

FRCol: Face Recognition Based Speaker Video Colorization

Rory Ward^a and John Breslin^b

Data Science Institute, School of Engineering, University of Galway, Galway, Ireland

Keywords: Colorization, Face Recognition, Generative AI, Computer Vision.

Abstract: Automatic video colorization has recently gained attention for its ability to adapt old movies for today’s modern entertainment industry. However, there is a significant challenge: limiting unnatural color hallucination. Generative artificial intelligence often generates erroneous results, which in colorization manifests as unnatural colorizations. In this work, we propose to ground our automatic video colorization system in relevant exemplars by leveraging a face database, which we retrieve from using facial recognition technology. This retrieved exemplar guides the colorization of the latent-diffusion-based speaker video colorizer. We dub our system FRCol. We focus on speakers as humans have evolved to pay particular attention to certain aspects of colorization, with human faces being one of them. We improve the previous state-of-the-art (SOTA) DeOldify by an average of 13% on the standard metrics of PSNR, SSIM, FID, and FVD on the Grid and Lombard Grid datasets. Our user study also consolidates these results where FRCol was preferred to contemporary colorizers 81% of the time.

1 INTRODUCTION

Colorization has a broad spectrum of applications, whether reimagining nostalgic Hollywood classics like *Casablanca* (Curtiz, 1942) to *Psycho* (Hitchcock, 1960), or simply feeling closer to one’s ancestors with docuseries such as *World War II in Colour* (Martin, 2009). It has a vast potential to bring nostalgia and joy to many people. If done poorly, it also has the power to offend an audience and even distort history. Therefore, colorization must be handled with care. However, this process can be tedious and expensive, requiring massive attention to minute detail (Pierre and Aujol, 2021).


To make colorization more accessible, automatic colorization has been developed for both images and videos. Automatic image colorization requires spatial consistency throughout the frame, but there is no need for temporal consistency, unlike automatic video colorization. Many tools and techniques have been created for video and image applications, with a rich literature associated with both (Chen et al., 2022). Some notable examples include histogram matching (Liu and Zhang, 2012), Convolutional Neural Network (CNN) (Zhang et al., 2016), Generative Adversarial Network (GAN) (Kouzouglidis et al., 2019),

Transformer (Weng et al., 2022) and Diffusion-based (Saharia et al., 2022) systems.

One particular category of videos that we will choose to pay particular attention to in this work is speaker videos. We make this decision because human faces are important in everyday life. Humans have evolved to pay special attention to faces over millennia as they can transmit non-verbal information from person to person (Erickson and Schulkin, 2003). Therefore, if a colorization system is poor at colorizing faces, it will struggle to convince any human evaluator of its authenticity.

With this in mind, there exists a significant challenge with automatic video colorization: limiting unnatural color hallucination (Zhao et al., 2024). As colorization is a poorly-constrained problem with multiple plausible colorizations for any given colorization, how do we guide the system to the “correct” output? We propose to incorporate exemplar frames into the colorization process. We suggest a facial recognition algorithm to retrieve the most relevant exemplar from a pre-populated exemplar frame database. We can then use this pertinent exemplar to guide the colorization process.

In addition to the massive increase in the capabilities of automatic colorization due to deep learning and artificial intelligence, there has also been a huge increase in the capabilities of the adjacent field

^a  <https://orcid.org/0009-0003-7634-9946>


^b  <https://orcid.org/0000-0001-5790-050X>



Figure 1: **Colorization of The Adventures of Sherlock Holmes (1984)**. The grayscale version is shown on the top, and the FRCol colorization is shown on the bottom. The output has been upscaled with Topaz Labs.

of face recognition. Face recognition is the process of matching a person's identity to a reference image stored in a database (Wang and Deng, 2021),(S, 2023). It has many applications, including fraud detection (Choi and Kim, 2010), cyber security (Dodson et al., 2021), airport and border control (Sanchez del Rio et al., 2016), banking (Jain et al., 2021) and healthcare (Sardar et al., 2023). While this technology has huge potential to benefit people's lives positively, some associated challenges and concerns exist. Some of the main issues with this technology have to do with privacy and representation (Raji et al., 2020). There may be issues around using personal information, such as images of faces without consent, and the systems being biased through the underrepresentation of groups within the training sets. In recognition of the advances in facial recognition technology, we propose to leverage it in our system to reduce the amount of unnatural colorization that plagues automatic video colorization. Summarizing the contributions of our work:

- We propose a novel automatic speaker video colorization system augmented by exemplars retrieved using facial recognition technology called FRCol.
- FRCol achieves state-of-the-art performance on the automatic speaker video colorization task across various datasets and metrics. Specifically, we achieved a 13% average increase across the Grid and Lombard Grid datasets on the PSNR, SSIM, FID, and FVD scores compared to the previous SOTA DeOldify. Our user study also con-

solidates these results where FRCol was preferred to contemporary colorizers 81% of the time.

- We developed an intuitive user application to interact with FRCol easily. It takes a grayscale video and an optional path to a custom faces database as input. It outputs the resultant colorization played parallel to the input grayscale video.

2 RELATED WORK

2.1 Automatic Image Colorization

Automatic image colorization is a well-established task with an extensive body of text associated with it (Liang et al., 2024),(Chang et al., 2023),(Cao et al., 2023). (Mohn et al., 2018) propose to use a random forest to train an automatic image classifier with orders of magnitude less training data required than would be required for a CNN-based colorizer. (Oh et al., 2014) propose to use colorization as a method to improve image coding based on local regression. Two of the main methods that we used to compare against are DeOldify (Antic, 2019) and Generative Color Prior (GCP) (Wu et al., 2022). DeOldify (Antic, 2019) is a self-attention generative adversarial network-based automatic image colorizer (Zhang et al., 2018). It is trained with a two-time scale update rule (Heusel et al., 2017). GCP (Wu et al., 2022) is a generative adversarial network-based automatic image colorization-based system which leverages a learned generative prior to colorizing images.

