UDC 004.415.2
DOI https://doi.org/10.32782/2663-5941/2024.6.2/21

Oleshchenko L.M.
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Tarelkina K.O.
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

USER-DRIVEN VIDEO COLORIZATION SYSTEM DEVELOPMENT USING GENERATIVE ADVERSARIAL NEURAL NETWORK

The article presents a method and software for automated video colorization using deep learning algorithms. The challenge of automated colorization lies in predicting the color channel values for each frame based on input brightness values (L channel in the Lab color model) while ensuring spatio-temporal consistency. The goal of this research is to develop a software system that integrates user input for color palette adjustments and ensures color consistency across frames.

The proposed method implements a two-stage process: colorizing keyframes using a Generative Adversarial Network (GAN) with U-Net-based generator and colorizing intermediate frames using keyframes and previously colorized frames as references. The use of a Generative Adversarial Network with U-Net-based generator is justified by its ability to effectively capture fine-grained details and global context, ensuring high-quality, realistic colorization with improved spatial precision and temporal consistency. To enhance the quality of the results, contextual losses, temporal consistency losses, and smoothness losses are applied. The proposed method was implemented using the Python programming language with TensorFlow, a deep learning framework, for building and training the model.

The developed software integrates user input in the form of color hints for keyframes, enabling the creation of customized color solutions. A distinctive feature of the proposed system is the use of an adaptive approach for determining keyframes based on a threshold SSIM value (0.4). This ensures efficient processing of large volumes of video data while maintaining temporal color consistency.

Experiments demonstrated the high quality of the system's performance, with an average L1 loss of 0.016±0.003 and SSIM of 0.93±0.1 on the training dataset.

Future research focuses on improving temporal consistency losses to achieve real-time efficiency. This would enable the proposed solution to be applied in areas such as filmmaking, the media industry, and the automation of old video restoration processes.

Key words: *software, automated video colorization, deep learning, generative adversarial networks, GAN, contextual losses, spatio-temporal consistency, color palette adjustments.*

Introduction. Problem statement. Automated video colorization is a transformative technology that minimizes the substantial manual effort traditionally required for assigning colors to grayscale videos. This process involves using computational methods to predict color channel values for every frame in a video sequence. Depending on the color model employed, these color channels could be in RGB format or in the chromatic channels (a and b) of the Lab color model. This process must maintain both spatial consistency – ensuring accurate and natural color representation within each individual frame – and temporal consistency, which ensures smooth and coherent color transitions between consecutive frames in a sequence.

The Lab color model, widely adopted in colorization tasks, separates color representation into three distinct components: L, a, and b. The L component represents the lightness or brightness of a color and is directly linked to grayscale intensity. The a channel measures the color range from green to red, while the b channel captures the range from blue to yellow. In video colorization, the grayscale information provided by the L channel serves as the foundation for predicting the values of the a and b channels, effectively reconstructing the color information for each frame. The challenge lies in ensuring that the reconstructed colors are visually natural, spatially consistent within each frame, and temporally consistent across the entire video.

To evaluate the effectiveness of video colorization models, several metrics are commonly utilized. Among these, L1 and L2 losses measure the numerical difference between predicted and actual color values, providing a quantitative assessment of accuracy. Structural similarity index (SSIM) evaluates the

perceptual similarity between the colorized output and the ground truth, capturing how natural the result appears to the human eye. Temporal loss specifically assesses the stability of color transitions across frames, while subjective metrics, such as human evaluation, gauge the overall visual appeal and realism of the generated video.

Automated video colorization methods can be broadly categorized into three groups: user-input-based, example-based, and fully automatic approaches. User-input-based methods rely on human-provided hints, such as scribbles or color strokes, to guide the colorization process. Example-based methods use reference images or videos to transfer colors to the target frames. Fully automatic methods, on the other hand, require no external input, relying entirely on the model to infer appropriate colors. Despite its promise, fully automatic colorization poses a significant challenge due to the inherent multimodality of the task: a single grayscale frame may correspond to multiple plausible colorizations, making it difficult for the model to decide on a single correct output.

The objective of this research is to develop an innovative video colorization system that integrates user input to enable customizable color palettes and enhance control over the final output. By leveraging user-provided guidance, the system strikes a balance between automation and creative flexibility. The proposed system is rigorously evaluated using quantitative metrics, such as L1 loss and SSIM, to ensure high-quality results.

The research explores strategies to improve temporal consistency loss, a crucial aspect for achieving smooth and coherent color transitions across frames. These advancements aim to optimize the system for real-time applications in fields like film production, media post-processing, and the restoration of archival video footage.

