 9. The qualitative comparison of colorization results from various systems, including DeOldify [6], ColTran [18], DeepRemaster [42], GCP [31], VCGAN [41], LatentColorization without the temporal consistency mechanism enabled (LatentColorization w/o TC), LatentColorization and the ground truth, for both the GRID [58] dataset (left) and the Lombard Grid [59] dataset (right) reveals differences in their performance.

of a temporally consistent colorization approach contribute to realistic and convincing colorization results, making the process more accessible and reducing the reliance on traditional human-driven colorization methods. This research provides insights into the potential of diffusion models for video colorization and opens up opportunities for further developments in this field.

VII. FUTURE WORK

Expanding on our research, adapting the system to work with various video styles, types, and content would be a promising direction for future work. This would enable a broader assessment of the approach's applicability in general video colorization. Decreasing the inference and training time of this system would also be beneficial. Currently, the

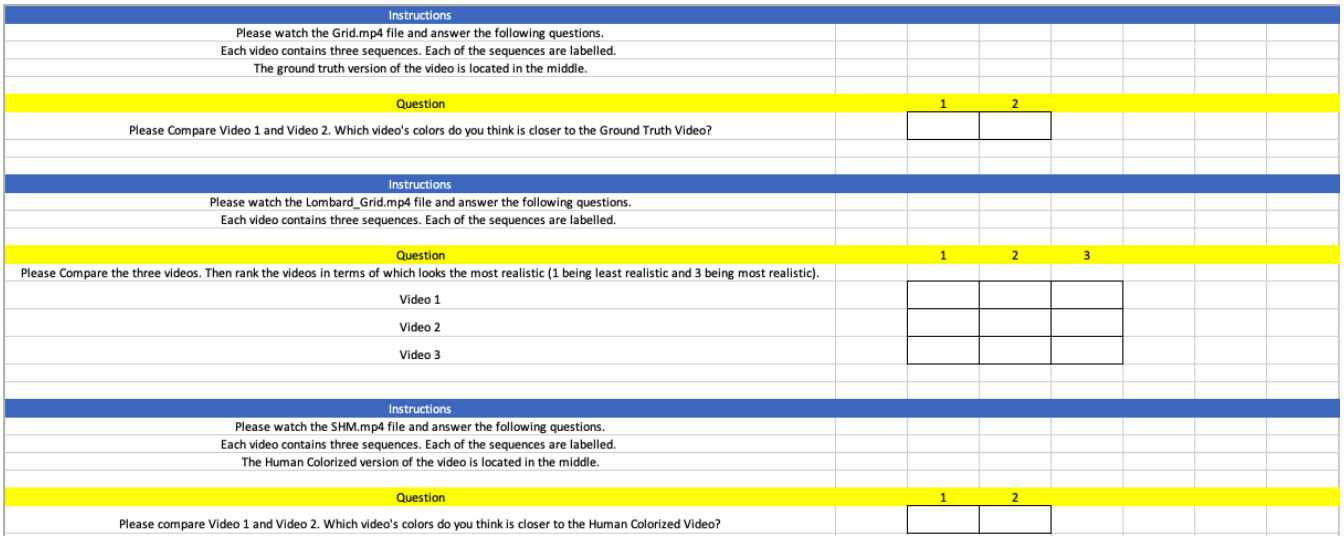


FIGURE 10. The survey questions asked. Question 1 compares LatentColorization to DeOldify [6] on the Grid [58] dataset. Question 2 compares LatentColorization, DeOldify [6] and the ground truth on the Lombard Grid [59] dataset. Question 3 compares LatentColorization, and DeOldify [6] on the Sherlock Holmes Movie dataset.

system must sample each frame multiple times, which limits real-time capabilities. There are also ethical concerns that further research must be conducted into. Two of the main ethical concerns are misuse and bias. These systems could be potentially used maliciously to distort history. They could also inherit bias from their datasets, which would result in systems that are not fair. One practical way of alleviating some of the ethical concerns regarding this work could be by developing a model card highlighting the potential biases present in the system. These endeavors will further enhance the practicality and versatility of our research in automatic video colorization.

APPENDIX A
ADDITIONAL EXPERIMENTS

See Figures 8 and 9.

APPENDIX B
SURVEY

See Figure 10.

REFERENCES

[1] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," 2020, *arXiv:2005.10825*.

[2] F. Pierre and J.-F. Aujol, *Recent Approaches for Image Colorization*. Cham, Switzerland: Springer, 2021, pp. 1–38.

[3] S. Liu and X. Zhang, "Automatic grayscale image colorization using histogram regression," *Pattern Recognit. Lett.*, vol. 33, no. 13, pp. 1673–1681, Oct. 2012.

[4] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, Aug. 2004.

[5] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8044–8053.