As none of these methods have temporal consistency developed they cannot colorize videos as well as a system like FRCol, which is designed specifically for videos.

2.2 Automatic Video Colorization

One of the simplest ways of attempting automatic video colorization is to decompose the video into a sequence of frames, colorize each frame individually using an automatic image colorizer and then recompile the video sequence from the colorized frames. The problem with this approach is that the frames are colorized independently, so temporal consistency is not ensured. This can result in colorization, which appears to change color or frequently flicker, giving a very unnatural finish to the colorizations. Some more sophisticated approaches exist that design for temporal consistency by default (Liu et al., 2023),(Wan et al., 2022),(Blanch et al., 2023). (Ramos and Flores, 2019) propose to colorize one frame of a sequence and then propagate that frame’s color through the video sequence by matching intensity and texture descriptors. (Ward et al., 2024) propose LatentColorization, a temporally consistent automatic speaker video colorization system which leverages latent-diffusion priors and a temporal consistency mechanism. Our approach improves over LatentColorization in that FRCol can accept the additional condition of retrieved exemplars, which can reduce color hallucinations. We compared against Video Colorization with Video Hybrid Generative Adversarial Network (VCGAN) (Zhao et al., 2023) in our evaluation section. VCGAN is a recurrent colorization system designed with temporal consistency in mind, as it uses a feed-forward feature extractor and a dense long-term temporal consistency loss. As VCGAN is a GAN-based system, it is susceptible to mode collapse and, in particular, bland colorizations, which our model is not as it is diffusion-based.

2.3 Exemplar Guided Video Colorization

One subsection of automatic video colorization particularly relevant to this work is exemplar-guided video colorization. Exemplar-guided video colorization takes an exemplar frame and grayscale video as input. It then uses the color information provided in the exemplar frame to guide the resultant colorization (Ward and Breslin, 2022),(Endo et al., 2021),(Xu et al., 2020),(Akimoto et al., 2020),(Lu et al., 2020),(Zhang et al., 2019). (Iizuka and Simo-Serra, 2019) propose DeepRemaster, an automatic

video colorization system based on temporal convolutional neural networks with attention mechanisms. It was trained with artificially deteriorated videos. DeepRemaster has no exemplar retrieval system incorporated into its design, so it is more susceptible to unnatural colorization than FRCol.

2.4 Face Recognition

There are generally four steps involved in face recognition: face detection (Kumar et al., 2019), normalization (Djamaluddin et al., 2020), feature extraction (Benedict and Kumar, 2016) and finally, face recognition. Plentiful textual resources exist on facial recognition technologies (Chen and Jenkins, 2017),(Filali et al., 2018),(Geetha et al., 2021). (Chen and Jenkins, 2017) propose using Principal Component Analysis (PCA) and K-Nearest Neighbours (KNN), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). (Filali et al., 2018) propose Haar-AdaBoost, LBP-AdaBoost, GF-SVM and GFNN. Haar-AdaBoost is a combination of Haar cascade classifiers and AdaBoost machine learning algorithm. Local binary patterns (LBP) are used instead of the Haar cascade classifiers in the LBP-AdaBoost formulation. Gabor Filters are used for GF-SVM and GFNN, with the difference between the two being that a support vector machine is used for GF-SVM and a neural network for GFNN. (Geetha et al., 2021) compare an Eigenface method, PCA, CNN, and SVM for face recognition. Technical challenges associated with face recognition technologies exist. Three of the most common ones are improper lighting (Fahmy et al., 2006), low-quality images (Li et al., 2019), and various angles of view (Troje and Bülthoff, 1996). More recently, there has been a tendency in the literature towards systems that leverage deep learning to handle specific constraints such as low power consumption (Alansari et al., 2023) or occlusions (Mare et al., 2021).

3 METHODOLOGY

3.1 Data Processing

Following on from (Ward et al., 2024), we use the Grid (Cooke et al., 2006) and Lombard Grid (Alghamdi et al., 2018) datasets. The Grid dataset consists of high-quality video recordings of 1000 sentences spoken by each of the 34 talkers. The Lombard Grid dataset is a high-quality collection of speaker videos of 54 subjects saying 5400 utterances. All of the frames were resized to 128x128 pixels. The orig-

 Algorithm 1: FRCol.

Require: Input: Face Database F ,
Grayscale Video V

Require: Modules:

Face Recognition Module $FR : V \rightarrow D$

Automatic Colorizer $AC : V \rightarrow \tilde{V}$

Exemplar Selection Module $ESM : V, D \rightarrow \tilde{e}$

Exemplar Guided Colorizer $EGC : V, \tilde{e} \rightarrow \tilde{V}$

Ensure: Colorized Video \tilde{V}

- 1: Prompt FR to generate the Decision D given the Grayscale Video V .
 - 2: If Decision D is no, use AC to colorize the Grayscale Video V without guidance.
 - 3: Else choose the most relevant Exemplar \tilde{e} from the Face Database F using the Exemplar Selection Module ESM .
 - 4: Then colorize the Grayscale Video V with the Exemplar Guided Colorizer EGC given the selected Exemplar \tilde{e} .
 - 5: **return** Colorized Video \tilde{V}
-

inal frames were in color and needed to be converted to grayscale. 10,000 frames were used for training and 1,500 for testing, giving approximately a standard 90/10 split.

3.2 FRCol System Description

The proposition is to guide colorization using exemplars retrieved from the face database using the exemplar selection module if the face recognition module identifies a face. The concept is that instead of relying on an end-to-end colorizer to learn what color particular objects are, it can be guided using exemplars retrieved via face recognition. See Fig. 2 and Algorithm 1. The black-and-white video is initially passed through the face recognition module, Step 1. If the face recognition module does not recognize a face in the video, it reverts to colorization without exemplar conditioning, Step 2a. If the face recognition module detects a face in the frames, it queries the faces database for the most similar face using the exemplar selection module. This face is passed onto the conditioning mechanism of the colorizer. Finally, the colorizer takes the conditions that it has been passed, the black-and-white video and the exemplar frame if a face has been detected, and it performs its colorization process. This results in the colorized video, Step 2b.

