Available online at www.garmian.edu.krd

Passer Journal of Basic and Applied Sciences

Passer 7 (Issue 1) (2025) 276-285

Original Article

Dynamic Reference Image Selection for Seamless Video Colorization Across Scenes

Zahoor M. Aydam *, [1], Nidhal K. El Abbadi [2]

[1] *Computer Science department, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq.*
[2] *Al-Mustaqbal Center for AI Applications, Al-Mustaqbal University, Babylon, Iraq*

Received 08 January 2025; revised 06 May 2025;
accepted 09 February 2025; available online 29 May 2025

DOI: 10.24271/PSR.2025.449298.1543

ABSTRACT

With the rapid advancement of information technology and the widespread availability of high-resolution videos, the demand for colorizing black-and-white footage from previous eras has increased significantly. However, using a single reference image to colorize a video is a challenging task due to the variations in objects, colors, textures, and lighting across different video frames. This paper proposes a novel method to address these challenges by selecting multiple reference images based on the distinct scenes within a video. The proposed approach consists of two key components. First, a large set of reference images is generated, and their corresponding features are extracted using a custom network that combines ResNet50 with Generalized Mean Pooling (GeM). Second, the method operates in two phases: (1) automatically counting the number of different scenes in the video by calculating the Structural Similarity Index (SSI) between frames, followed by computing the mean and standard deviation to establish a threshold for scene differentiation; and (2) selecting the most suitable reference image for each scene by comparing the features of the scene, extracted via ResNet50-GeM, with those of the reference images. The dot product is used to evaluate feature similarity, with the highest-ranked reference image selected for each scene. The proposed method achieved an accuracy of 99% in identifying scene changes and selecting appropriate reference images, outperforming existing techniques in both accuracy and efficiency.

Keywords: ResNet50-Gem, gray-level videos, image processing, image features.

1 Introduction

The demand for converting legacy black-and-white videos into modern, pleasant-colored videos has grown rapidly in the recent decade. This process is important due to its wide applications across various fields [1]. The aim of colorizing monochrome footage is to produce realistic, colorized versions of videos while preserving temporal consistency across frames [2]. Traditional video colorization techniques select a frame to use as a reference image for colorizing the subsequent frames [3]. Despite, the effectiveness of these techniques, but still struggle with complex videos including multiple scenes, where each with various contents and color palettes. Selecting an optimal reference image has a significant impact on the colorized video quality, it is a critical challenge [4].

Most of the current techniques rely on selecting the reference images randomly, and this may lead to producing unpleasant videos, with inconsistencies in colors, and visual artifacts [5]. For that, improving the techniques to select the optimal reference image(s) is essential to produce high-quality colored video.

Recently, the advancement in machine learning provided a promising solution to improve the selection of reference images. Techniques based on machine learning leveraging the dynamic scene, color distribution, and information about the various objects' motion to select reference images that can enhance color consistency across varied scenes. However, most current techniques still use a single reference image for all video frames, ignoring the scene diversity across video frames [6&7]. The current study suggests a new framework to address this gap by fusing advanced feature extraction methods and machine learning techniques to enhance the selection of the best reference image for each video scene.

The remainder of this paper is organized as follows: Section Two introduces a brief explanation of some of the related work in the field. Section Three outlines the theoretical background necessary for understanding the proposed approach. Section Four describes the performance metrics used to evaluate our methodology. Section Five provides a detailed description of the methodology. Results and an in-depth discussion are presented in Section Six, followed by conclusions and future directions in Section Seven.

* Corresponding author
E-mail address zahoorm.alfraih@student.uokufa.edu.iq (Instructor).
Peer-reviewed under the responsibility of the University of Garmian.

2 Related Works

Articles specifically addressing the selection of reference images for video colorization are scarce. Most papers that involve auto-selection of reference images do not thoroughly explain the methods used for this purpose, often mentioning it only in the context of video colorization. We were unable to find any articles dedicated to explaining how to choose the best reference image for video colorization.

