CAVC: Cosine Attention Video Colorization

Leandro Stival[1][a], Ricardo da Silva Torres[2,3][b] and Helio Pedrini[1][c]
[1]*Institute of Computing, University of Campinas, Av. Albert Einstein 1251, Campinas, SP, 13083-852, Brazil*
[2]*Wageningen Data Competence Center, Wageningen University and Research, Wageningen, The Netherlands*
[3]*Norwegian University of Science and Technology, Larsgårdsvegen 2, 6009 Alesund, Norway*

Keywords: Video Colorization, Deep Learning, Cosine Similarity, Attention Mechanism.

Abstract: Video colorization is a challenging task, demanding deep learning models to employ diverse abstractions for a comprehensive grasp of the task, ultimately yielding high-quality results. Currently, in example-based colorization approaches, the use of attention processes and convolutional layers has proven to be the most effective method to produce good results. Following this way, in this paper we propose Cosine Attention Video Colorization (CAVC), which is an approach that uses a single attention head with shared weights to produce a refinement of the monochromatic frame, as well as the cosine similarity between this sample and the other channels present in the image. This entire process acts as a pre-processing of the data from our autoencoder, which performs a feature fusion with the latent space extracted from the referent frame, as well as with its histogram. This architecture was trained on the DAVIS, UVO and LDV datasets and achieved superior results compared to state-of-the-art models in terms of FID metric in all the datasets.

1 INTRODUCTION

The process of colorization is present since the popularization of analog photographs were originally had produced in a grayscale, having register in the century 14. As the photos, the videos are originally produced without the presence of color in the sequence o frames. So currently due to the advances of the computer vision we can create this information that are missing in the original capture processing. More specifically, the colorization was benefited by the growth of the deep learning techniques, this creating the area denoted Deep Learning Video Colorization (DLVC). This approach used the capacity of models to learn the patterns present in the video frames and a way to produce the missing information, in this case the color.

Currently, in the literature is observed an increase in the number of possible solution, as presented by Stival and Pedrini (2023) in which example-based and fully automated colorization techniques were being demonstrated, as well as the machine learning techniques that are generally used in this process. Another point that underscores the importance of this area of research for the visual computing community can be observed in the New Trends in Image Restoration and Enhancement (NTIRE) challenge (Kang et al., 2023), where solutions for video colorization are proposed.

Nevertheless, despite the growth of research papers in the field of DLVC, several aspects remain open for enhancement, with primary focus on elevating color fidelity within sequences of frames and mitigating instances of color bleed, where object colors exceed their designated boundaries. The objective is to improve the current state-of-the-art results, predominantly exemplified by the NTIRE challenge, while simultaneously devising innovative solutions to address the existing shortcomings in current approaches.

Instead of merely augmenting the complexity of existing models, our emphasis lies in introducing solutions that leverage the problem-solving approach, yielding enhancements that translate into tangible improvements in current outcomes. Hence, we introduce two key innovations. Firstly, we leverage cosine similarity to measure the alignment between each input channel and the output of a singular attention module. Secondly, we employ a shared-weight Transformer block to process individual input channels, marking another innovative stride in our approach. In the following sections, we delve into a more comprehensive

[a] https://orcid.org/0000-0002-3379-6813
[b] https://orcid.org/0000-0001-9772-263X
[c] https://orcid.org/0000-0003-0125-630X

Stival, L., Torres, R. and Pedrini, H.
CAVC: Cosine Attention Video Colorization.
DOI: 10.5220/0012348500003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 385-392
ISBN: 978-989-758-679-8; ISSN: 2184-4321

description of our methodology, elucidating the implementation details of each module.

The main contributions of our work are summarized as follows:

(i) proposition of a Single Channel Attention (SCA) method with shared weights, responsible for producing a better feature space of the gray scale frame and making the colorization of the frames more accurate when compared to the original version;

(ii) exploration of the benefits stemming from the utilization of cosine similarity among the channels within the scaled frame reveals that this novel feature space offers an improved input quality when subjected to processing by the primary colorization network;

(iii) creation of a channel-invariant input methodology introduces a novel approach that can be adapted for use in various computer vision tasks, even when the number of channels is not uniform, addressing a broader spectrum of problems beyond the immediate context;

(iv) validation of the use of color histograms in the colorization process, demonstrating the significant advantages yielded by this straightforward technique within the DLVC process.

