

Video Colorization Using VGG16 Autoencoder

Devyn Miller
Chapman University
Orange, United States
devmiller@chapman.edu

Abstract—This work presents a deep learning approach to colorize monochromatic media using a VGG16-based autoencoder architecture. The model leverages transfer learning from ImageNet-pretrained VGG16 as an encoder and a custom decoder network to predict color channels in the LAB color space. The system addresses key challenges in media colorization including plausible color inference, semantic preservation, and temporal consistency. Training was conducted over 20 epochs, achieving accuracy levels above 77% with consistent loss reduction. The model demonstrates the capability to effectively transform grayscale images into colorized representations, with applications in historical media restoration, educational materials, and creative processes. Results indicate that leveraging pre-trained feature extraction capabilities combined with a carefully designed upsampling decoder can successfully learn complex grayscale-to-color transformations.

Keywords—image colorization, deep learning, VGG16, transfer learning, computer vision, autoencoder, LAB color space

I. INTRODUCTION

A. Problem Statement

In digital media, color is pivotal in enhancing content’s visual appeal and interpretability. A vast repository of historical and archival footage, educational materials, and artistic works exists in grayscale, limiting the potential impact on contemporary audiences accustomed to color media. This limitation affects these works’ aesthetic value and ability to convey emotions, highlight details, and engage viewers effectively.

Our project provides a deep-learning approach to induce compelling and visually appealing colors in monochromatic footage. Media colorization presents several challenges:

- 1) **Plausible Color Inference:** Inferring plausible colors involves understanding the context, objects, and scenes within the footage. This challenge is compounded by multiple color possibilities for any given pixel in a grayscale image, making choosing the most appropriate colors non-trivial.
- 2) **Semantic Preservation:** Ensuring that the colorization process respects the semantic integrity of the original footage is crucial. This involves accurately identifying and coloring objects, backgrounds, and human subjects in a manner that aligns with their real-world appearances.
- 3) **Temporal Consistency (Video):** One of the most significant challenges in video colorization is maintaining color consistency across consecutive frames. Unlike static images, videos require that colorization be tem-

porally coherent to avoid flickering and color bleeding, which detract from the viewer’s experience.

By developing an advanced colorization algorithm, our project addresses these challenges by leveraging deep learning techniques and state-of-the-art image processing technologies. Our model architecture is based on the VGG16 framework, chosen for its robust feature extraction capabilities. By doing so, we aim to unlock the full potential of grayscale media, making them more accessible, engaging, and informative for modern audiences.

B. Significance of the Project

The significance of transforming grayscale media touches upon media consumption, historical documentation, and the utilization of artificial intelligence in creative processes.

Historical footage and photographs serve as a visual testament to the past, capturing moments that define human history. However, the monochromatic nature of such media can create a barrier to engagement for modern audiences accustomed to the vividness of color content. By colorizing these historical artifacts, we can bridge the gap between past and present, offering a more relatable and immersive glimpse into bygone eras. This not only aids in educational and cultural preservation but also helps in generating renewed interest and understanding of historical events.

Color adds layers of information, making it more intuitive for viewers to discern details and understand complex concepts. This is particularly important in a world where digital media is an integral part of education and where the quality of visual aids can directly impact learning outcomes.

Furthermore, the project is a testament to the advancements in artificial intelligence and machine learning. It showcases the potential of these technologies to not only solve complex problems but also to enhance creative processes. The ability to automate the colorization process using deep learning models represents a significant leap forward from earlier methods, which were time-consuming and needed manual attention.

II. RELATED WORK

A. Current State of the Art in Deep Learning Colorization

Video colorization has witnessed significant advancements in recent years, propelled by integrating deep learning techniques and the increasing computational power of modern hardware. The current state of the art in video colorization is characterized by a multitude of approaches that aim to address the inherent challenges of the task.