Vijetha, U. & Geetha, V. suggested approach an image colorization based on dividing the image into superpixels and combining it with an automated reference image selection mechanism is proposed. This methodology aims to overcome the limitations associated with traditional image colorization methods, which often rely on manual selection of reference images. The method is based on a hybrid texture matching model that enables the automatic selection of the most appropriate reference image based on the characteristics of the image to be colorized. This integration has contributed to improved color transfer quality, resulting in more realistic and attractive colored images, while reducing human intervention in the colorization process [8].

Jinliang Yao, et al. The suggested approach overcomes eliminates difficulties present in conventional content-based image retrieval because it removes requirements for complex hard sample mining operations during training. The GS model unites two components to enable training at every stage given only image-level annotations. It merges Generalized-Mean (GeM) pooling with a differentiable smoothed average precision (AP) approximation. The improved version GSA extends retrieval performance through combined optimization of two loss functions to advance feature discovery. The Revisited Oxford and Paris datasets display experimental evidence showing that the proposed models deliver strong performance along with effectiveness [9].

Gherbi, T et. al. Suggest established a private method requiring users to select color-balanced reference images for diagnosis precision. Researchers gain full access to staining patterns through this particular approach. The method employs two approaches known as global and unsupervised normalization to achieve standardized color and intensity diversity standards. The performance evaluation of colorization showed SSIM reaching 0.92 as one of the measured metrics. Reference image selection for particular classes provides enhanced image quality by replacing random selection. The reference image selection process provides limited effectiveness because it depends on dataset diversity especially when observing staining differences between images [10].

Ali, O. & Ibrahim, S. The proposed approach uses a two-step methodology to enhance Information Retrieval effectiveness through optimized term weight optimization. An evolutionary gradient method develops Global Term Weights in the first step for enhancing term discrimination across all documents within the collection. The evaluation of term contributions in individual documents utilizes the Local Term Weight of Term Frequency-Average Term Occurrence (TF-ATO) in the second stage of the process. By implementing this method organization can minimize very complex weight optimization algorithms while achieving acceptable retrieval results. The experimental outcome validates that the proposed method achieves better performance than Okapi BM25 and TF-ATO with DA weighting through its MAP, AP and NDCG metrics. Its performance changes based on how complete the relevance judgments are when training takes place [11].

Panghurian, F et. al. Suggest approach brings forward a deep learning technique for mapping oil palm plantations across Indonesia to solve the problems within traditional manual monitoring systems that require extensive work hours as well as produce unreliable results. Observers can utilize Sentinel satellite imagery at medium resolution because it provides both ample spatial clarity and extensive availability. The research uses a UNet-based semantic segmentation model that includes ResNet-50 as backbone networks for more effective segmentation. The specific network connections facilitate better feature recognition and improved detection of plantation areas. The method yielded superior results because ResNet-50 obtained F1 scores of 0.94 together with an inference accuracy of 0.92 indicating deep learning combined with remote sensing produces effective mapping capabilities [12].

Y. Yang et al. Suggest approach BiSTNet an automated reference image selection to output both accurate and temporally continuous colorized frames during the video colorization process. The process seeks to perform forward and backward colorization using two reference images that are automatically identified per frame. The forward stage starts by employing the first reference image to continuously colorize video frames. The color propagation process in the video timeline is tracked properly by the Bidirectional Temporal Feature Fusion technique. The Mixed Expert Block consists of methods that resolve challenges by fighting color spreading while adding semantic prior knowledge and preventing inconsistent frame colors [13].