Our primary objective in this work was to surpass the benchmarks provided by NTIRE. To achieve this, we leveraged the methodologies outlined in prior research contributions, introducing the Cosine Attention Video Colorization (CAVC) method. In essence, our goals can be summarized as follows: to attain superior results through the amalgamation of the SCA technique, cosine similarity, and the utilization of color histograms extracted from the reference image.

2 RELATED WORK

In this section, we provide an examination of the references that have made significant contributions to our research. We illustrate their relevance to DLVC and illustrate examples of how these references have enriched and complemented our methodology.

2.1 Colorization

When examining the current landscape of DLVC approaches, two primary categories have garnered significant attention: fully-automatic and reference-based methods. In the former, no colorized frames are provided as references during the inference process, while the latter involves the use of colored frames as references.

In this paper, we have chosen to align with the example-based approach, wherein our methodology involves the model's prediction and the subsequent propagation of color throughout the entire video.

2.2 Similarity

The practice of analyzing image similarity or representing images within latent spaces holds significant prevalence in the field of visual computing. This widespread adoption is primarily motivated by the necessity to quantitatively measure images while preserving vital information related to textures and objects.

Prior to the emergence of deep learning, researchers predominantly relied on methods that prioritized pixel-level distances, including metrics like Euclidean distance. Some advanced techniques, such as those demonstrated in the work of Russakoff et al. (2004), utilized Regional Mutual Information (RMI). RMI is an approach that assesses the amalgamation of pixels, image histograms, and the distribution of evidential pixels to quantify image similarity.

Machine learning models deployed for computer vision tasks predominantly rely on architectures that prioritize pattern recognition during their training. This trend has led to a near-complete displacement of techniques that were not specifically tailored for deep learning, underscoring the transformative impact of deep learning approaches in this domain.

This shift in the paradigm of image similarity analysis is conspicuous, as direct examination of image data has given way to its representation in latent spaces. An illustrative case can be found in the research of Lee et al. (2023), which introduced the Structural Embedding Network (SENet). SENet adeptly combines image embeddings with self-similarity information extracted from the frames, exemplifying the evolution in image analysis techniques.

Working within the domain of latent spaces, cosine similarity has emerged as a pivotal component in studies requiring the generation of latent representations for samples. Its application involves comparing these representations, with the aim of preserving the proximity of similar samples while increasing the separation from negative samples within the new space. Cosine similarity is a prevalent technique in computer vision tasks, notably enhancing overall model performance, particularly in scenarios where the latent representations need to discern subtle distinctions in small regions and intricate details within images, as

highlighted in the work developed by Nakagawa et al. (2023).

3 PROPOSED VIDEO COLORIZATION METHOD

Our methodology primarily focused on the creation of a robust latent representation capable of encompassing the requisite information for colorizing frames. This approach has resulted in a potent method that exhibits invariance to the number of input channels and a pronounced ability to accentuate the similarity and intricate details within an individual frame. The colorization process is composed of a monochromatic frame, denoted as $s_g \in \mathbb{R}^{3\times H\times W}$, and a reference frame, denoted as $s_r \in \mathbb{R}^{3\times H\times W}$.

3.1 Model Architecture

Our colorization pipeline, known as CAVC, is constructed from three integral modules: Single Channel Attention (SCA), Color Extraction, and U-Colorization. These modules interact as follows: U-Colorization employs a conventional visual autoencoder architecture, featuring a pretrained Vision Transformer (ViT) to extract color information from the s_r frame.

Our innovation is encapsulated in two pivotal aspects of the architecture. Firstly, the Single Channel Attention (SCA) module plays a crucial role in preprocessing all channels of s_g, thereby elevating the quality of feature representation. Secondly, we introduce the concept of cosine similarity between the inputs s_g and the output of the SSA (Single Channel Attention) module. This information is concatenated and subsequently fed into the autoencoder. The remainder of this section details the interactions between the modules and how they work internally.