Related research. The RGB color model is one of the most commonly used color representations in computer vision. It has a significant limitation: the interdependence of its channels makes it impossible to reconstruct one channel based on the others, leading to challenges in interpretability [1]. On the other hand, the Lab color model provides a distinct advantage by separating luminance (L) from chromatic information. In the Lab model, the L channel represents brightness, while the a and b channels capture color ranges (green-to-red and blue-to-yellow, respectively). This separation simplifies colorization tasks, as predicting two chromatic channels (a and b) based on the luminance channel (L) is computationally more efficient compared to predicting all three RGB channels. Transforming between RGB and Lab involves non-linear operations, which can result in data loss during model training. Similarly, the YUV model also separates chroma and luminance while employing a linear transformation from RGB. Despite this linearity, YUV models can produce unexpected artifacts, such as color stains, reducing their reliability for automated colorization tasks [2]. As a result, the Lab color model remains the preferred choice for tasks requiring accurate and consistent automatic colorization.

Early approaches to video and image colorization heavily relied on user input to guide the process. These initial algorithms used color hints, such as user-drawn scribbles, and propagated the color to adjacent regions based on pixel intensity and texture similarities [3]. Over time, advancements were made to improve the effectiveness of these methods. For instance, research [4] introduced a convolutional neural network (CNN) for automatic colorization. While CNN-based methods generally outperformed non-deep-learning techniques in terms of color accuracy and quality, they occasionally struggled with color bleeding into unrelated regions of a frame [5]. To address this issue, the Hybrid Scribble Propagation algorithm [6] was proposed. This method combines permeability-guided filtering (PGF) and an innovative entropy metric to ensure color is propagated only within the intended regions, preventing color mixing between distinct objects.

Example-based colorization methods, also known as reference-based techniques, use a reference image or video to transfer colors to grayscale frames. For instance, the VCGAN model described in [7] utilizes a generative adversarial network (GAN) to achieve example-based video colorization. The discriminator in VCGAN employs a PatchGAN architecture with fewer parameters to optimize performance, while the generator is based on a U-Net architecture. Two ResNet-50-IN feature extractors are integrated into the generator – one processes the input grayscale frame to extract high-level features, and the other processes the previous frame to ensure temporal consistency. The outputs from these networks are then combined to produce the final colorized frame. The authors implemented a two-stage training process: first, they pre-trained the model on the ImageNet dataset to establish robust initial weights; second, they fine-tuned the model on video datasets to ensure both spatial and temporal coherence.

In another study [8], researchers developed a multi-GAN approach to tackle the complexity of video colorization. This technique divides frames into regions based on pixel intensity, creating two

primary classes: C1 for low-intensity regions and C2 for high-intensity regions. The C2 class is further split into clusters, each processed by a separate GAN. By breaking down the task into smaller, more manageable segments, this approach achieved more accurate results for visually complex images and videos. To maintain visual harmony across frames, the authors proposed a technique where each region "inherits" its primary color from the corresponding region in the previous frame. This strategy minimizes significant color shifts across frames, improving the overall temporal consistency.

Modern GAN-based techniques for video colorization focus on enhancing both frame-level quality and sequence-level consistency. One approach employs a conditional GAN (cGAN) with 3D convolutions to evaluate the realism of individual frames and the continuity of the entire sequence. A new metric, termed Color Consistency, was introduced to measure the stability of colors across consecutive frames. Another innovative framework uses two networks, referred to as "f" and "g," to address the one-to-many nature of video colorization. Network *f* generates multiple plausible colorizations for a given frame, producing four distinct solutions in parallel. Network *g* refines these outputs by taking into account temporal inconsistencies, guided by confidence maps. These confidence maps quantify the degree of temporal mismatch between adjacent frames, with values ranging from 0 (low consistency) to 1 (high consistency). The g network adjusts the colors to reinforce consistency across the video. This network can be applied iteratively during testing to achieve even greater temporal harmony.

By incorporating advanced architectures and innovative loss functions, these methods aim to produce visually realistic and temporally consistent colorized videos, paving the way for practical applications in film restoration, media production, and beyond.

Proposed method. The proposed software method focuses on solving two primary issues in video colorization: preserving temporal consistency and generating visually realistic outputs that align with user expectations. The video sequence is initially divided into individual frames. The algorithm identifies the keyframes, which are essentially the initial frames of each scene. In the current implementation, keyframes are determined by comparing the SSIM values between consecutive frames. When the SSIM value of a frame differs from the previous one by a margin exceeding a predefined threshold (set experimentally at 0.4), that frame is designated as the next keyframe (Fig. 1).