3 Theoretical Background

Theoretical Background encompasses a variety of algorithms, that together contribute to the achievement of the desired objective. These algorithms include:

3. 1 Residual Networks with 50 Layers

ResNet is a prominent CNN model proposed by He et al., designed to address the vanishing gradient issue common in deep learning architectures. In ResNet-50, the numeral '50' represents the total number of convolutional layers in the network. A key innovation that improved ResNet's performance was the introduction of shortcut connections, enabling the skipping of layers. These connections effectively transform traditional networks into residual networks, facilitating the efficient training of deeper architectures. ResNet-50 features residual blocks with shortcut connections that forward inputs to subsequent layers without modification. This design helps mitigate the vanishing gradient problem, ensuring stable training in deeper networks. Another major enhancement in ResNet-50 is the use of a bottleneck architecture within its residual blocks. Unlike standard blocks with two layers, the bottleneck structure employs three layers, utilizing 1×1 convolutions to reduce parameters and

computational complexity, thereby speeding up training. Residual connections also play a critical role in avoiding bottlenecks and enhancing performance as network depth increases, enabling ResNet-50 to outperform other networks in image recognition tasks. Additional components, such as pooling layers, batch normalization, and fully connected layers, further contribute to its efficiency. Pooling layers reduce spatial dimensions to lower computational complexity while preserving essential features. Batch normalization normalizes inputs across mini-batches, accelerating training and improving convergence. Finally, the fully connected layers handle classification tasks using higher-level features extracted from earlier layers. Together, these architectural elements make ResNet-50 highly effective for image classification and related tasks [4]. Figure 1 illustrates the architecture, highlighting its key components and their integration to achieve this performance.

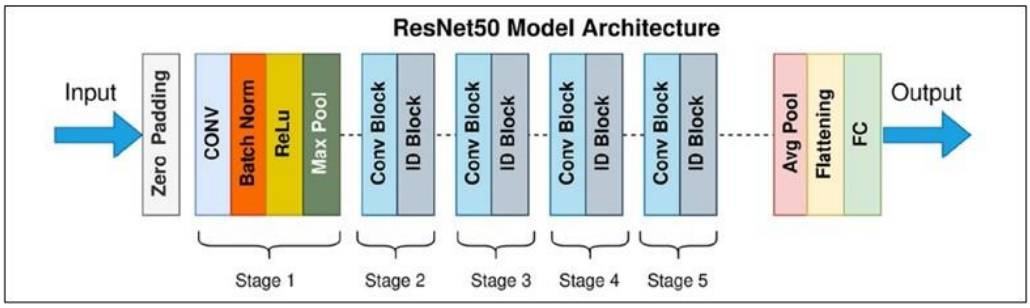

Figure 1: ResNet50 Architectures [5].

3. 2 Generalized Mean Pooling

Generalized Mean Pooling (Gem) is a flexible pooling method that extends traditional pooling techniques by exploiting the concept of the generalized mean. The pooling parameter p can be either trainable or fixed, and it is shared across all channels in practice to avoid distraction during training. The pooling operation is formalized as:

$$D = \frac{1}{H*W}\sum_{(i,j)\in(H,W)} f_p(i,j,c))^{\frac{1}{p}},\ \ c \in \{1,2\cdots C\} \quad (1)$$

when p =1 it becomes average pooling, and when $p \to \infty$, it becomes max pooling. Given an input image $I \in Rh\times w\times 3$, the output feature map $f = \theta(I) \in \mathbb{R}$H×W ×C is processed through Gem, producing a descriptor $D \in \mathbb{R}$C×1 [6].
Larger values of p result in more localized feature responses, we proposed to use value p = 3.0 in the current work. Using p =3 offers a balance between average and max pooling, providing flexibility to emphasize important features while still maintaining some contribution from the entire feature map.

4 Evaluation Metrics

To measure the effectiveness of image retrieval systems, several metrics are commonly used:

4. 1 Precision

Measures the ratio of corrected cases that are predicted to all the prediction cases [7].

$$\mathrm{P} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FP}) \quad (2)$$

Where TP is a true positive and FP is a false positive.

4. 2 Recall

Indicates the ratio of corrected cases that are predicted to all the positive prediction cases [14].

$$\mathrm{R} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \quad (3)$$

where FN is false negatives.