3.1.1 Single Channel Attention

Creating a robust feature space for image representation is a pivotal aspect of computer vision, and the same holds true for frame representation. In our pursuit of enhancing existing methods in the literature, we introduce the Single Channel Attention (SCA) module within our pipeline, positioned before the Encoder. SCA serves as a preprocessing step, aiding the Encoder in generating superior feature representations.

The functionality of SCA can be delineated into two primary components. First, it performs an attention process on each channel of s_g, yielding a new feature space with identical dimensions to the input channel, denoted as $channel_{att} \in \mathbb{R}^{1\times H\times W}$. An important aspect of the SCA implementation is its invariance to the number of channels present in the input sample, achieved through the utilization of shared weights. Consequently, SCA can be employed as a texture enhancer in diverse domains where the number of input channels differs from the conventional 3 channels often found in visual computing models.

The second complementary aspect to the attention process in the SCA module is the application of cosine similarity between the current channel being processed and the adjacent channel, yielding $channel_{similarity} \in \mathbb{R}^{1\times H\times W}$. This information proves valuable in accentuating image details. Subsequently, we concatenate $channel_{att}$ with $channel_{similarity}$ to construct a new feature space, denoted as $SCA_{features} \in \mathbb{R}^{64\times H\times W}$. This feature space effectively represents s_g in a more robust manner for the subsequent colorization process. The architecture and information flow of SCA are illustrated in Figure 2.

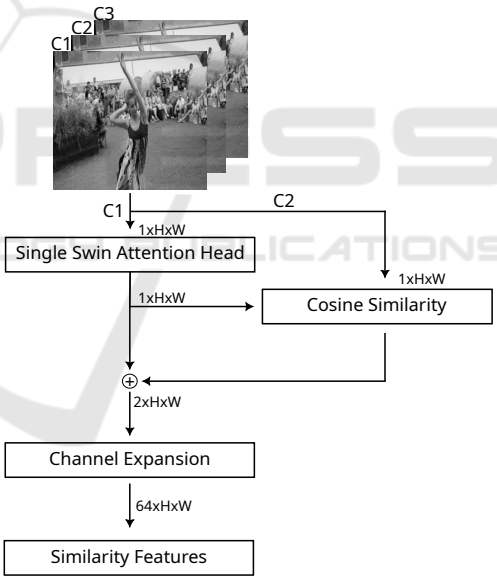

Figure 1: The figure depicts the process of refining input frame characteristics, situated before the U-Colorization encoder. It involves the Single Swin Attention Head ((a Transformer that has only one attention head with weights shared between channels)), which processes each channel of s_g to generate $channel^{att}$. Subsequently, cosine similarity with the next s_g channel computes $channel^{similarity}$, and the two spaces are concatenated, expanding the sample channels by 64, serving as input to the coloring model.

3.1.2 Color Extraction

The quality of the feature space employed to represent the colors within s_r holds paramount importance for

the model's ability to discern the appropriate colors to apply during frame colorization. To optimize this representation's quality, we chose to utilize a pre-trained ViT model, specifically the *vit_b_32* implementation available in the PyTorch library (Falbel, 2023). This ViT model has been trained on the ImageNet dataset, and we extract the output from its final layer in the encoder. Consequently, the reference frame s_r is processed through this pre-trained network, yielding a feature vector denoted as $color_{features} \in \mathbb{R}^{50\times768}$. This representation is crucial to the decoding process within our U-Colorization model.

3.1.3 U-Colorization

The frame coloring process was implemented through an autoencoder approach, comprising four convolutional layers and incorporating Swin attention mechanisms in both the encoder and decoder. In addition, skip connections were established between these two modules. Within this colorization architecture, there are two noteworthy aspects that merit emphasis, as they significantly contribute to the quality of the results.

The first noteworthy aspect involves the fusion of $color_{features}$ with the output of the encoder-generated feature, enhancing the model's comprehension of color application. Additionally, a modification to the original decoder architecture was made by incorporating the color histogram extracted from s_r. This addition led to a marked improvement in color quality, particularly for individual frames within the video.