The unique feature of this framework lies in its two-stage process: colorizing keyframes and then filling in the remaining inner frames. Keyframes are processed first using an image colorization network. The inputs for the network include the grayscale lightness channel (L channel) of the keyframe (dimensions H x W x 1), user-provided color hints (ab channels, dimensions H x W x 2), and a binary mask indicating the locations where the user added color suggestions (dimensions H x W x 1). Once a keyframe is colorized, the video colorization

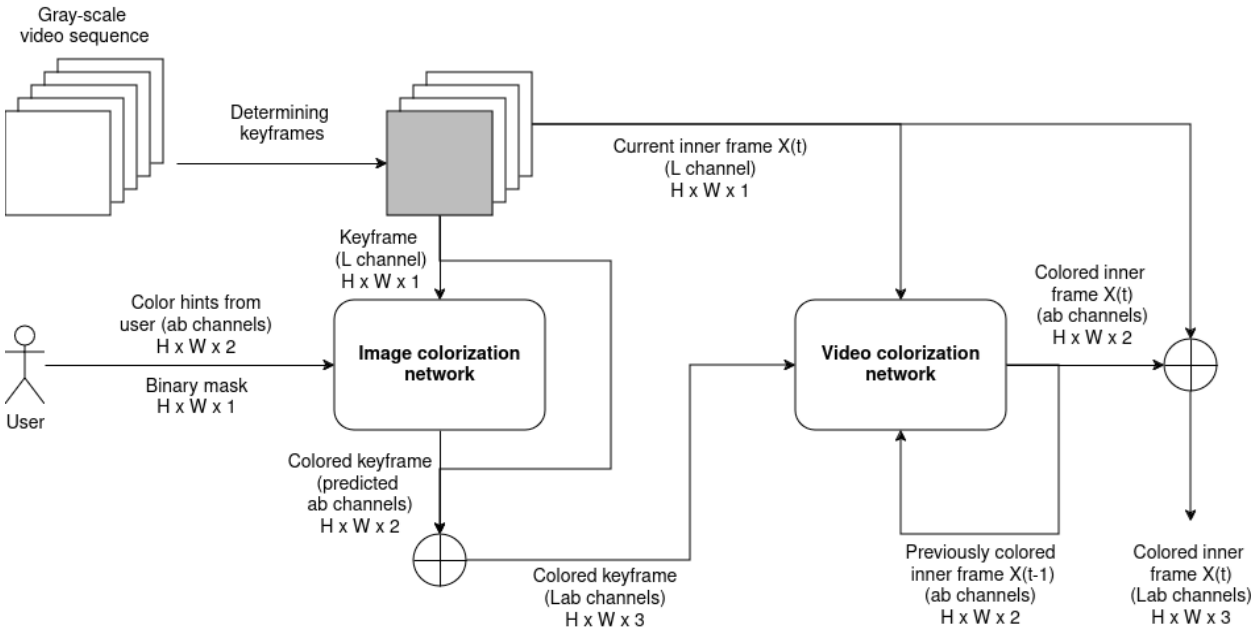

Fig. 1. A schema of the proposed software method

network proceeds to process the inner frames within each scene. For this task, the network relies on two reference frames: the already colorized keyframe and the most recently colorized inner frame. This strategy ensures that both short-term and long-term temporal consistency is maintained across the sequence. The network generates predictions for the ab channels of the current frame, which are then combined with its L channel to reconstruct the complete colorized frame.

The framework allows for flexibility in the choice of network architecture for image and video colorization. In this implementation, a Generative Adversarial Network (GAN) (Fig. 2) with a U-Net-based generator (Fig. 3) is used to achieve high-quality results. Alternative architectures, such as Vision Transformers or more advanced GAN configurations, can also be utilized. The video colorization process can be viewed as reference-based image colorization with two reference inputs.