3.3 Face Recognition Module

The face recognition algorithm used for this project is a pre-trained ResNet-34 similar to that used in (He

et al., 2015). It was trained on 3 million faces taken from the FaceScrub (Ng and Winkler, 2014) and VGG (Parkhi et al., 2015) datasets. It was then tested on the Labelled Faces in the Wild (Huang et al., 2007) benchmark, where it achieved an accuracy of 99.38%.

3.4 Exemplar Selection Module

The exemplar selection module calculates the minimum Euclidean distance $\min(\|\cdot\|)$ between the embedding of the black-and-white face Z_{bw} and the embedding of every exemplar face Z_i in the faces database $\forall i \in I$, see eqn 1. It then returns the exemplar with the lowest value Z_e .

$$Z_e = \min(\|Z_{bw} - Z_i\|) \forall i \in I \quad (1)$$

3.5 Face Database

The face database consists of faces taken from the train set of the Grid and Lombard Grid datasets. The relevancy of the faces in the face database substantially impacts the quality of the resultant colorizations. For our experiments, we allowed the model to use subjects from the train set of the datasets on the test set inferences. This limits the application of this approach to cases where similar exemplar images exist of the faces of the persons in the video being colorized.

3.6 Colorizer

During training, the current frame ground truth, the black-and-white current frame, the previous frame, and the exemplar frame are input to the colorizer. During inference, the current frame ground truth is replaced with Gaussian noise as the model will not have access to the ground truth. See Fig. 3. The critical elements of the colorizer are:

Image Encoder. This component (implemented as a Vector Quantised-Variational AutoEncoder (VQ-VAE) (van den Oord et al., 2018)) encodes the input frames into embedding representations. It generates the ground truth embedding or the Gaussian noise embedding Z_T depending on whether the system is in training or inference mode, the embedding of the current black-and-white frame Z_{BW} , the embedding of the previous color frame Z_P , and the embedding of the exemplar frame Z_E .

Denoising U-Net. The denoising U-Net is responsible for denoising the embeddings generated by the image encoder Z_{T-1} . It is sampled T (timesteps) until a satisfactory level of noise removal has occurred. T

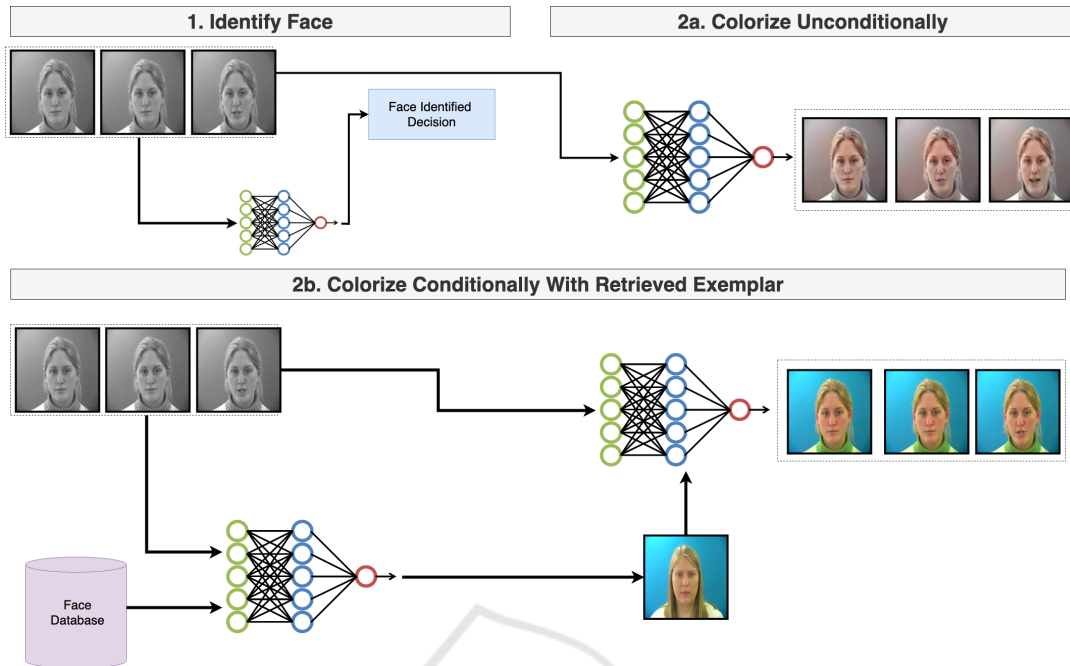


Figure 2: **This diagram depicts the overall system architecture.** Initially, face recognition is performed on the black-and-white video to check whether a face exists in the video, Step 1. If a face is nonexistent, the black-and-white video is colorized without exemplar conditioning, Step 2a. If a face is detected, the face database is queried for the closest exemplar face. This exemplar face is then used to guide the latent diffusion-based colorizer to colorize the video, Step 2b.

is a hyperparameter set to 1000 for training and 50 for inference in our experiments.

Conditioning Mechanism. The conditioning mechanism provides contextual information and conditioning signals to guide the colorization process. It concatenates the various embeddings, including Z_{BW} , Z_P , Z_T , and Z_E , which represent the black-and-white input frame, the output of the model for the previous frame, the noisy frame to be denoised, and the exemplar frame.

Image Decoder. This component (the same VQVAE as the image encoder) decodes the predicted frames from their embedding representations. It generates the predicted frame from the predicted frame embedding Z_T .

4 EVALUATION AND DISCUSSION

FRCol was tested under various circumstances to determine its performance. The metrics used to parametrize the evaluation are defined in subsection 4.1. The colorizers are compared visually in subsection 4.2. This is followed by a numeric evaluation using the objective metrics in subsection 4.3. An ablation study is conducted to determine the impor-

tance of the various aspects of FRCol in subsection 4.4. This is followed by gathering user opinions in subsection 4.5. A real-world example concludes this section in subsection 4.6.