4. 3 Mean Average Precision (MAP)

Evaluates the ranking quality of retrieved images [15].

$$\mathrm{MAP} = \Sigma(\text{Average Precision(query)}) / \text{Number of Queries} \quad (4)$$

4. 4 F1-Score

Combines precision and recall into a single measure [16].

$$\mathrm{F1} = 2 * (\mathrm{P} * \mathrm{R}) / (\mathrm{P} + \mathrm{R}) \quad (5)$$

4. 5 Accuracy

Measures the ratio of overall correctness cases to all cases [17].

$$\mathrm{Accuracy} = (\mathrm{TP} + \mathrm{TN}) / (\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}) \quad (6)$$

where TP is true positive, FP is false positive, FN is false negative, and TN is true negative.

4. 6 Peak Signal-to-Noise Ratio (PSNR)

A measure uses to measure the proportion of the similarity or extent of the difference between two images of the same structure. It is determined by:

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{M}^2}{\mathrm{MSE}}\right) \quad (7)$$

The resulting value is 100 when two images are identical and begin to fall below that amount when the two images are different. The value is zero when two images are completely different [18].

4. 7 Structure Similarity Index (SSIM)

The SSIM index assesses how similar a tested image p is, to the original image q visually using a formula.

$$\mathrm{SSIM}(p,q) = [K(p,q)]^{\alpha}.[Q(p,q)]^{\beta}.[W(p,q)]^{\gamma} \quad (8)$$

Where α, β, and γ are the parameters that determine the significance of each component.

$$\mathrm{K}(p,q) = (2 * \mathrm{M_p} * \mathrm{M_q} + \mathrm{S_1})/(\mathrm{M_p^2} + \mathrm{M_q^2} + \mathrm{S_1}) \quad (9)$$

$$\mathrm{Q}(p,q) = (2 * \sigma_p * \sigma_q + \mathrm{S_2})/(\sigma_p^2 + \sigma_q^2 + \mathrm{S_2}) \quad (10)$$

$$\mathrm{W}(p,q) = (\sigma_{pq} + \mathrm{S_3})/(\sigma_p * \sigma_q + \mathrm{S_3}) \quad (11)$$

Where constants S_1, S_2 and S_3 are utilized to prevent any instabilities that may average from the pixel value $(M_p^2 + M_q^2)$, standard deviation $(\sigma_p^2 + \sigma_q^2)$ or $(\sigma_p * \sigma_q)$ is approaching zero. SSIM (p, q) varies between 0 (indicating dissimilarity) 1 (representing identical patches) [19].

5 Proposed Method

The proposed method is designed to select the optimal reference images for colorizing monochrome videos with multiple scenes. The method included two main parts: (1) Training the Proposed Network, which focuses on training the suggested network to rehabilitate for future works. (2) Feature extraction, provides a feature vector containing potential reference images, and (3) Scene analysis, performs two functions: first, analyze the video to Count the Number of Scenes in the Video, which is related to the number of reference images that must be selected. The second function is the automatic selection of the best reference images for each scene by using the trained networks.

5. 1 Training the Proposed Network

Feature extraction aims to identify the optimal feature vector for selecting the optimal reference image for color transfer, which is a key objective in automatic video scene colorization. The two stages are:

The goal of the training phase is to create a reference image vector with corresponding features to assist the user in selecting the optimal image for each video scene. In this proposal, a combined ResNet50 and a Generalized Mean Pooling (GeM) model are suggested for feature extraction. Initially, the ResNet50-GeM is provided with the ability to recognize various scenes by training it on the public video dataset. The flowchart of the proposed model is shown in Fig. 2.
The main goal of ResNet50 is to extract discriminative features (spatial and semantic) from the input images through its deep layers, which can be used for several purposes. To mitigate the problem of vanishing gradient, the ResNet50 used the residual connections, the skip connections permit the gradient to effectively flow during the backpropagation, and enable the ResNet to train without performance degradation. In general, the early layers of ResNet50 capture the low features (edge, texture), while the deep layers capture the high-level features (objects, scenes).