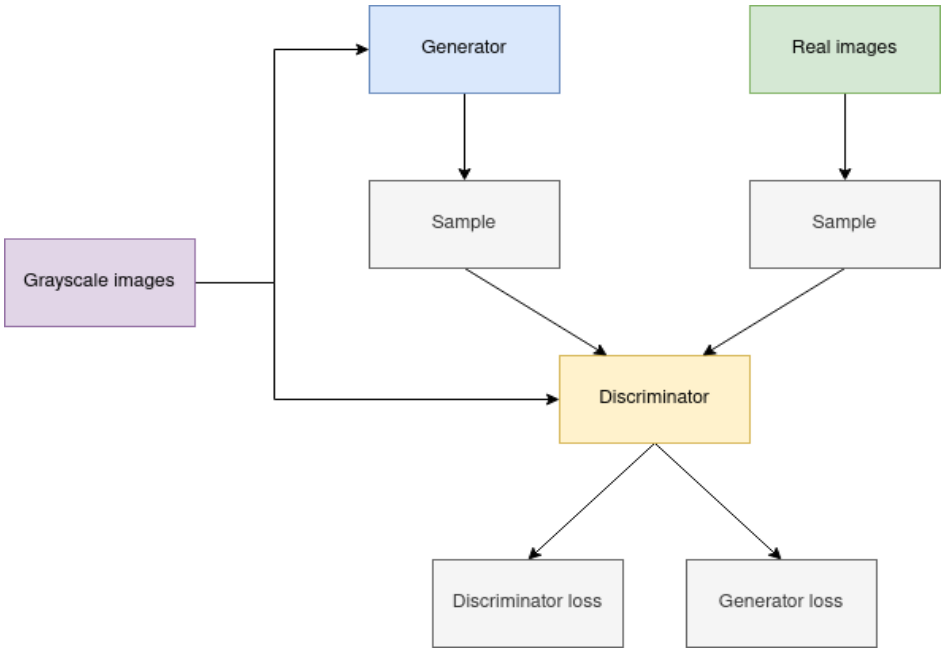

Fig. 2. Conditional GAN. Schema

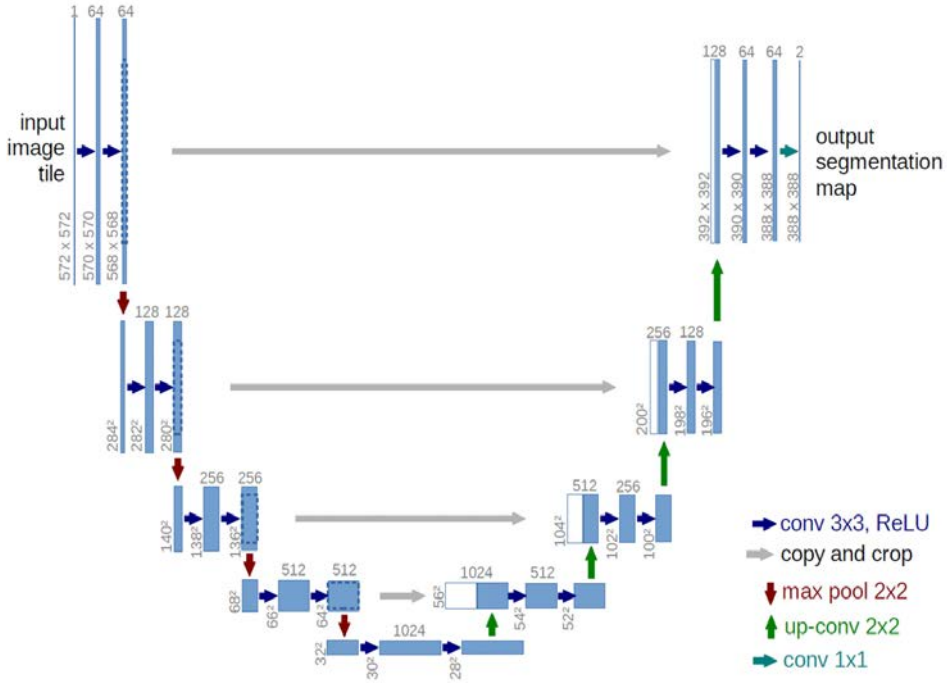

Fig. 3. Generator architecture (U-net). Schema

To enhance the quality of the results, the proposed method incorporates both contextual and temporal consistency losses into the overall loss function for training. Contextual loss is calculated as the mean absolute difference between features extracted from the original and predicted images, specifically using layers 2 to 5 of the VGG19 network. Temporal consistency loss, on the other hand, penalizes discrepancies in color propagation detected through optical flow. A smoothness loss is introduced to promote the spatial coherence of color transitions within each frame.

The developed software is built using a three-tier architecture comprising a web-based client application developed with Angular (TypeScript), a web server powered by FastAPI (Python), and a data layer housing the colorization models implemented in TensorFlow (Python). This design ensures both portability and scalability: the lightweight web application is accessible from most devices, while the web server supports multi-user scalability. The neural networks can be utilized, trained, and tested independently of the web application. A schematic representation of the application architecture is provided in the Fig. 4.

The model integrates a U-Net architecture with attention mechanisms. The U-Net employs an encoder-decoder structure: the encoder progressively reduces the resolution of the input image, extracting contextual features at multiple levels, while the decoder restores the spatial resolution by upsampling the feature maps. Skip connections are used to preserve high-resolution details from the encoder. Alongside the U-Net structure, the model incorporates spatial and channel-wise attention mechanisms. Spatial attention prioritizes regions in the feature maps based on their spatial relevance, while channel attention emphasizes key features by adjusting the importance of individual channels in the feature maps. The inclusion of attention mechanisms allows the network to focus more effectively on color hints and distribute them consistently across image regions.