4.1 Metrics

Evaluating colorizers is challenging as it is a subjective task with no consensus on the best way to achieve it. We will follow the most standard practice of employing subjective and objective metrics. Specifically, we choose to use four objective and one subjective metric. The objective metrics are Peak Signal to Noise Ratio (PSNR) (Fardo et al., 2016), Structural Similarity Index (SSIM) (Wang et al., 2004), Fréchet Inception Distance (FID) (Heusel et al., 2018) and Fréchet Video Distance (FVD) (Unterthiner et al., 2019). The subjective metric we used is Mean Opinion Score (MOS) (Mullery and Whelan, 2022). PSNR compares a source and target image on a per-pixel basis. A higher PSNR indicates two more similar images from a pixel difference perspective. The difficulty with this metric is that humans do not evaluate images on a per-pixel metric but instead on more of a per-image basis. This means that PSNR sometimes does not correlate with human perception. SSIM improves this limitation by comparing a source and target image on an object similarity level instead of per

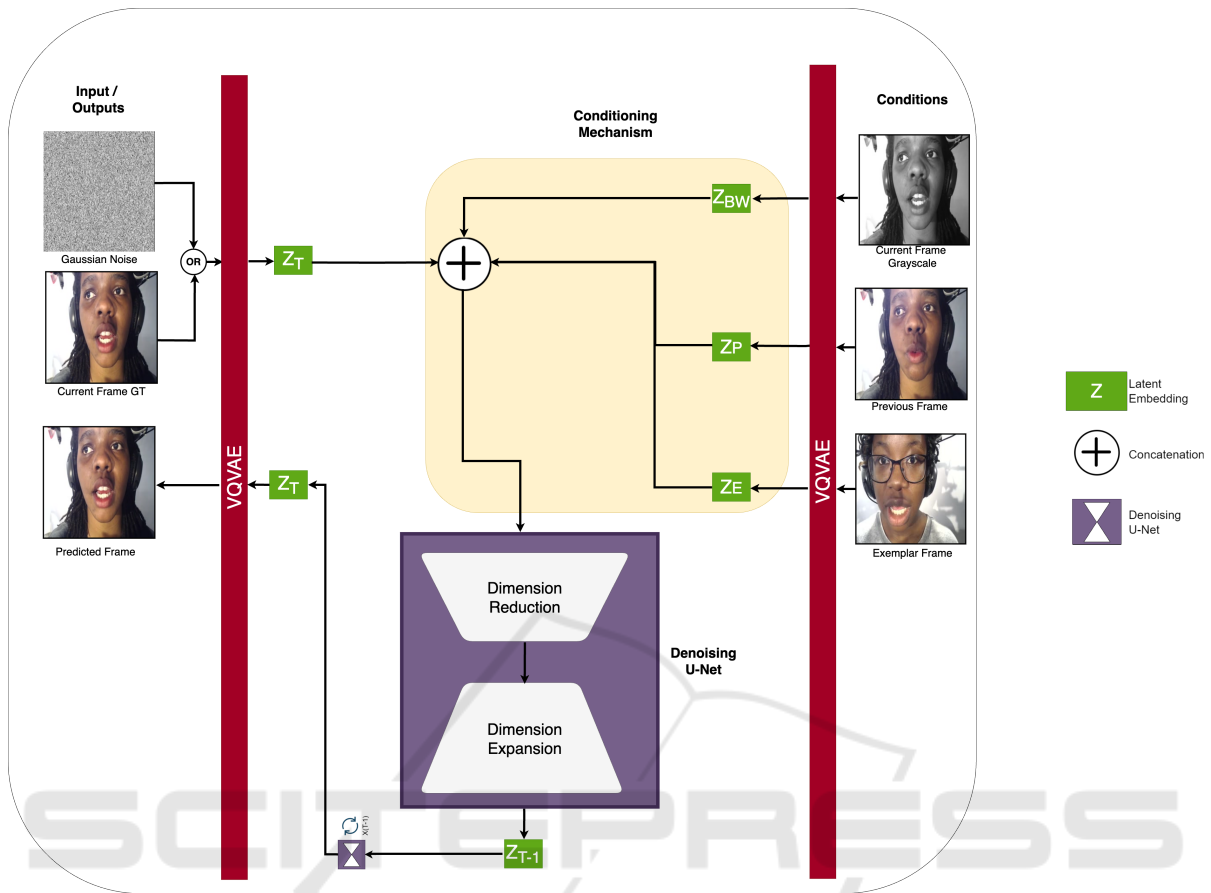


Figure 3: **The colorizer architecture during training and testing is depicted in the diagram.** This illustrates the network’s key elements and interactions: image encoder and decoder (VQVAE), denoising U-Net and conditioning mechanism.

pixel. A higher SSIM indicates two images that have more similar objects. The challenge with SSIM is that it compares images pairwise instead of their distributions. FID improves upon this limitation by considering the distribution of the colorizations instead of pairwise image comparison. A lower FID indicates two color distributions which are more closely aligned and therefore a better colorization. The issue with using FID is that it is designed to compare images and does not account for temporal consistency, which is essential for automatic video colorization. FVD builds upon this limitation in that, as well as considering the distributions of the colorizations; it also considers the temporal consistency between frames. Each metric mentioned above is objective, calculating a difference from a ground truth. However, as colorization is subjective, we must also deploy a subjective metric, specifically MOS. We calculate MOS as the percentage preference of a specific method in a user study. In recognition of the different capabilities of each of the metrics, we have chosen to report on all of them to give a holistic evaluation.

4.2 Qualitative Analysis

Fig. 4 provides a visual representation of the comparison of FRCol with contemporary automatic video colorization methods. The Grid (top) and the Lombard Grid (bottom) datasets are used to evaluate the methods. The observations mirror each other for both datasets. The outputs of each of the systems are shown column-by-column. Each previous state-of-the-art has either colorized the outputs dull (DeOldify, DeepRemaster) or with poor fidelity to the ground truth (GCP, VCGAN, LatentColorization) apart from FRCol. DeOldify’s lack of colorfulness is consistent with the idea that GANs, which DeOldify is based on, can be susceptible to mode collapse, where they produce limited and less diverse color variations. GCP has produced colorful output but 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 dataset. This could potentially be a result of its retrieval mechanism. VCGAN has produced a blue filter-type effect on the frames.