[6] J. Antic. (2019). *Deoldify*. [Online]. Available: https://github.com/jantic/DeOldify

[7] Z. Wan, B. Zhang, D. Chen, and J. Liao, "Bringing old films back to life," 2022, *arXiv:2203.17276*.

[8] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," 2018, *arXiv:1806.09594*.

[9] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[11] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," 2021, *arXiv:2105.05233*.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[13] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2205, *arXiv:2205.11487*.

[14] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[16] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," 2019, *arXiv:1906.09909*.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[18] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," 2021, *arXiv:2102.04432*.

[19] E. Casey, V. Pérez, Z. Li, H. Teitelman, N. Boyajian, T. Pulver, M. Manh, and W. Grisaitis, "The animation transformer: Visual correspondence via segment matching," 2021, *arXiv:2109.02614*.

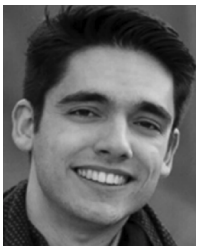
[20] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH Conf.*, 2022, pp. 1–10.

[21] B. Natarajan, E. Rajalakshmi, R. Elakkiya, K. Kotecha, A. Abraham, L. A. Gabralla, and V. Subramaniyaswamy, "Development of an end-to-end deep learning framework for sign language recognition, translation, and video generation," *IEEE Access*, vol. 10, pp. 104358–104374, 2022.

[22] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9396–9416, Dec. 2022.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2016, *arxiv:1611.07004*.
[24] C. Zou, H. Mo, C. Gao, R. Du, and H. Fu, "Language-based colorization of scene sketches," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–16, Nov. 2019.
[25] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–14, Dec. 2018.
[26] X. Kuang, X. Sui, C. Liu, Y. Liu, Q. Chen, and G. Gu, "Thermal infrared colorization via conditional generative adversarial network," 2018, *arXiv:1810.05399*.
[27] W. Chen and J. Hays, "SketchyGAN: Towards diverse and realistic sketch to image synthesis," 2018, *arXiv:1801.02753*.
[28] P. Hensman and K. Aizawa, "CGAN-based Manga colorization using a single training image," 2017, *arXiv:1706.06918*.
[29] C. W. Seo and Y. Seo, "Seg2pix: Few shot training line art colorization with segmented image data," *Appl. Sci.*, vol. 11, no. 4, p. 1464, Feb. 2021.
[30] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," 2017, *arXiv:1702.06674*.
[31] Y. Wu, X. Wang, Y. Li, H. Zhang, X. Zhao, and Y. Shan, "Towards Vivid and diverse image colorization with generative color prior," 2022, *arXiv:2108.08826*.
[32] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*.
[33] F. Mameli, M. Bertini, L. Galteri, and A. Del Bimbo, "A NoGAN approach for image and video restoration and compression artifact removal," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, 2021, pp. 9326–9332, doi: 10.1109/ICPR48806.2021.9413095.
[34] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*.
[35] P. Kouzouglidis, G. Sfikas, and C. Nikou, "Automatic video colorization using 3D conditional generative adversarial networks," 2019, *arXiv:1905.03023*.
[36] R. Endo, Y. Kawai, and T. Mchizuki, "A practical monochrome video colorization framework for broadcast program production," *IEEE Trans. Broadcast.*, vol. 67, no. 1, pp. 225–237, Mar. 2021.
[37] N. Akimoto, A. Hayakawa, A. Shin, and T. Narihira, "Reference-based video colorization with spatiotemporal correspondence," 2020, *arXiv:2011.12528*.
[38] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "VEEGAN: Reducing mode collapse in GANs using implicit variational learning," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017.
[39] T. Che, Y. Li, A. Paul Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," 2016, *arXiv:1612.02136*.
[40] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, *arxiv:1606.03498*.
[41] Y. Zhao, L.-M. Po, W. Y. Yu, Y. Abbas Ur Rehman, M. Liu, Y. Zhang, and W. Ou, "VCGAN: Video colorization with hybrid generative adversarial network," *IEEE Trans. Multimedia*, vol. 25, pp. 3017–3032, 2023.
[42] S. Iizuka and E. Simo-Serra, "DeepRemaster: Temporal source-reference attention networks for comprehensive video enhancement," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–13, Dec. 2019.
[43] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," 2021, *arXiv:2103.14031*.
[44] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020, *arXiv:2006.11239*.
[45] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, and T. Salimans, "Imagen video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.
[46] R. Villegas, "Phenaki: Variable length video generation from open domain textual descriptions," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–14.
[47] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," 2020, *arXiv:2009.09761*.
[48] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," 2022, *arXiv:2207.09983*.
[49] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," 2021, *arXiv:2111.05826*.
[50] H. Liu, M. Xie, J. Xing, C. Li, and T.-T. Wong, "Video colorization with pre-trained text-to-image diffusion models," 2023, *arXiv:2306.01732*.
[51] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," 2021, *arXiv:2106.05931*.
[52] T. Dockhorn, A. Vahdat, and K. Kreis, "Score-based generative modeling with critically-damped Langevin diffusion," 2021, *arXiv:2112.07068*.
[53] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," 2021, *arXiv:2112.07804*.
[54] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-a-video: Text-to-video generation without text-video data," 2022, *arXiv:2209.14792*.
[55] M. Otani and H. Hioki, "Video colorization based on optical flow and edge-oriented color propagation," in *Proc. SPIE*, vol. 9020, C. A. Bouman and K. D. Sauer, Eds., 2014, Art. no. 902002.
[56] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng, "Intrinsic colorization," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 1–9, Dec. 2008.
[57] R. Ward and J. Breslin, "Towards temporal stability in automatic video colourisation," in *Proc. 24th Irish Mach. Vis. Image Process. Conf.*, Aug. 2022, pp. 1–8.
[58] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "The grid audio-visual speech corpus," Zenodo, Jan. 2006, doi: 10.5281/zenodo.3625687.
[59] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *J. Acoust. Soc. Amer.*, vol. 143, no. 6, pp. EL523–EL529, Jun. 2018.
[60] A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *Image Process. Line*, vol. 1, pp. 208–212, Sep. 2011.
[61] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," 2021, *arXiv:2102.09672*.
[62] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–9.
[63] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," 2017, *arXiv:1711.00937*.
[64] D. Bigioi, S. Basak, M. Stypułkowski, M. Zieba, H. Jordan, R. McDonnell, and P. Corcoran, "Speech driven video editing via an audio-conditioned diffusion model," 2023, *arXiv:2301.04474*.
[65] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, and M. Pantic, "Diffused heads: Diffusion models beat GANs on talking-face generation," 2023, *arXiv:2301.03396*.
[66] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 110:1–110:11, 2016.
[67] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Alamitos, CA, USA, Jun. 2020, pp. 7965–7974.
[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