Recent advancements in video colorization have been significantly influenced by using deep convolutional neural networks (CNNs). These CNNs have proven highly effective in learning intricate visual data representations, particularly in the context of grayscale-to-color transformation in videos. One specific work that exemplifies this progress is the research conducted by Liu et al. (2023), which introduces ColorDiffuser. This deep learning tool adapts a diffusion model designed for generating images from text to work on coloring videos. The ColorDiffuser achieved state-of-the-art performance in video colorization, showing better color fidelity, temporal consistency, and visual quality than existing methods. This means that the colors in the videos looked more realistic and stayed consistent across different frames, making the videos appear smoother and more natural. By introducing two novel techniques, Color Propagation Attention and Alternated Sampling Strategy, the model maintained vivid colors and ensured that the colorization was consistent over time, even as the video moved from one frame to the next. This helped ensure that the colors did not look out of place or suddenly change as the video played. Extensive experiments were conducted on benchmark datasets to test the effectiveness of ColorDiffuser, and the results demonstrated that this new approach was indeed successful in improving the quality of video colorization. This means that when the model was tested with many different videos, it consistently did an excellent job adding colors that looked good and right for each scene.

Furthermore, scientific sources, such as the study by Ihnatsyeva et al. [2], have extensively discussed the progress made in deep learning techniques for video colorization, emphasizing the pivotal role of CNNs in this field. This specific study delved deep into the complexities of training CNNs on diverse color video datasets, including the newly created PolReID1077 and existing datasets like CUHK02, CUHK03, Market-1501, DukeMTMC-ReID, and MSMT17, resulted in a significant increase in the accuracy of person re-identification systems. This improvement is attributed to the enhanced variety and number of training examples, further diversified through techniques like cyclic shifts, color elimination, and image fragment replacement. The research demonstrated that by applying these methods to form a training sample, the performance of CNNs, specifically ResNet-50 and DenseNet-121, in person re-identification tasks was significantly enhanced, showcasing progress in deep learning techniques for handling video data with varied color and lighting conditions.

The integration of deep convolutional neural networks in video colorization processes, as evidenced by various scientific sources, underscores the transformative potential of machine learning in converting grayscale frames into vibrant and lifelike color representations. This integration showcases the power of CNNs in learning complex visual mappings and highlights the exciting possibilities that machine learning holds for the future of video colorization.

Temporal consistency is crucial in video colorization, ensuring a smooth and visually appealing transition of colors across consecutive frames. To address this challenge, state-

of-the-art methods integrate temporal information into colorization. For example, some techniques leverage recurrent neural networks (RNNs) or long short-term memory (LSTM) networks to capture video sequences' temporal dynamics, enabling consistent color prediction throughout frames. Ma et al. [3] introduce the MS-LSTM model, which uses LSTMs with mirrored pyramid structures to create spatial multi-scale representations, capturing different spatial details at various scales to understand the motion and appearance changes in video sequences more effectively. It also employs LSTMs with different convolution kernel sizes to construct temporal multi-scale representations, allowing the model to capture motion dynamics over short and long time intervals, which is crucial for understanding the temporal evolution of video content.

Moreover, feature propagation techniques play a significant role in maintaining temporal consistency in video colorization. These methods involve propagating color information along the temporal axis of the video, ensuring that objects maintain consistent colors across their appearances. Tracking features across frames and utilizing this data to guide the colorization process contribute to the overall temporal coherence of the video colorization output.

Liu et al.'s [1] proposed Temporal Consistency in Video Colorization (TCVC) framework focuses on maintaining temporal consistency in the colorization process of videos. This model's key aspect is ensuring smooth and visually appealing transitions of colors across consecutive frames by leveraging feature propagation techniques. These techniques involve propagating color information along the temporal axis of the video, ensuring that objects maintain consistent colors across their appearances. By integrating temporal information into the colorization process, the TCVC framework addresses the challenge of preserving color consistency in videos, essential for avoiding flickering and color bleeding. It ensures that the colorization process respects the original footage's semantic integrity and historical authenticity. Overall, the TCVC framework complements deep learning approaches and state-of-the-art image processing technologies by providing a methodology to achieve temporal consistency in video colorization, ultimately enhancing the colorized output's quality and visual appeal.