Figure 4: **The qualitative comparison of colorization results from various systems.** Included in this diagram are DeOldify, GCP, VCGAN, LatentColorization, DeepRemaster, FRCol, and the ground truth for both the GRID dataset (top) and the Lombard Grid dataset (bottom) is shown.

DeepRemaster performs better when given plentiful exemplars; when it does not have this, it resorts to bland, dull colors. LatentColorization has colorized with high color fidelity to the ground truth. One comment that can be made is that frame 3 of the Grid dataset LatentColorization has failed to ensure spatial consistency of the subject’s top, with one shoulder being red and the other navy. It is challenging to differentiate between FRCol and the ground truth visually.

4.3 Quantitative Analysis

An important point to make before comparing the methods quantitatively is that the amount of computing each method has used in training should be proportional to their results. This is particularly relevant for LatentColorization, which has used 33 more epochs to train than FRCol. In light of this, we chose the next highest-performing system, DeOldify, as the previous state-of-the-art.

Comparing the approaches quantitatively in Table 1, we can see that FRCol has achieved strong results across all datasets and metrics. FRCol achieves the best score on all metrics except PSNR in the Grid dataset experiment, where it is only bested by LatentColorization. In the Lombard Grid dataset, FRCol achieves the best FVD score. FRCol achieves the optimal FVD score on average across the experiments. Normalizing and comparing the averaged scores shows that our approach performs 13% better than the previous SOTA, DeOldify. On the Grid dataset, FRCol performs on average 17% better than DeOldify. On the Lombard Grid dataset, FRCol performs, on average, 8% better than DeOldify.

Although LatentColorization achieves the optimal score in many metrics, on average, across the datasets, FRCol performs 1% better, indicating that even with less training compute, it can perform at a similar level to LatentColorization.

Table 1: **The quantitative comparisons provide a detailed evaluation of different colorization methods across various datasets.** These methods include DeOldify, GCP, VCGAN, LatentColorization, DeepRemaster, and FRCol. The evaluation criteria encompass several metrics, including PSNR, SSIM, FID, and FVD. It also outlines what conditions the approaches accept and how much computing was used to train them. Arrows indicate the optimal direction of the score, i.e. \uparrow indicates higher is better, \downarrow indicates lower is better.

Dataset	Conditions	Compute	Method	PSNR \uparrow	SSIM \uparrow	FID \downarrow	FVD \downarrow
Grid	None	Unknown	DeOldify (Antic, 2019)	28.16	0.81	58.04	694.62
	None	ImageNet @ 20 epochs	GCP (Wu et al., 2022)	27.92	0.80	79.78	844.93
	None	Unknown	VCGAN (Zhao et al., 2023)	27.95	0.85	63.15	931.00
	None	Grid + Lombard Grid @ 376 epochs	LatentColorization (Ward et al., 2024)	30.00	0.85	38.63	311.73
	Exemplar	Unknown	DeepRemaster (Iizuka and Simo-Serra, 2019)	27.83	0.79	90.15	993.49
	Exemplar	Grid + Lombard Grid @ 343 epochs	FRCol	29.69	0.85	37.60	280.25
Lombard Grid	None	Unknown	DeOldify (Antic, 2019)	29.73	0.92	35.08	385.21
	None	ImageNet @ 20 epochs	GCP (Wu et al., 2022)	30.01	0.96	36.12	314.55
	None	Unknown	VCGAN (Zhao et al., 2023)	29.19	0.97	57.24	813.83
	None	Grid + Lombard Grid @ 376 epochs	LatentColorization (Ward et al., 2024)	31.14	0.94	25.67	245.71
	Exemplar	Unknown	DeepRemaster (Iizuka and Simo-Serra, 2019)	30.55	0.93	99.50	460.36
	Exemplar	Grid + Lombard Grid @ 343 epochs	FRCol	30.51	0.94	27.20	218.57
Overall	None	Unknown	DeOldify (Antic, 2019)	28.95	0.86	46.56	539.92
	None	ImageNet @ 20 epochs	GCP (Wu et al., 2022)	28.96	0.88	57.95	579.74
	None	Unknown	VCGAN (Zhao et al., 2023)	28.57	0.91	60.20	872.41
	None	Grid + Lombard Grid @ 376 epochs	LatentColorization (Ward et al., 2024)	30.57	0.89	32.15	278.72
	Exemplar	Unknown	DeepRemaster (Iizuka and Simo-Serra, 2019)	29.19	0.86	94.82	726.92
	Exemplar	Grid + Lombard Grid @ 343 epochs	FRCol	30.10	0.89	32.40	249.41

Table 2: **Ablation test of the FR module.** \uparrow and \downarrow indicates the direction of optimal performance. The best scores are highlighted in bold. - FR refers to the method that does not leverage face recognition.

Dataset	Method	PSNR \uparrow	SSIM \uparrow	FID \downarrow	FVD \downarrow
Grid	FRCol	29.69	0.85	37.60	280.25
	- FR	28.03	0.71	56.64	571.33
Lombard Grid	FRCol	30.51	0.94	27.20	218.57
	- FR	30.03	0.94	35.19	247.90
Overall	FRCol	30.10	0.89	32.40	249.41
	- FR	29.03	0.82	45.91	409.61

4.4 Ablation Study

An ablation study was also carried out to evaluate the significance of certain system aspects on overall performance; see Table 2. The central element of the system being ablated was the face recognition module. To achieve this, FRCol was compared against the system without facial recognition technology, namely - FR. - FR was constructed by using a random face as the condition so the impact of a relevant exemplar could be investigated. FRCol performs on average 16% better across the metrics than - FR on the Grid dataset. FRCol performs on average 3% better across the metrics than - FR on the Lombard Grid dataset. FRCol performs on average 9% better across the metrics than - FR on the overall dataset. The face recognition has much more of a performance gain on the Grid than the Lombard Grid dataset. This could be due to Grid having more similar faces and, therefore, a more relevant set of exemplars.