Research results

The developed video colorization software system incorporates a unique feature that allows users to actively contribute to the colorization process by providing input. Fig. 5 illustrates from left to right: original RGB image, binary mask, color hints, Lchannel with marks applied.

This approach enables the customization of color palettes, making it possible to adapt the colorization to specific user preferences or artistic requirements. By integrating user guidance, the system achieves greater flexibility and control over the final output, ensuring that it meets both technical standards and aesthetic expectations (Fig. 6, 7).

Following the completion of the training process, the video colorization model demonstrates impressive performance. It achieves an L1 loss value of 0.016 with a standard deviation of ±0.003, indicating a high level of accuracy in reconstructing pixel color values. The network attains SSIM score of 0.93 with a deviation of ±0.1, signifying the system's capability to produce visually coherent and detailed outputs. These metrics were calculated using normalized pixel values within the [0, 1] range, reflecting the model's robust ability to handle the nuances of video colorization tasks.

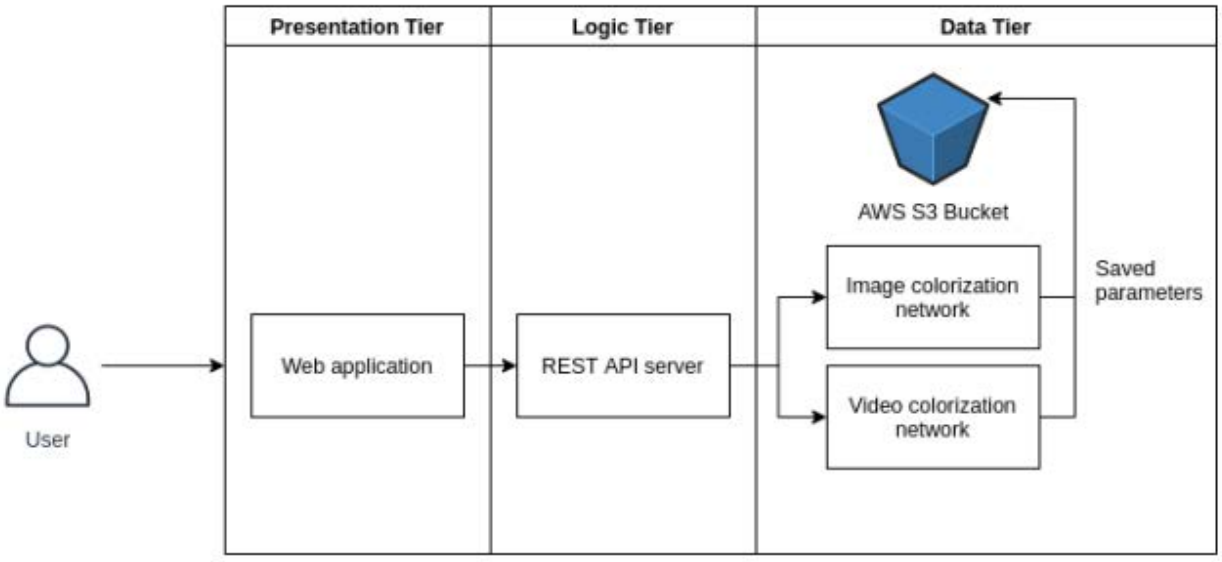

Fig. 4. 3-tier software architecture

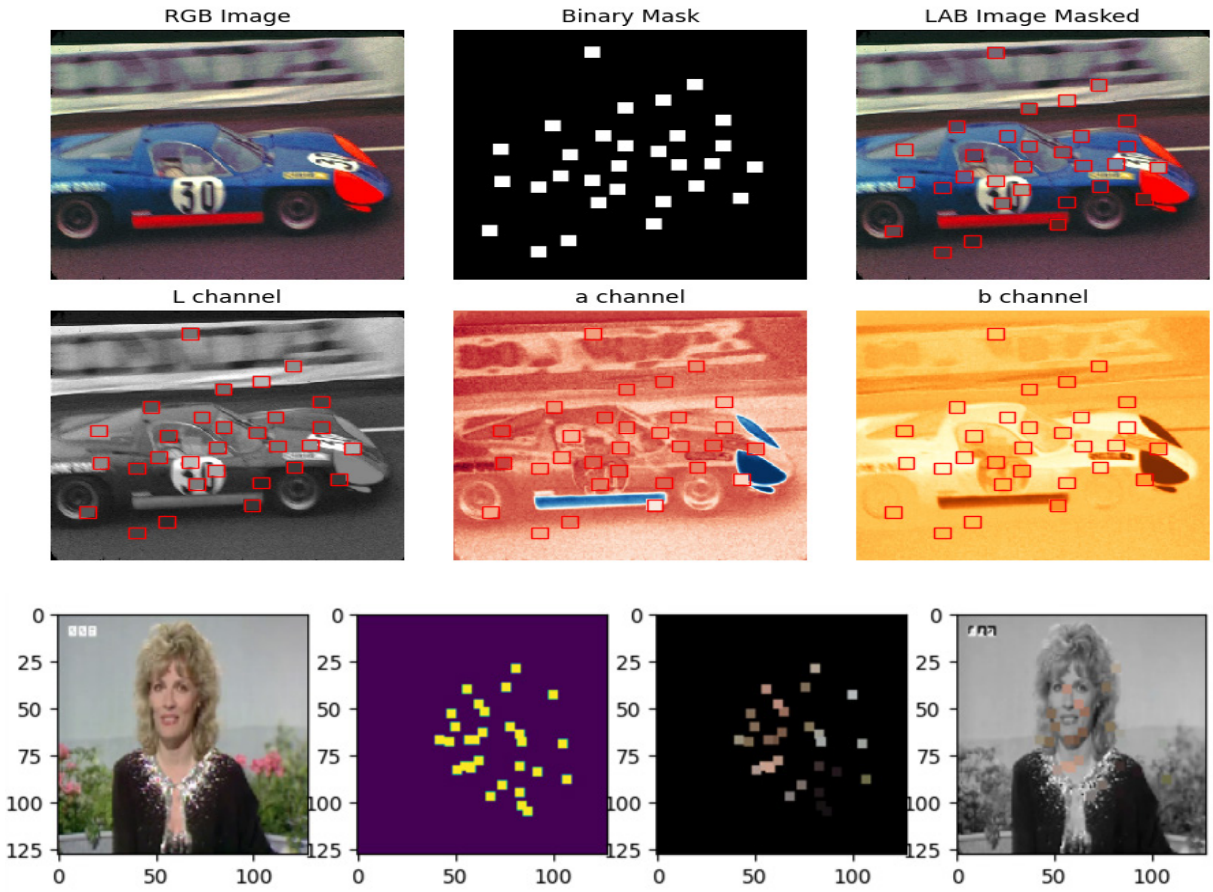

Fig. 5. Simulation of user input on training and testing images (adding color hints)