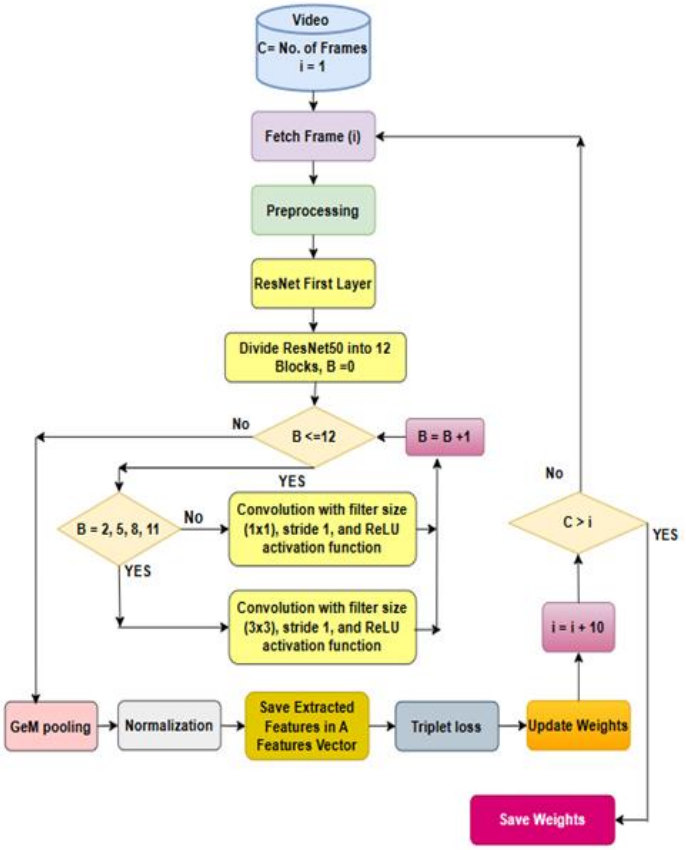

Figure 2: Training ResNet50Gem

On the other way, the goal of the pooling technique GeM is to aggregate the spatial features into a compact global fixed-size descriptor, to improve the feature representation by capturing more discriminative information. GeM is robust to outliers and retains more spatial information, unlike max and average pooling.

The combination of ResNet50 and GeM has many advantages for this proposal, such as:

1. Benefits of hierarchical features extracted by ResNet and the ability of GeM to aggregate these features into a fixed size descriptor which is useful for retrieving and matching data.
2. Unlike traditional pooling, the Gem is more flexible to pool the features extracted by ResNet50 which is rich in spatial and semantic information.
3. The combination model is effective in capturing local and global context, which makes it generalized better across different scenes.
4. This combination is crucial in capturing both fine-grained details and global context, leading to better performance, robustness, and adaptability. This, makes ResNetGem well-suited, particularly for scene recognition.

Stage 1.1: Preprocessing

The first stage in the training phase is the preprocessing stage and involves many steps:

1. Loading a video: The video dataset is loaded, and each video is split into multiple frames, which are then used as input to the model sequence.

2. Image Resizing: Each input frame to ResNet50 is resized to 224×224 pixels to meet the network requirement. Frames resizing ensures that the input images into the network have a fixed dimension appropriate for feature extraction.

3. Image Normalization: In this step, the pixel value of each frame is scaled into a range of [0 .. 1]. In the training phase, the large variation in pixel intensities may lead to unstable gradients and slower convergence. So, the normalization process ensures consistent input to the network.

4. Image Enhancement: Denoising techniques are applied to enhance the images and reduce noise in the frames. Denoising is implemented by using the mean filter.

5. Handling the Vanishing Gradient Problem: Deep networks like ResNet50, with 50 layers, can suffer from the vanishing gradient problem, where gradients become too small, making it difficult to update weights in earlier layers. To counter this, ResNet50 is divided into 12 blocks, each consisting of multiple layers (see Figure 2). These blocks include skip connections, convolution, and Rectified Linear Unit (ReLU) layers, which help prevent the vanishing gradient issue. Skip connections bypass one or more layers, allowing the network to learn residuals (the difference between input and output) instead of learning the complete output at once. This helps in sharing and reusing parameters across different layers and improves feature extraction.