4.5 User Study

A user study was conducted to get a more subjective view of FRCol’s performance. This study aimed to evaluate the difference in performance between our proposed approach, FRCol, and the previous SOTA DeOldify. 16 participants were shown two sets of three videos and asked a question on each set.

For the Grid 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 FRCol, the other by DeOldify, and the third was the ground truth. The ground truth video was labelled as such, whereas the FRCol and DeOldify versions of the video were anonymous. To distinguish the FRCol version of the video from the DeOldify version 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 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 FRCol, the other by DeOldify, 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 based on which one looked the most realistic. Therefore, this question (Question 2) acted as a visual Turing test (Tur-

ing, 1950) 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.

We then collated, analysed, and visualized the user study results; see Fig. 5 and Fig. 6. In Fig. 5, the X axis represents the MOS score for each method, and the Y axis differentiates between DeOldify and FRCOL. The MOS score is the percentage preference for each technique. In Fig. 6, the X axis represents the average score for each method, and the Y axis differentiates between DeOldify, the ground truth and FRCOL. The average score is the tally of each score per method divided by the number of participants. We used the average score for this figure as this more accurately displayed the relevant information for a multi-class ranking question.

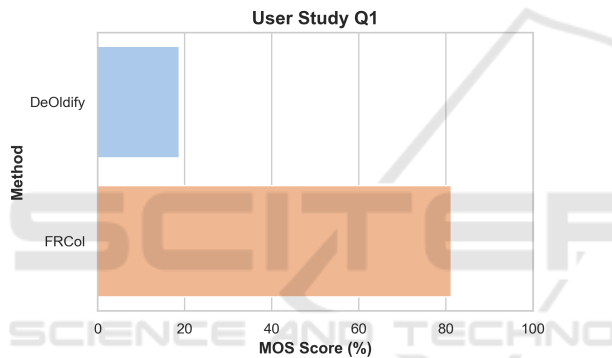


Figure 5: The head-to-head user study results between FRCOL and DeOldify on the Grid dataset. The X-axis represents the (MOS) for each question’s methods. The Y axis indicates the relevant method. The participants were unaware of which video was from which colorizer. They were asked which video was closer to the ground truth.

From the graph, we can see that overall, FRCOL was preferred to DeOldify. For Question 1, DeOldify received a MOS score of 19%, and FRCOL received a MOS score of 81%, indicating a strong preference for FRCOL on this question. For Question 2, the ground truth received the highest average score of 2.81, followed by FRCOL at 1.94 and DeOldify at 1.25. Summarising this result, the ground truth was followed by FRCOL and finally DeOldify in terms of average score.

4.6 Real World Example

To fully evaluate an automatic video colorization system, it must work on authentic archival material as well as dataset videos. In recognition of this, we colorize an excerpt from “The Adventures of Sherlock Holmes (1984)”; See Fig. 1. The output of FRCOL

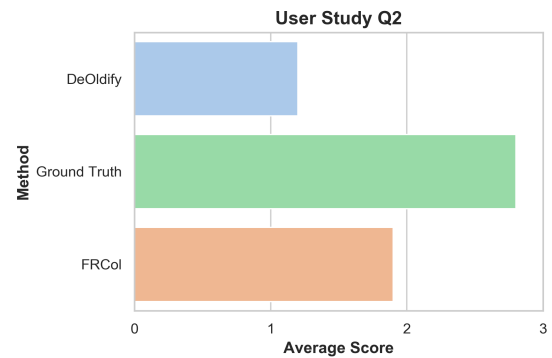


Figure 6: The user study results for Question 2 (Lombard Grid). The X-axis represents the average score, with 0 being the worst and 3 best. The Y axis indicates the relevant method. The participants were unaware of which video was which. They were asked to rate each video regarding its realism and consistency.

is shown at the bottom, and the grayscale version is shown at the top. The comparison demonstrates that FRCOL applies to authentic archival material. It correctly segmented the subject from the background and applied realistic colors to both the actor and the background.

We developed a user interface to facilitate interaction with the FRCOL system; see Fig. 7. The interface allows the user to specify the grayscale video they wish to colorize and a file path to a custom faces database from which they would like the algorithm to choose the most relevant exemplar. The system defaults to the standard faces database if no file path is provided. Once the grayscale video and optional faces database file path have been entered into the user interface, there is a simple colorization button to submit the request to colorize. Once the colorization has been performed, the colorized video is returned to the user interface, where it is presented beside the input grayscale video.

5 CONCLUSION

Automatic speaker video colorization performance can be improved by augmenting a system with exemplars retrieved using facial recognition technology. This performance gain has been demonstrated to span various datasets and metrics. Specifically, we achieved a 13% average increase across both datasets on the Grid and Lombard Grid on the PSNR, SSIM, FID, and FVD scores compared to the previous SOTA DeOldify. This objective evaluation was further shown in our subjective user study, where FRCOL was preferred to contemporary colorizers 81% of the time. Such a system applies to authentic histor-

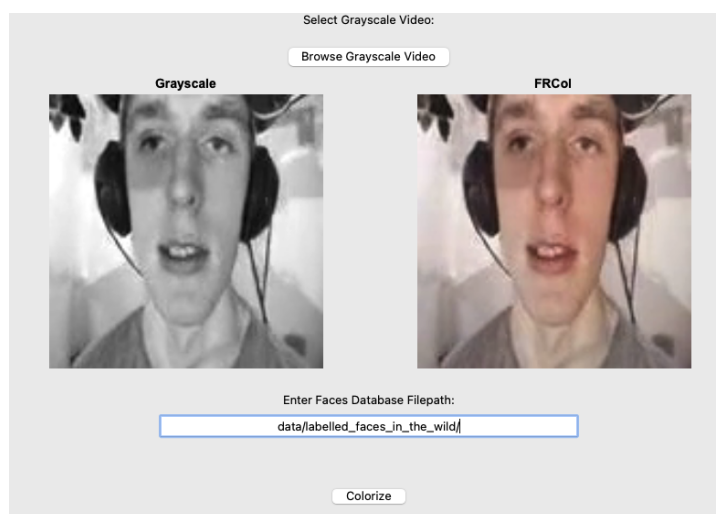


Figure 7: User interface for the FRCol application. It takes a grayscale video and an optional path to a custom faces database as input. It outputs the resultant colorization played parallel to the input grayscale video.

ical material, such as old Sherlock Holmes movies and modern datasets. It can also be easily deployed in an intuitive user application, which colorize grayscale videos based on custom face databases.