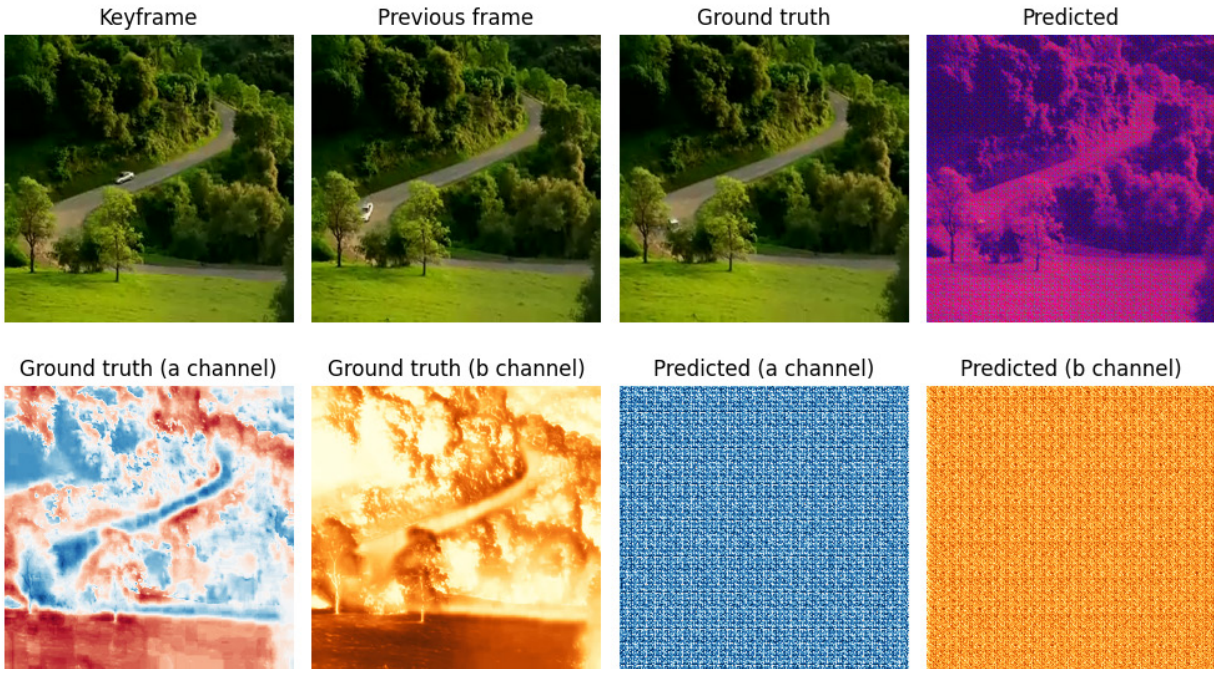

Fig. 6. Values at the beginning of training

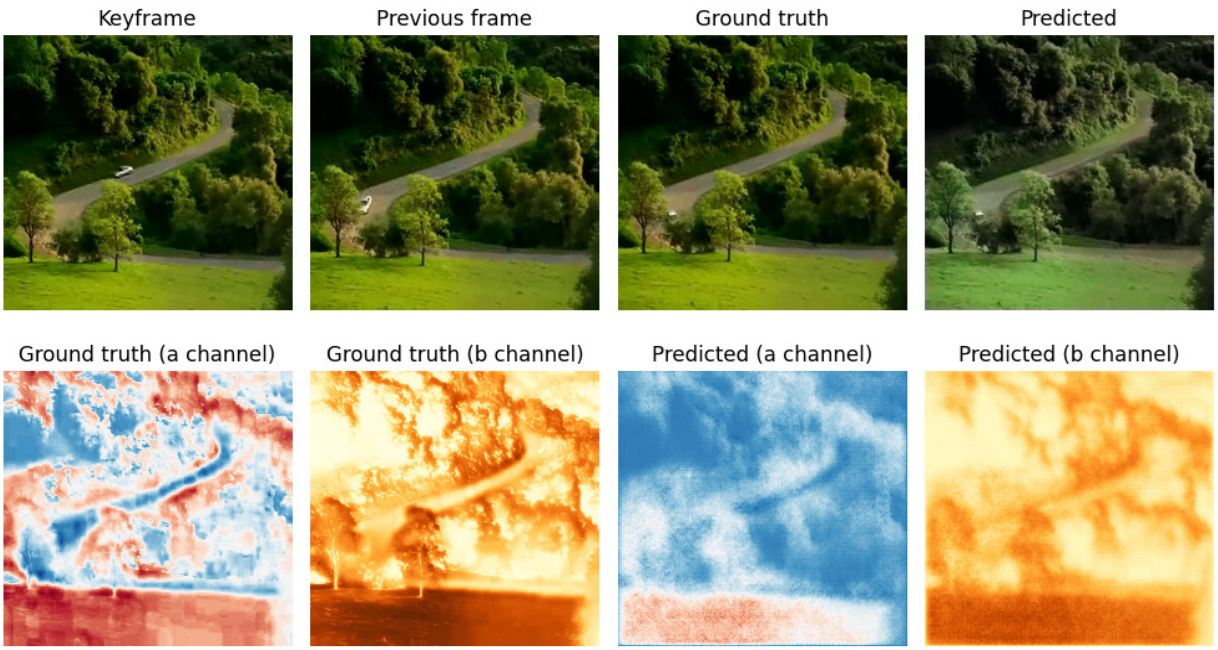

Fig. 7. Values at the end of training

Conclusions and future work. The developed video colorization framework presents a novel solution for enhancing the quality and usability of automated video colorization. By integrating user input through customizable color palettes and leveraging the Lab color model, the method provides a robust balance between automation and manual control. The dual-stage process – keyframe colorization followed by inner frame propagation –effectively addresses the challenges of spatial and temporal consistency, producing visually realistic and coherent results across video sequences. The use of a GAN with a U-Net-based generator, coupled with contextual and temporal consistency losses, ensures high-quality outputs that meet user expectations. The modular design allows for flexibility in adopting alternative architectures, such as transformers or advanced GANs, to further enhance results. The inclusion of smoothness loss and temporal consistency optimization makes the framework well-suited for practical applications in film and media production, where seamless colorization is critical.

Future improvements could focus on refining temporal consistency and incorporating additional user-guided features, such as region-specific coloring or real-time preview capabilities. Expanding the framework to include contextual factors like scene transitions, object tracking, or semantic understanding could enhance its applicability across diverse use cases. The proposed method establishes a strong foundation for integrating AI-driven video colorization into creative workflows, offering a cost-effective and efficient alternative to traditional manual approaches.

Bibliography:

1. Stival L., Pedrini H. Survey on Video Colorization: Concepts, Methods and Applications. *Journal of Signal Processing Systems*. 2023. № 95(6). P. 679–702. DOI: https://doi.org/10.1007/s11265-023-01872-w.

2. Ballester C., Bugea A., Carrillo H., Clément M., Giraud R., Raad L., Vitoria P. Influence of color spaces for deep learning image colorization. *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*. 2023. P. 847-878. DOI: https://doi.org/10.1007/978-3-030-98661-2_125.