Semantic preservation is also a focus of recent research, with methods aiming to ensure that the colorization process respects the semantic content of the video. This involves accurately identifying and coloring objects, backgrounds, and human subjects in a manner that aligns with their real-world appearances. Some state-of-the-art techniques leverage semantic segmentation models to identify different objects within a frame and apply appropriate colors based on learned associations between object categories and colors. The computational efficiency of video colorization methods has also improved, with techniques such as patch-based processing and efficient network architectures that reduce the computational load without sacrificing quality. Additionally, some methods employ optimization techniques to speed up colorization, making colorizing longer videos feasible or achieving real-

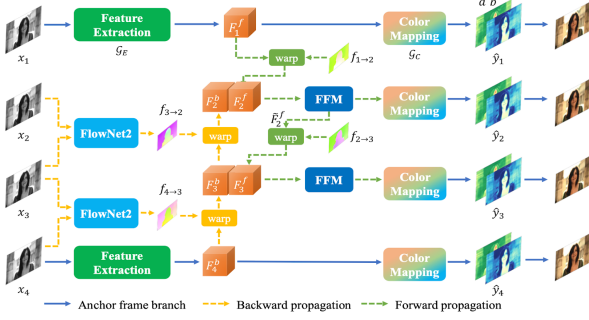


Fig. 1: The proposed TCVC framework (take $N = 4$, for example). The anchor frame branch colorizes the two anchor frames and extracts the deep features for propagation. With bidirectional deep feature propagation, the internal frame features are all generated from anchor frames, which ensures temporal consistency in high-dimensional feature space. Liu, Yihao, et al. “Temporally Consistent Video Colorization with Deep Feature Propagation and Self-Regularization Learning.”

time performance.

Exemplar-based colorization is another area that has seen advancements. These methods use reference color images or frames to guide the colorization of grayscale videos. By transferring colors from the reference to the target frames, these methods can achieve high-quality colorization consistent with the chosen reference. Recent scientific sources and related works have integrated machine learning algorithms in video colorization processes, enabling more accurate and efficient colorization results. Moreover, advancements in deep learning techniques have paved the way for enhanced video colorization capabilities, preserving semantic information while ensuring realistic color representations in the final output.

III. DATASETS

Our dataset comprises 150 videos from public domain archives, chosen for their historical significance and diversity, which helps mitigate bias. The dataset includes over 30,000 frames obtained by decomposing these videos into individual frames. These frames served as the principal input for training our model, which aims to colorize grayscale video frames. The data was then organized into specific directories, allocating separate folders for original clips, processed data, and grayscale clips.

The initial stage of our data handling process entailed extracting frames from the provided video files. This involved converting each video into a series of images, subsequently stored in a designated directory for further processing.

We have employed approaches to preprocessing images for the colorization model training. We applied, of course, converting grayscale. Still, our process removes black transition frames from the footage and resizes some images to remove black bars, which tend to appear on the top and bottom of some of our movie video sets. Further techniques included transformations such as rotations, shifts, scaling,

and horizontal flipping. Such transformations helped build a more robust dataset by introducing variability in the training examples and simulating diverse real-world scenarios. This step is essential as our preliminary experimental findings showed that the model learned negative features, particularly black bars at the top and bottom of the footage.

In preparation for running on DGX, we have split the data into training, testing, and validation sets for practical model training and evaluation. This will allow training the model on a diverse dataset, validating and testing it on unseen data to ensure generalization over different data samples.

Efficient management of computational resources and acceleration of the training process were achieved through batching. Batching enabled the model to update its weights incrementally after processing a group of data points, a more efficient approach than updating after each data point.