In essence, U-Colorization entails two key components: the encoder, responsible for creating a feature space that captures essential characteristics from the monochromatic input image s_g, and the decoder, tasked with reconstructing this feature space using inputs such as the encoder's output, $SCA_{features}$, $color_{features}$, and the color histogram derived from s_r. The CAVC pipeline, encompassing these stages, is depicted in Figure 2.

4 EXPERIMENTS

For the evaluation of our method, we opted to employ two metrics featured in the NTIRE challenge. The first metric is the Fréchet Inception Distance (FID), which assesses the quality of color within the frames. The second metric is the Color Distribution Consistency (CDC), which evaluates the effectiveness of color propagation throughout the videos.

4.1 Datasets

To train the weights of CAVC and FCeB, we initially conducted pre-training on the CAVC model using the Densely Annotated Video Segmentation (DAVIS) dataset (Voigtlaender et al., 2019). Following this pre-training phase, we proceeded with fine-tuning on the Large-scale Diverse Video (LDV) dataset (Yang and Timofte, 2021) and an individual process in the Unidentified video objects (UVO) Wang et al. (2021).

The DAVIS dataset encompasses 120 video sequences, segmented into 60 training videos, 30 validation videos, and 30 test videos. Conversely, the LDV dataset comprises 200 training videos and 20 validation videos and UVO for 5,000 videos for training and 250 for test. For the purpose of evaluating our model, we exclusively utilize the DAVIS test set and the LDV validation set at the first experiments and UVO to check the generalization of the model.

4.2 Quantitative Methods

Our architecture underwent evaluation using two prominent techniques that are prevalent in the latest state-of-the-art methods in DLVC and the NTIRE competition. This choice allows for a direct and meaningful comparison of our results with other contemporary approaches. The utilization of both FID and CDC proves to be highly effective, as FID assesses the coloring quality within individual frames, while CDC quantifies color propagation throughout the video. Subsequent sections provide detailed insights into the computation of both FID and CDC for our test sets.

4.2.1 Fréchet Inception Distance

The Fréchet Inception Distance (FID), as introduced by Heusel et al. (2017), is a metric that normalizes its values within the range of 0 to 1. It assesses the level of similarity between images, with a score of 0 signifying complete identity between the compared images.

An advantage of using FID for colorization evaluation is that it does not rely on a pixel-by-pixel comparison. Instead, it measures the similarity between images in the latent space generated by a pre-trained Inception V3 model (Szegedy et al., 2015).

4.2.2 Color Distribution Consistency

While FID serves as a valuable metric for assessing colorization quality, an additional metric is essential to evaluate the consistency of color propagation across video frames. To address this requirement, we