LIMITATIONS & FUTURE WORK

FRCol, like any system, has its limitations. Firstly, it is fine-tuned on speaker data and has limited generalizability to out-of-domain data. Secondly, training and testing large computer vision models are compute-intensive and costly for the environment. Thirdly, there are ethical implications associated with colorization, the most prominent being concerns around the model learning biases from the datasets and reflecting that in its colorizations. Finally, the quality of the colorizations is highly dependent on the relevancy of the exemplar images contained in the faces database. This approach assumes that exemplar images from the train portion of the same dataset being tested are available in the faces database. If this assumption is untrue, there is a degradation in performance.

In the future, we would like to improve this work's limitations. The system should be able to generalize to out-of-domain data. We plan to achieve this by enhancing the diversity of data on which the system is trained and incorporating an object detection module. We want to improve our system's efficiency by investigating more effective sampling methods to reduce the number of iterations required to train and infer. We plan to consider the ethical implications of our work more deeply. An actionable item in this topic

could be creating a model card describing the system, dataset, biases and limitations. Some work can be done on the model's ability to perform when less relevant exemplars are exclusively available.

ACKNOWLEDGEMENTS

This work was conducted with the financial support of Taighde Éireann - Research Ireland through the Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223 and the Insight Research Ireland Centre for Data Analytics under Grant No. 12/RC/2289_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We would like to thank the reviewers for their valuable insights.

REFERENCES

- Akimoto, N., Hayakawa, A., Shin, A., and Narihira, T. (2020). Reference-based video colorization with spatiotemporal correspondence.
- Alansari, M., Hay, O. A., Javed, S., Shoufan, A., Zweiri, Y., and Werghi, N. (2023). Ghostfacenet: Lightweight face recognition model from cheap operations. *IEEE Access*, 11:35429–35446.
- Alghamdi, N., Maddock, S., Marxer, R., Barker, J., and Brown, G. (2018). A corpus of audio-visual lombard speech with frontal and profile views. *Journal of the Acoustical Society of America*, 143.
- Antic, J. (2019). Deoldify. <https://github.com/jantic/DeOldify>.

- Benedict, S. R. and Kumar, J. S. (2016). Geometric shaped facial feature extraction for face recognition. In *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, pages 275–278.
- Blanch, M. G., O'Connor, N., and Mrak, M. (2023). Scene-adaptive temporal stabilisation for video colourisation using deep video priors. In Karlinsky, L., Michaeli, T., and Nishino, K., editors, *Computer Vision – ECCV 2022 Workshops*, pages 644–659, Cham. Springer Nature Switzerland.
- Cao, Y., Meng, X., Mok, P. Y., Liu, X., Lee, T.-Y., and Li, P. (2023). Animediffusion: Anime face line drawing colorization via diffusion models.
- Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., and Shi, B. (2023). L-coins: Language-based colorization with instance awareness. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19221–19230, Los Alamitos, CA, USA. IEEE Computer Society.
- Chen, J. and Jenkins, W. K. (2017). Facial recognition with pca and machine learning methods. In *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 973–976.
- Chen, S.-Y., Zhang, J.-Q., Zhao, Y.-Y., Rosin, P. L., Lai, Y.-K., and Gao, L. (2022). A review of image and video colorization: From analogies to deep learning. *Visual Informatics*, 6(3):51–68.
- Choi, I. and Kim, D. (2010). Facial fraud discrimination using detection and classification. In Bebis, G., Boyle, R., Parvin, B., Koracin, D., Chung, R., Hammound, R., Hussain, M., Kar-Han, T., Crawfis, R., Thalmann, D., Kao, D., and Avila, L., editors, *Advances in Visual Computing*, pages 199–208, Berlin Heidelberg. Springer Berlin Heidelberg.
- Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006). The grid audio-visual speech corpus. Collection of this dataset was supported by a grant from the University of Sheffield Research Fund.
- Curtiz, M. (1942). *Casablanca (1942) - in color*. Starring Humphrey Bogart, Ingrid Bergman, Paul Henreid, Claude Rains, Sydney Greenstreet.
- Djamaluddin, M., Hamonangan, N. M., and Editri, S. A. (2020). Normalization of facial pose and expression to increase the accuracy of face recognition system. *Journal of Physics: Conference Series*, 1539(1):012035.
- Dodson, C. T. J., Soldera, J., and Scharcanski, J. (2021). Some information geometric aspects of cyber security by face recognition. *Entropy*, 23(7).
- Endo, R., Kawai, Y., and Mchizuki, T. (2021). A practical monochrome video colorization framework for broadcast program production. *IEEE Transactions on Broadcasting*, 67(1):225–237.
- Erickson, K. and Schulkin, J. (2003). Facial expressions of emotion: A cognitive neuroscience perspective. *Brain and Cognition*, 52(1):52–60. Affective Neuroscience.
- Fahmy, G., El-Sherbeeny, A., Mandala, S., Abdel-Mottaleb, M., and Ammar, H. (2006). The effect of lighting direction/condition on the performance of face recognition algorithms. In Flynn, P. J. and Pankanti, S., editors, *Biometric Technology for Human Identification III*, volume 6202, page 62020J. International Society for Optics and Photonics, SPIE.
- Fardo, F. A., Conforto, V. H., de Oliveira, F. C., and Rodrigues, P. S. (2016). A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms.
- Filali, H., Riffi, J., Mahraz, A. M., and Tairi, H. (2018). Multiple face detection based on machine learning. In *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–8.
- Geetha, M., Latha, R., Nivetha, S., Hariprasath, S., Gowtham, S., and Deepak, C. (2021). Design of face detection and recognition system to monitor students during online examinations using machine learning algorithms. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2018). Gans trained by a two time-scale update rule converge to a local nash equilibrium.
- Hitchcock, A. (1960). *Psycho (1960) - in color*. <https://archive.org/details/psycho-1960-in-color>. Starring Anthony Perkins, Janet Leigh, Vera Miles, John Gavin, Martin Balsam.
- Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.
- Iizuka, S. and Simo-Serra, E. (2019). DeepRemaster: Temporal Source-Reference Attention Networks for Comprehensive Video Enhancement. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia 2019)*, 38(6):1–13.
- Jain, A., Arora, D., Bali, R., and Sinha, D. (2021). Secure authentication for banking using face recognition. *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, 2(2):1–8.
- Kouzouglidis, P., Sfikas, G., and Nikou, C. (2019). Automatic video colorization using 3d conditional generative adversarial networks.
- Kumar, A., Kaur, A., and Kumar, M. (2019). Face detection techniques: A review. *Artificial Intelligence Review*, 52.
- Li, P., Prieto, L., Mery, D., and Flynn, P. (2019). Face recognition in low quality images: A survey.
- Liang, Z., Li, Z., Zhou, S., Li, C., and Loy, C. C. (2024). Control color: Multimodal diffusion-based interactive image colorization.
- Liu, H., Xie, M., Xing, J., Li, C., and Wong, T.-T. (2023). Video colorization with pre-trained text-to-image diffusion models.