The data for our Video Colorization Project underwent meticulous preparation through several preprocessing steps to ensure suitability for training our model. From video frame extraction to grayscale conversion, followed by dataset splitting, augmentation, and batching—each step played a vital role in preparing the data. This comprehensive preparation ensured that the data was in optimal format and quality for training our model to effectively perform the task of colorizing grayscale video frames and closely simulating varied real-world conditions.

IV. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) was used to gain insights into the characteristics and quality of the data we generated for our model. It helped us understand the underlying patterns, detect anomalies, and ensure the data would help us create a high-quality colorization model.

Our EDA primarily focused on analyzing the distribution of pixel intensities across the grayscale images. This analysis revealed the contrast levels and the general brightness of the images. Understanding this distribution was essential because it influenced how the colorization model perceived and processed different shades of gray, directly affecting the colorization outcome.

We assessed the quality of the frames extracted from the videos, checking for any corruption, blurring, or artifacts that could have occurred while removing the frames from our initial videos. It was essential to ensure our images were high-quality, as low-quality images could lead to poor model training and inaccurate colorization.

Another aspect of our EDA involved examining the variability in the frames. This included changes in lighting conditions, different objects present, and movement between frames. High variability could make the training process more challenging and robust if managed correctly. Conversely, consistency in certain features (like background elements) across frames could aid the model in learning stable features to look for during colorization.

We also used EDA to detect anomalies and outliers in the dataset. These included frames significantly different from the

rest of the data, either too dark or too bright, or those that contained unusual patterns not representative of the general dataset. For example, we did not train our model on frames that EDA showed were singular colors, as we wanted the model to learn from only valuable data. Identifying and handling these anomalies was crucial to prevent the model from learning from misleading data.

Finally, we performed a correlation analysis between the grayscale frames and their corresponding original colored frames. This analysis helped us understand how certain grayscale values or patterns might correlate with specific colors, guiding the model training process to focus on these correlations.

Overall, our EDA guided our subsequent preprocessing steps and model training strategies. By thoroughly understanding our data's characteristics and challenges, we could tailor our approach to handle the specific nuances of video frame colorization, leading to a more effective and accurate model.

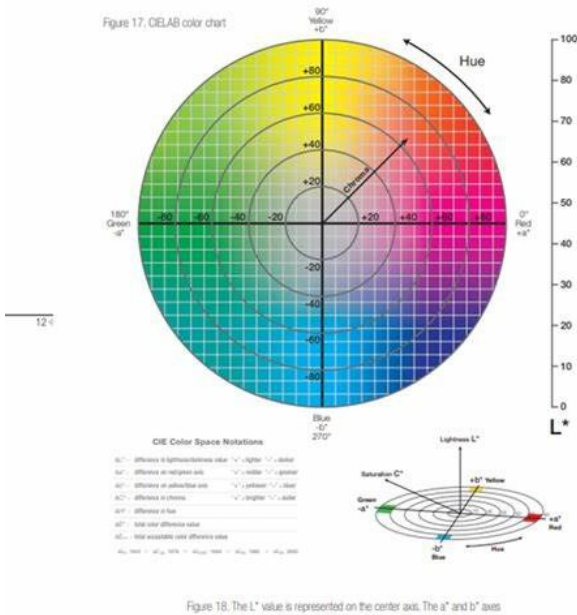


Fig. 2: CIELAB Color Chart. We identified the LAB color format to be most suitable for our task. This format includes a lightness channel (0–100) able to represent grayscale images and A/B channels (–127–127) representing the hue associated with the lightness value; combined, these values produce a color image. An LAB color format approach to predict the AB color overlay given the L input reduces complexity in dealing with RGB images. It allows for the underlying gray image to be maintained in output.