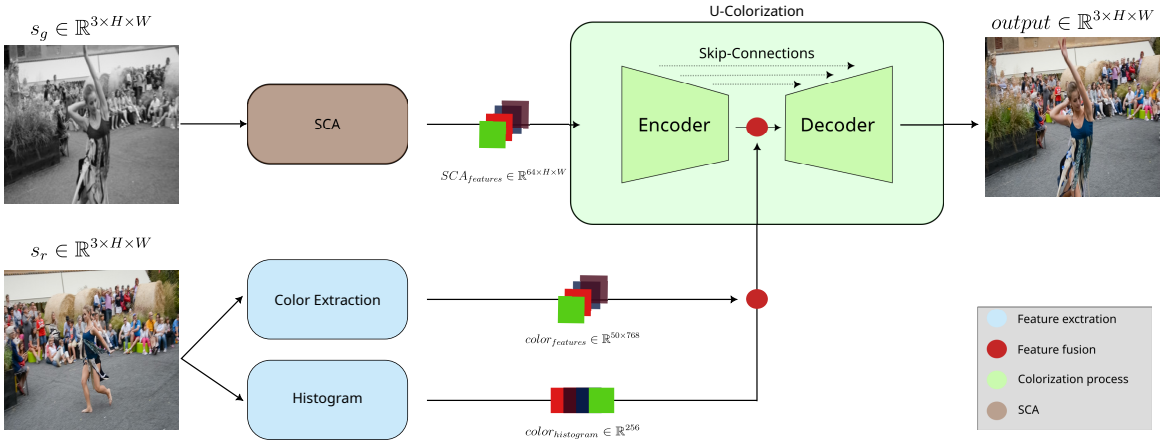

Figure 2: Overview of the complete colorization process, involving two input frames: the initial monochrome frame to be colorized, denoted as s_g, and the reference frame, denoted as s_r. Initially, s_g undergoes processing by the Single Channel Attention (SCA) module, which extracts the $SCA_{features}$ to serve as input for the U-Colorization model. Simultaneously, s_r is subjected to two color extraction modules, generating both the $color_{features}$ and $color_{histogram}$ representations. These two feature spaces are integrated with the output from the U-Colorization encoder. Following decoding, the result is represented by the colored version, denoted as $output$, of the initial s_g frame.

opted for the Color Distribution Consistency (CDC) metric.

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

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

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

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

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

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

4.3 Training

The training of the CAVC modules was initiated with the DAVIS dataset, encompassing 100 epochs, primarily for pre-training purposes due to its smaller size. Subsequently, the LDV dataset was employed for 10 epochs to further refine the model results. Overall, both phases of training were conducted over 300,000 epochs, ensuring comprehensive model training and optimization.

The input images were resized to dimensions of 256×256 to create batches that were well-suited for our infrastructure. For optimization, we selected the classic AdamW optimizer. The training process began with a learning rate of 10^{-4} and was linearly decayed to 10^{-6} over the course of training. All experiments were conducted on a Windows 11 computer with the following hardware specifications: an AMD Ryzen 5600g CPU featuring 12 cores running at 3.90 GHz and equipped with 32 GB of RAM. Additionally, the system was equipped with an NVIDIA GeForce GTX 1080 Ti GPU, which boasted 11,000 MB of GDDR5 memory (NVIDIA et al., 2020).

5 RESULTS

We evaluated the quality of our model using the test set from the DAVIS dataset and the validation set from

LDV. This selection of datasets for evaluation was made to ensure that our results remained comparable to those presented in the NTIRE competition. The quantitative results achieved in DAVIS dataset (test set) and LDV dataset (validation set) are presented in Table 1. The qualitative results obtained with the proposed method in the DAVIS dataset are shown in Figure 3.

Table 1: Results of our implementation on both the DAVIS test set and the LDV validation dataset. These results are compared with state-of-the-art methods from the literature and the NTIRE challenge. The methods considered state of the art include BiSTNet (Zhang et al., 2019), ColorVid (Wan et al., 2020), and DeOldify (Salmona et al., 2022).

	DAVIS		LDV Validation	
Method	**FID↓**	**CDC↓**	**FID↓**	**CDC↓**
BiSTNet	$4.5e^{-4}$	$5.9e^{-3}$	$7.2e^{-4}$	$3.3e^{-3}$
ColorVid	$3.7e^{-4}$	$\mathbf{4.6}e^{-3}$	$2.6e^{-4}$	$2.7e^{-3}$
DeOldify	$6.4e^{-4}$	$5.2e^{-3}$	$4.9e^{-4}$	$3.8e^{-3}$
Ours	$\mathbf{3.0}e^{-4}$	$5.0e^{-3}$	$\mathbf{2.4}e^{-4}$	$\mathbf{1.9}e^{-3}$

The Unidentified Video Objects (UVO) has used to measure the capacity of the model to learn colorization in large datasets with very different scenes and huge diversity of objects and colors. The training is realized with 10 epochs in 5,000 videos and the test has used 250 to evaluate. The results are presented in the Table 2 and the qualitative results in Figure 4.

Table 2: Quantitative analysis comparing model results in the UVO dataset was performed using our approach, BisTnet, and ColorVid. The use of DeOldfy for testing was limited.