- Liu, S. and Zhang, X. (2012). Automatic grayscale image colorization using histogram regression. *Pattern Recognition Letters*, 33(13):1673–1681.
- Lu, P., Yu, J., Peng, X., Zhao, Z., and Wang, X. (2020). Gray2colormet: Transfer more colors from reference image. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 3210–3218, New York, NY, USA. Association for Computing Machinery.
- Mare, T., Duta, G., Georgescu, M.-I., Sandru, A., Alexe, B., Popescu, M., and Ionescu, R. T. (2021). A realistic approach to generate masked faces applied on two novel masked face recognition data sets.
- Martin, J. (2009). Ww2 in color. A documentary on World War II, featuring colorized footage.
- Mohn, H., Gaebelein, M., Hänsch, R., and Hellwich, O. (2018). Towards image colorization with random forests. In *VISIGRAPP (4: VISAPP)*, pages 270–278.
- Mullery, S. and Whelan, P. F. (2022). Human vs objective evaluation of colourisation performance.
- Ng, H.-W. and Winkler, S. (2014). A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347.
- Oh, P., Lee, S. H., and Kang, M. G. (2014). Local regression based colorization coding. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 153–159. IEEE.
- Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.
- Pierre, F. and Aujol, J.-F. (2021). *Recent Approaches for Image Colorization*, pages 1–38. Springer International Publishing, Cham.
- Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 145–151, New York, NY, USA. Association for Computing Machinery.
- Ramos, A. P. and Flores, F. C. (2019). Colorization of grayscale image sequences using texture descriptors. In *VISIGRAPP (4: VISAPP)*, pages 303–310.
- S, P. (2023). Detailed survey of machine learning algorithms for face recognition. *International Journal of Creative Research Thoughts*, 11:b832–b836.
- Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. (2022). Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10.
- Sanchez del Rio, J., Moctezuma, D., Conde, C., Martin de Diego, I., and Cabello, E. (2016). Automated border control e-gates and facial recognition systems. *Computers & Security*, 62:49–72.
- Sardar, A., Umer, S., Rout, R. K., Wang, S.-H., and Tanveer, M. (2023). A secure face recognition for iot-enabled healthcare system. *ACM Trans. Sen. Netw.*, 19(3).
- Troje, N. F. and Bühlhoff, H. H. (1996). Face recognition under varying poses: The role of texture and shape. *Vision Research*, 36(12):1761–1771.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(October):433–60.
- Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. (2019). FVD: A new metric for video generation.
- van den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2018). Neural discrete representation learning.
- Wan, Z., Zhang, B., Chen, D., and Liao, J. (2022). Bringing old films back to life.
- Wang, M. and Deng, W. (2021). Deep face recognition: A survey. *Neurocomputing*, 429:215–244.
- Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.
- Ward, R., Bigioi, D., Basak, S., Breslin, J. G., and Corcoran, P. (2024). Latentcolorization: Latent diffusion-based speaker video colorization.
- Ward, R. and Breslin, J. G. (2022). Towards temporal stability in automatic video colourisation. In *The 24th Irish Machine Vision and Image Processing Conference (IMVIP 2022)*.
- Weng, S., Sun, J., Li, Y., Li, S., and Shi, B. (2022). Ct2: Colorization transformer via color tokens. In *ECCV*.
- Wu, Y., Wang, X., Li, Y., Zhang, H., Zhao, X., and Shan, Y. (2022). Towards vivid and diverse image colorization with generative color prior.
- Xu, Z., Wang, T., Fang, F., Sheng, Y., and Zhang, G. (2020). Stylization-based architecture for fast deep exemplar colorization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9360–9369.
- Zhang, B., He, M., Liao, J., Sander, P. V., Yuan, L., Bermak, A., and Chen, D. (2019). Deep exemplar-based video colorization.
- Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-attention generative adversarial networks.
- Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization.
- Zhao, P., Chen, Y., Zhao, Y., Jia, W., Zhang, Z., Wang, R., and Hong, R. (2024). Audio-infused automatic image colorization by exploiting audio scene semantics.
- Zhao, Y., Po, L.-M., Yu, W.-Y., Rehman, Y. A. U., Liu, M., Zhang, Y., and Ou, W. (2023). Vcgan: Video colorization with hybrid generative adversarial network. *IEEE Transactions on Multimedia*, 25:3017–3032.