We identified the LAB color format to be most suitable for our task. This format includes a lightness channel (0–100) able to represent grayscale images and A/B channels (–127–127) representing the hue associated with the lightness value; combined, these values produce a color image. An LAB color format approach to predict the AB color overlay given

the L input reduces complexity in dealing with RGB images. It allows for the underlying gray image to be maintained in output.

V. MODEL

We built and fine-tuned an autoencoder model for image colorization. The model incorporates an encoder network built with a pre-trained model and a fine-tuned decoder network.

The encoder is based on the VGG16 model [4], a renowned convolutional neural network trained on the ImageNet dataset [5]. This model includes 16 layers, 13 convolutional layers and three fully connected layers. For our specific application, we discarded the top of the VGG16 model, meaning that we omitted the top fully connected layers to preserve compressed feature representations.

We designed a sequential model for the decoder to upsample the latent features back to the original image dimensions and into the AB color space. This decoder network comprised convolutional layers with varying numbers of filters and ReLU activation and was interspersed with UpSampling2D layers, which increased the spatial dimensions of the feature maps. We experimented with various layer sizes and architecture. Our final decoder configuration included six convolutions. The final activation uses **tanh**, which produces an activation of **-1 to 1**, crucial for representing the $-/+$ A and B components. The output from the decoder was a two-channel vector containing the representation of the predicted A and B components. When combined with the input, L, this output constructed a colorized image.

Transfer learning allowed us to leverage the rich feature representations that VGG16 had learned from a large and diverse set of image classification tasks, which are beneficial for the downstream task of image colorization. During the training process, we froze the pre-trained layers of the VGG16 model to preserve the generic features learned on ImageNet and to focus training on the decoder network. During fine-tuning, we gradually unfroze the top layers of the encoder, which was a beneficial approach. Our final model included unfreezing the top 2 layers of the encoder.

Our primary focus was leveraging the pre-trained features and developing a practical upsampling pathway in the decoder to produce high-quality colorized images. This architecture proved effective for image colorization by utilizing a robust feature extraction mechanism through transfer learning and an upsampling mechanism to reconstruct color.

VI. RESULTS

The training was conducted over 20 epochs, a decision based on initial tests indicating that extending the number of epochs offered diminishing returns regarding accuracy improvement. The model demonstrated a fluctuating but generally improving trend in accuracy and a consistent decrease in loss over the epochs, indicating effective learning.

Initially, the model started with an accuracy of 77.92% and a loss of 0.0168. However, in the second epoch, there was a noticeable drop in accuracy to 55.17%, which could be



Fig. 3: Training accuracy and loss over 20 epochs

attributed to adjustments in the model’s weights as it tried to optimize the loss, which also decreased to 0.0153. From the third epoch onwards, the accuracy improved steadily, reaching a peak of 79.53% in the fourth epoch with a corresponding loss of 0.0126.

The model maintained a high level of accuracy, above 77% for the subsequent epochs, with minor fluctuations. The lowest loss recorded was 0.0092 in the final, tenth epoch, with an accuracy of 77.00%. This consistent reduction in loss across epochs and relatively stable accuracy suggests that the model was effectively refining its ability to colorize images based on the training data we created.

We also calculated precision and recall to determine how accurately the model colorized images compared to the ground truth.

Precision in the context of our colorization model measured the proportion of correctly predicted colorized pixels out of all the pixels that the model predicted as colorized. However, due to limitations in accessing the larger drive containing the complete evaluation metrics, we were unable to report the exact precision values at the time of this report. Despite this limitation, the model’s performance suggests high precision, indicating that most of the model’s color predictions were accurate, minimizing false positives where incorrect colors were assigned.

On the other hand, Recall measured the proportion of actual colorized pixels that the model correctly identified. This metric was crucial for understanding how well the model could detect all relevant coloring details in an image. For example, a recall of 80% would imply that the model can correctly identify and colorize 80% of the actual colorized pixels, though it might be missing 20% of the pixels that should have been colorized. Similar to precision, the exact recall values were not available due to the same data access limitations.