RORY WARD (Member, IEEE) received the bachelor's degree in electronic and computer engineering from the National University of Ireland Galway, in 2021. He is currently pursuing the Ph.D. degree with the University of Galway. His Ph.D. study is sponsored by the SFI Centre for Research Training in Artificial Intelligence. His research interest includes automatic visual media colorization.



DAN BIGIOI (Graduate Student Member, IEEE) received the bachelor's degree in electronic and computer engineering from the University of Galway, in 2020, where he is currently pursuing the Ph.D. degree. Upon graduating, he was a Research Assistant with NUIG, studying the text-to-speech and speaker recognition methods under the Data-Center Audio/Visual Intelligence on-Device (DAVID) Project. His Ph.D. study is sponsored by D-REAL and the SFI Centre for Research Training in Digitally Enhanced Reality. His research interests include novel deep learning-based techniques for automatic speech dubbing and discovering new ways to process multimodal audio/visual data.

SHUBHAJIT BASAK (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from West Bengal University of Technology, India, in 2011, and the M.Sc. degree in computer science from the University of Galway, Ireland, in 2018, where he is currently pursuing the Ph.D. degree in computer science. He has more than six years of industrial experience as a Software Development Professional. He is also with FotoNation/Xperi. His research interest includes deep learning tasks related to computer vision.

JOHN G. BRESLIN (Senior Member, IEEE) is currently a Personal Professor (a Personal Chair) of electronic engineering with the College of Science and Engineering, University of Galway, Ireland, where he is the Director of the TechInnovate/AgInnovate Programs. He is associated with two SFI Research Centers, where he is a Co-Principal Investigator with Insight (Data Analytics) and a Funded Investigator with VistaMilk (AgTech). He has taught electronic engineering, computer science, innovation, and entrepreneurship topics during the past two decades. He has written more than 250 peer-reviewed academic publications with a H-index of 50 and 10 000 citations. He has coauthored the books *Old Ireland in Colour Volume 1, 2 and 3*, *The Social Semantic Web*, and *Social Semantic Web Mining*. He has co-created the SIOC framework, which is implemented in hundreds of applications (by Yahoo, Boeing, and Vodafone) on at least 65 000 websites with 35 million data instances. He received the various best paper awards from conferences and journals.

PETER CORCORAN (Fellow, IEEE) was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He is currently holding the Personal Chair of electronic engineering of the College of Science and Engineering, University of Galway. He has more than 600 cited technical publications and patents, more than 120 peer-reviewed journal articles, 160 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is also a member of the IEEE Consumer Technology Society, for more than 25 years, and the Founding Editor of *IEEE Consumer Electronics Magazine*.