3. Levin A., Lischinski D., Weiss Y. Colorization using optimization. *ACM Transactions on Graphics*. 2004. № 23(3). P. 689–694. DOI: https://doi.org/10.1145/1015706.1015780.

4. Zhang R., Zhu J., Isola P., Geng X., Lin A., Yu T., Efros A. A. Real-Time User-Guided Image Colorization with Learned Deep Priors. *ACM Transactions on Graphics (TOG)*. 2017. Vol. 36, Issue 4. P. 1-11. DOI: https://doi.org/10.1145/3072959.3073703.

5. Li S., Liu Q., Yuan H. Overview of Scribbled-Based colorization. *Art And Design Review*. 2018. № 06 (04). P. 169–184. DOI: https://doi.org/10.4236/adr.2018.64017.

6. Dogan P., Aydin T. O., Stefaniski N., Smolic A. Key-Frame based spatiotemporal scribble propagation. *WICED 2015*. 2015. P. 13-20. https://doi.org/10.2312/wiced.20151073.

7. Zhao Y., Po L., Yu W. Y., Rehman Y. A. U., Liu M., Zhang Y., Ou W. VCGAN: Video colorization with Hybrid Generative Adversarial Network. *IEEE Transactions on Multimedia*. 2023. № 25. P. 3017–3032. DOI: https://doi.org/10.1109/tmm.2022.3154600.

8. Jampour M., Zare M. R., Javidi M. Advanced multi-GANs towards near to real image and video colorization. *Journal of Ambient Intelligence and Humanized Computing*. 2022. № 14 (9). P. 12857–12874. DOI: https://doi.org/10.1007/s12652-022-04206-z.

Олещенко Л.М., Тарелкіна К.О. РОЗРОБКА КЕРОВАНОЇ КОРИСТУВАЧЕМ СИСТЕМИ РОЗФАРБОВУВАННЯ ВІДЕО З ВИКОРИСТАННЯМ ГЕНЕРАТИВНОЇ ЗМАГАЛЬНОЇ НЕЙРОННОЇ МЕРЕЖІ

У статті представлено метод і програмне забезпечення для автоматизованої колоризації відео з використанням алгоритмів глибокого навчання. Проблема автоматизованої колоризації полягає у передбаченні значень кольорових каналів для кожного кадру на основі вхідних значень яскравості (канал L в кольоровій моделі Lab) із забезпеченням просторово-часової узгодженості. Метою даного дослідження є розробка програмної системи, яка дозволить інтегрувати користувацький ввід для налаштування кольорових палітр та забезпечувати узгодженість кольорів між кадрами. Запропонований метод реалізує двоетапний процес: колоризацію ключових кадрів за допомогою генеративної змагальної мережі GAN (Generative Adversarial Network) із генератором на основі U-Net і колоризацію внутрішніх кадрів із використанням ключових кадрів та попередньо колоризованих кадрів як посилань. Використання генеративної змагальної мережі із генератором на основі U-Net обґрунтоване її здатністю ефективно захоплювати дрібні деталі та глобальний контекст, забезпечуючи якісну, реалістичну колоризацію з покращеною просторовою точністю та часовою узгодженістю. Для покращення якості результатів застосовуються контекстуальні втрати, втрати узгодженості в часі та втрати згладженості. Запропонований метод було реалізовано за допомогою мови програмування Python із використанням TensorFlow, фреймворку для глибокого навчання, для створення та навчання моделі.

Розроблене програмне забезпечення інтегрує в робочий процес користувацьке введення у вигляді кольорових підказок для ключових кадрів, що дозволяє створювати кастомізовані кольорові рішення. Відмінною рисою запропонованої системи є використання адаптованого підходу до визначення ключових кадрів, що базується на пороговому значенні SSIM (0.4). Це забезпечує ефективну обробку великих обсягів відеоданих, зберігаючи часову послідовність кольорів. Проведені експерименти продемонстрували високу якість роботи системи, зокрема, середнє значення втрат L1 склало 0.016±0.003, а SSIM – 0.93±0.1 на навчальному наборі даних.

Подальші дослідження спрямовані на покращення втрат узгодженості в часі для досягнення ефективності в реальному часі. Це дозволить використовувати запропоноване рішення у таких сферах, як кіновиробництво, медіаіндустрія та автоматизація процесів відновлення старих відеоматеріалів.

Ключові слова: *програмне забезпечення, автоматизована колоризація відео, глибоке навчання, генеративні змагальні мережі, GAN, контекстуальні втрати, просторово-часова узгодженість, налаштування кольорових палітр.*