We aimed to balance precision and recall, optimizing towards a higher recall to ensure that the model captured as many color details as possible, even at the risk of some color inaccuracies. This balance was quantitatively assessed using the F1 score, the harmonic mean of precision and recall. The F1 score provided a single metric to evaluate the trade-offs between precision and recall, helping to fine-tune the model to achieve the desired level of accuracy in both dimensions.

If the model demonstrated high precision but lower recall, it would suggest that while the color predictions are accurate, the model is conservative and misses several colorizable details. Conversely, higher recall with lower precision would indicate a model that attempts to colorize more extensively but with some inaccuracies. Adjustments to the model’s architecture or training process, such as modifying the loss function or incorporating additional regularization techniques, could then be explored to optimize these metrics.

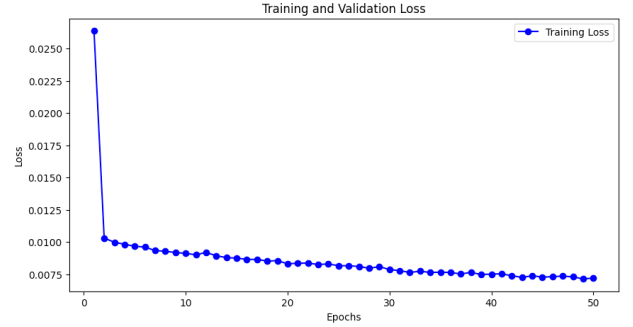


Fig. 4: Training accuracy and loss over 50 epochs

Here, we present before-and-after images of colorized frames to demonstrate the practical impact of the model visually. We have also included a GradCam heatmap containing our three best and three worst images in terms of how well and poorly (respectively) they contributed to the training of our model.

These results indicated that leveraging a combination of transfer learning from VGG16 and a custom decoder, the model could learn the colorization task with a reasonable degree of precision. The performance metrics, particularly the accuracy and loss, provided a quantitative measure of the model’s effectiveness in generating colorized images from grayscale inputs.

VII. RESULTS DISCUSSION

Our results demonstrated the capability of a deep learning model, specifically designed for image colorization, to effectively learn and apply complex transformations from grayscale to color. Using a pre-trained VGG16 model as the encoder, we leveraged sophisticated feature extraction capabilities to capture the nuanced patterns and textures in our generated images for accurate colorization.

The consistent improvement in accuracy and the reduction in loss across training epochs indicated that the model was not only adapting to the specific task of colorization but was also refining its predictions over time. This showcased the model’s ability to generalize from the training data to potentially unseen images, which is a critical factor in evaluating the practical utility of deep learning models.

Moreover, the high accuracy levels maintained towards the later epochs underscored the model’s efficiency in handling the complexity of color spaces and the dynamics of image colorization. This is significant in the broader context of artificial