	UVO	
Method	**FID↓**	**CDC↓**
ColorVid	$\mathbf{1.15}e^{-4}$	$6.74e^{-3}$
BisTNet	$1.97e^{-4}$	$1.08e^{-2}$
Ours	$1.69e^{-4}$	$\mathbf{6.04}e^{-3}$

6 ABLATION

During the development of this work, we identified various methods that could potentially enhance the colorization process. In this section, we present the alternative methods that were tested in lieu of SCA and showcase the comparative improvements achieved by our approach during the training phase on the DAVIS dataset.

6.1 Cosine Similarity Impact

Initially, the SCA implementation did not incorporate cosine similarity between the s_g and $channel_{att}$ channels. In practice, this omission resulted in the colorization process yielding inferior results compared to the current ones.

The quantitative values for the evaluation metrics achieved in the DAVIS dataset from this analysis are presented in Table 3.

Table 3: Quantitative analysis that compares the effectiveness of incorporating cosine similarity between the $channel_{att}$ and the adjacent channel in s_g.

	DAVIS	
Method	**FID↓**	**CDC↓**
With Similarity	$4.35e^{-3}$	$5.42e^{-2}$
Without Similarity	$6.56e^{-3}$	$5.73e^{-2}$

Examining the activation levels of the layers, we can observe that the feature production effectively preserves long-range textures and object shapes when cosine similarity is applied between the output of $channel_{att}$ and the adjacent channel in s_g, as depicted in Figure 5.

6.2 Color Extraction Impact

Throughout the course of this work, the representation of colors present in s_r was also thoroughly examined and evaluated. Initially, we opted to utilize a pre-trained VGG-19 model, but the results did not meet our expectations. As reported in Table 4, the colorization results exhibited a substantial improvement when we transitioned to a more robust and recent model, specifically a ViT.

Table 4: Quantitative analysis that compares the model results when the color information from s_r is extracted using either VGG or ViT.

	DAVIS	
Method	**FID↓**	**CDC↓**
VGG	$7.5e^{-3}$	$5.35e^{-2}$
VIT	$5.69e^{-3}$	$4.75e^{-2}$

6.3 Color Histogram

The decision to include a histogram of colors present in s_r occurred due to an issue encountered during implementation, where some samples exhibited a low presence of colors in frames. Consequently, the inclusion of the histogram proved to be a valuable guide

Ground Truth | s_g | R | Deoldify (Salmona et al., 2022) | ColorVid (Wan et al., 2020) | BistNet (Zhang et al., 2019) | Ours

Figure 3: Comparison of the results obtained on the DAVIS dataset with our CAVC model and state-of-the-art methods reveals that our model achieves outcomes that are nearly identical to the original colorization, with colorization that appears more natural than the current state of the art. The examples presented in this comparison depict the 20th frame of each video.

Ground Truth | s_g | R | ColorVid (Wan et al., 2020) | BistNet (Zhang et al., 2019) | Ours

Figure 4: Results obtained with our CAVC model in the UVO dataset compared with the BisTNet and ColorVid.

for distributing colors during the coloring process, as illustrated by the results in Table 5.

Table 5: Quantitative analysis that compares the impact of using color histogram of s_r in the decoder of the colorization process.

	DAVIS	
Method	**FID↓**	**CDC↓**
With Histogram	$3.30e^{-4}$	$5.10e^{-3}$
Without Histogram	$4.73e^{-4}$	$5.31e^{-3}$

7 CONCLUSIONS

The application of deep learning techniques has proven to be the optimal approach for video colorization, as exemplified by the NTIRE competition's adoption of this method. This work incorporates traditional image processing methods, encompassing histograms, texture refinement, and filtering, to enhance existing techniques and illustrate advancements in state-of-the-art results. The addition of cosine similarity and histograms has notably enhanced the quality of frame colorization results.

Additionally, we found that the texture refinement process in the monochrome frame is effective when combined with single-head attention as preprocessing of the input channels. Our findings suggest that this approach can be applied for enhancing image colorization in future research.

ACKNOWLEDGEMENTS

The authors would like to thank FAPESP (#2022/12294-8 and #2023/11556-1) for the support.

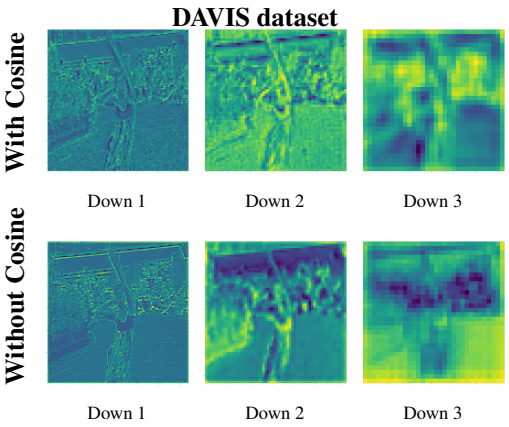

Figure 5: The qualitative results provide a visual representation of how object information and texture are propagated throughout the encoder network were evaluated on the DAVIS dataset. In the first row, we see the complete model architecture, while the second row represents the version without cosine similarity. It is evident that details can be effectively propagated through feature extraction when cosine similarity is employed, reinforcing why results with cosine similarity exhibit superior outcomes.

REFERENCES

Falbel, D. (2023). *torchvision: Models, Datasets and Transformations for Images*. https://github.com/mlverse/torchvision.

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. *Advances in Neural Information Processing Systems*, 30.

Kang, X., Lin, X., Zhang, K., Hui, Z., Xiang, W., He, J.-Y., Li, X., Ren, P., Xie, X., Timofte, R., Yang, Y., Pan, J., Peng, Z., Zhang, Q., Dong, J., Tang, J., Li, J., Lin, C., Li, Q., Liang, Q., Gang, R., Liu, X., Feng, S., Liu, S., Wang, H., Feng, C., Bai, F., Zhang, Y., Shao, G., Wang, X., Lei, L., Chen, S., Zhang, Y., Xu, H., Liu, Z., Zhang, Z., Luo, Y., and Zuo, Z. (2023). NTIRE 2023 Video Colorization Challenge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1570–1581.

Lee, S., Lee, S., Seong, H., and Kim, E. (2023). Revisiting Self-Similarity: Structural Embedding for Image Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23412–23421.

Nakagawa, T., Sanada, Y., Waida, H., Zhang, Y., Wada, Y., Takanashi, K., Yamada, T., and Kanamori, T. (2023). Denoising Cosine Similarity: A Theory-Driven Approach for Efficient Representation Learning. *arXiv preprint arXiv:2304.09552*.

NVIDIA, Vingelmann, P., and Fitzek, F. H. (2020). CUDA, release: 10.2.89. https://developer.nvidia.com/cuda-toolkit.

Russakoff, D. B., Tomasi, C., Rohlfing, T., and Maurer, C. R. (2004). Image Similarity using Mutual Information of Regions. In *8th European Conference on Computer Vision*, pages 596–607. Springer.

Salmona, A., Bouza, L., and Delon, J. (2022). DeOldify: A Review and Implementation of an Automatic Colorization Method. *Image Processing On Line*, 12:347–368.

Stival, L. and Pedrini, H. (2023). Survey on Video Colorization: Concepts, Methods and Applications. *Journal of Signal Processing Systems*, pages 1–24.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper with Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., and Chen, L.-C. (2019). Feelvos: Fast End-to-End Embedding Learning for Video Object Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490.

Wan, Z., Zhang, B., Chen, D., Zhang, P., Chen, D., Liao, J., and Wen, F. (2020). Bringing Old Photos Back to Life. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2747–2757.

Wang, W., Feiszli, M., Wang, H., and Tran, D. (2021). Unidentified Video Objects: A Benchmark for Dense, Open-World Segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 10776–10785.

Yang, R. and Timofte, R. (2021). NTIRE 2021 Challenge on Quality Enhancement of Compressed Video: Dataset and Study. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 667–676.

Zhang, B., He, M., Liao, J., Sander, P. V., Yuan, L., Bermak, A., and Chen, D. (2019). Deep Exemplar-based Video Colorization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8052–8061.