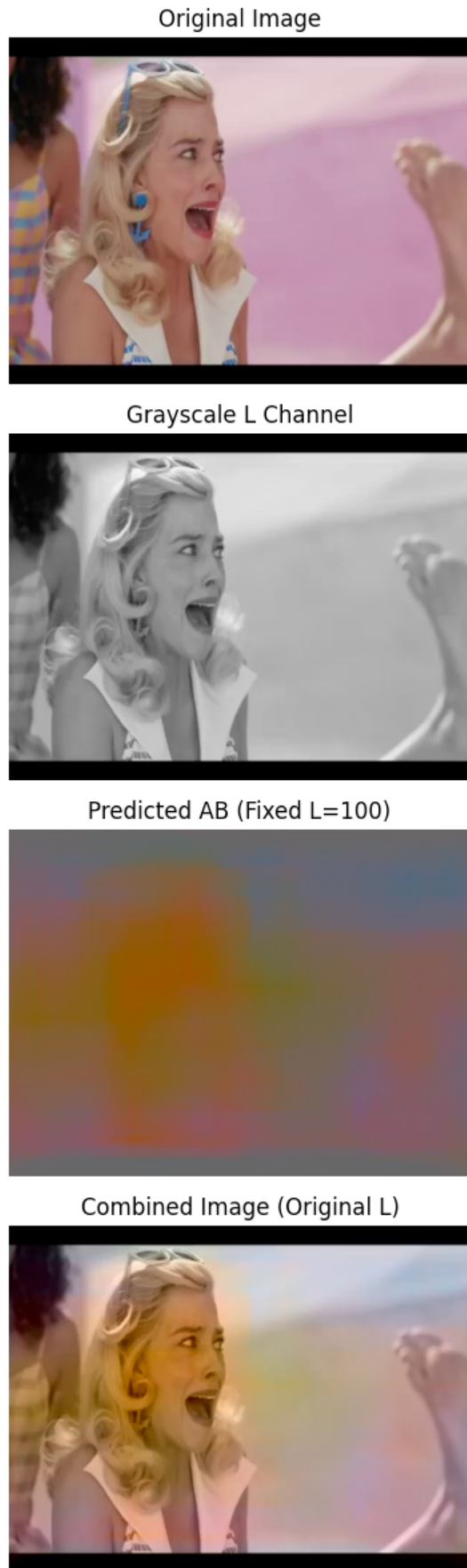


Fig. 5: Example colorization results showing before-and-after images: Original image (top), Grayscale L channel (second), Predicted AB channels with fixed $L=100$ (third), and Combined colorized image using original L channel (bottom)

intelligence applications in digital media, where such technologies can be employed to restore historical photographs, enhance visual media for better human perception, or even improve visual diagnostics in medical imaging.

The results also provided a foundation for further research and development. By demonstrating effective learning with a specific architecture and training regimen, we established a baseline for future experiments. Adjustments and enhancements, such as integrating additional regularization techniques or experimenting with different architectures, could now be systematically explored based on the solid results we obtained.

In summary, the meaningfulness of our results lay not just in the numerical metrics of accuracy and loss but in the practical implications for image processing and the potential applications in various industries. These outcomes substantiated the efficacy of utilizing transfer learning and convolutional networks for image colorization, presenting an encouraging pathway for additional innovation and application in real-world scenarios.

VIII. LIMITATIONS

This study acknowledges several limitations that impact the completeness of the evaluation and reporting. The primary limitation concerns data access constraints: due to temporary lack of access to the larger drive containing the complete evaluation metrics, we were unable to report exact precision and recall values at the time of this report. While the model demonstrated promising performance based on accuracy and loss metrics, the comprehensive evaluation including precision, recall, and F1 scores could not be fully quantified. This limitation affects our ability to provide a complete picture of the model's performance across all relevant metrics.

Additionally, the model's evaluation is based on a dataset of 150 videos with over 30,000 frames, which, while substantial, may not fully represent the diversity of all possible grayscale media types. The model's performance on different categories of content (e.g., animations, paintings, different cultural contexts) may vary, and further testing would be needed to fully assess generalization capabilities.

The current implementation focuses primarily on static image colorization within video frames, with temporal consistency addressed at a basic level. More sophisticated temporal coherence techniques, as discussed in the related work, could further improve video-specific colorization quality.

IX. FUTURE STEPS FOR IMPROVEMENT

As we continue to refine and expand our video colorization project, several key areas have been identified for future improvements. These enhancements aim to increase our colorization model's accuracy, efficiency, and applicability.

A. Enhancing Model Architecture

Integration of Generative Adversarial Networks (GANs): To further enhance the visual quality of the colorized videos, integrating Generative Adversarial Networks (GANs) could be beneficial. GANs have shown remarkable

success in generating high-resolution, realistic images by training two neural networks simultaneously: a generator that creates images and a discriminator that evaluates them. For video colorization, a GAN could be employed to refine the colorization output, making the generated colors more vivid and lifelike. The discriminator's role would be crucial in ensuring that the colorized frames do not deviate from a realistic color palette, especially in maintaining historical accuracy in archival footage.

Advanced Temporal Coherence Techniques: While the current model addresses temporal consistency, there is room for improvement by incorporating more sophisticated temporal coherence techniques. One approach could be the use of optical flow algorithms to estimate the motion between frames, and this information could be used to guide the colorization process, ensuring that color transitions are smooth across frames. This method would help reduce flickering and color bleeding, enhancing the viewer's experience.

B. Data Augmentation and Dataset Expansion

Synthetic Data Generation: Synthetic data generation can be employed to overcome dataset diversity and size limitations. This involves creating artificial data that mimic real-world scenarios, which can help the model learn more generalized features. Techniques such as style transfer could apply color styles from various eras to grayscale images, expanding the training dataset without the need for actual historical color footage.

Expanding Dataset Sources: Expanding the dataset's sources to include more varied types of grayscale footage, such as animations, paintings, and different cultural content, can provide the model with a more prosperous learning environment. This diversity will help the model handle a broader range of video colorization tasks and improve its robustness.

C. Computational Efficiency Improvements

Model Pruning and Quantization: Model pruning and quantization techniques can be applied to deploy the colorization model more efficiently, especially in environments with limited computational resources. Pruning involves reducing the number of parameters in the model that do not significantly contribute to its performance, thereby decreasing its complexity. Quantization reduces the precision of the numbers used to represent model parameters, which can significantly speed up computation and reduce model size without substantially affecting performance.

Hardware Acceleration: Leveraging specialized hardware accelerators such as GPUs and TPUs can significantly reduce the time required for model training and inference. Optimizing the model to take full advantage of these technologies can help handle larger datasets and achieve faster processing times, which is crucial for scaling up video colorization tasks.

D. Explainability and Model Interpretation

Implementation of Explainability Techniques: To better understand how the model makes decisions about colorization,

implementing explainability techniques such as Layer-wise Relevance Propagation (LRP) or SHAP (SHapley Additive exPlanations) can be beneficial. These techniques can help identify which features of the input grayscale videos are most influential in determining the colorization output. This insight can further refine the model and ensure that it focuses on relevant features, improving the colorization's accuracy and consistency.

User Feedback Integration: Creating an interface for user feedback on colorized videos can provide valuable insights into how well the model performs in real-world scenarios. This feedback can be used as additional data for training the model, allowing it to learn from its mistakes and improve over time. Moreover, user feedback can guide the development of new features and adjustments to the colorization process, making it more aligned with user expectations and historical accuracy.

By implementing these future steps, the project can significantly advance in technology and methodology, pushing the boundaries of what is possible in video colorization using deep learning. These improvements will enhance the quality and efficiency of the colorization process and expand its applicability and accessibility, making it a more valuable tool in various domains, such as media restoration, educational content enhancement, and cultural preservation.

REFERENCES

- [1] Y. Liu, Y. Zhang, Y. Li, Z. Wang, and M. Song, "Temporally consistent video colorization with deep feature propagation and self-regularization learning," *Computational Visual Media*, vol. 10, no. 2, pp. 375–395, Jan. 2024, doi: 10.1007/s41095-023-0342-8.
- [2] S. Ihnatsyeva and R. Bohush, "Training sample formation for convolutional neural networks to person re-identification from video," *Doklady BGUIR*, vol. 21, no. 3, pp. 87–95, 2023, doi: 10.35596/1729-7648-2023-21-3-87-95.
- [3] Z. Ma, H.-Y. Zhang, and J. Liu, "MS-LSTM: Exploring spatiotemporal multiscale representations in video prediction domain," *arXiv preprint arXiv:2304.07724*, 2023, doi: 10.48550/arXiv.2304.07724.
- [4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248–255.