Stage 1.2: Applying ResNet50

After preprocessing, the processed frames are passed into the ResNet50 architecture, starting with the first layer:

1. First Layer Operations: A 7×7 convolution is applied with a stride of 2, followed by ReLU activation. After the convolution, max pooling is performed to reduce the spatial size and help the model focus on the most important features.

2. Block Structure: The layers of ResNet50 are organized into blocks (as discussed earlier). Two types of convolution layers are used:

- 1×1 convolution layers for dimensionality reduction and expansion.
- 3×3 convolution layers for feature extraction.

3. Block Details

- Block 1: In this block image dimensions are changed while preserving essential features, by applying a 1×1 convolution, ReLU, and a stride of 1.
- Blocks 2, 5, 8, and 11: In these blocks the more detailed features are extracted from the image, by applying a 3×3 convolution, ReLU, and a stride of 1.
- Other Blocks: These blocks are utilized to fine-tune the extracted features while maintaining spatial dimensions, by applying a 1×1 convolution, ReLU, and a stride of 1.

Stage 1.3: Generalized Mean Pooling (GeM)

The generalized mean of the feature maps is computed by the GeM Pooling layer that uses the output from layer 49th as input. GeM pooling is crucial to reduce the outlier influence on creating strong features. Unlike the traditional pooling techniques, GeM improves the overall robustness.

Stage 1.4: Feature Vector Normalization

L2 normalization is utilized to normalize the extracted features. This makes the features appropriate for the subsequent processes like image retrieval, by ensuring that the norm is equal to 1 for all feature vectors.

Stage 1.5: Triplet Loss

The extracted features are input into the Triplet Loss function. The Triplet Loss function is a metric learning method that ensures similar images have the closer vectors in the feature space, while the dissimilar are far apart. Triplet Loss is optimum for training models where the aim is to learn meaningful representations for image retrieval. The triplet loss function stimulates the model to generate distinct clusters for various categories, which increases the efficiency of retrieval. Using Triplet loss before updating weights prevents overfitting to a single triplet and instead learns a robust feature representation.

Stage 1.6: Updating Weights

The model is updating the weights (backpropagation process) to ensure that the new frame benefits from learning the previous frames.

The embedding space in the last stage improved by reducing the distance of the intra-class and increasing the distance of the inter-class. Updating weights helps the model to learn before seeing a new frame.

All the frames in the video are processed and each frame finally will have a contribution to updating the network weights by a certain percentage. The model at the end of the training learns to create feature vectors.

5. 2 Feature Extraction

In this part of the proposal, the same video dataset that is used in the training is reused with the trained network to generate a feature vector dataset. The feature extraction process utilizes the trained ResNet50GeM network, as shown in Fig. 2, with the key difference that weights remain fixed (i.e., no further updates occur). Instead of saving network weights, the output consists of feature matrices representing each frame.

Theoretically, features could be extracted during the training process and saved for the next stages. In this proposal, the network is reused for feature extraction to ensure that using the final optimized model extracts all features consistently. During the training, feature representations changed according to the network learned, meaning features in the early stage may not align with those in the later learned representations. The post-training features extracted, guarantee that a stable and fully optimized network is used to process all frames, leading to a more structured and reliable feature space.
To reduce processing time, frame redundancy, and maintain temporal consistency, frames are processed at trained networks at regular intervals (one frame every ten frames). A structured feature dataset is created by storing the extracted feature vectors, where dissimilar images are mapped farther apart, and similar images have a smaller Euclidean distance. This dataset provides feature matching efficiently, ultimately helping the selection of the optimal reference image for color transfer.

5. 3 Scence Analysis and Network Application

This part consists of two phases, the first one is for counting the number of scenes in the video, and the second phase is the application of a trained network to suggest the best reference image for a particular scene.

Phase 1: Counting the Number of Scenes in the Video.

The number of scenes in a video is critical for selecting the best reference images and achieving accurate video colorization. Most existing works and articles rely on a single reference image, disregarding variations in scenes, objects, colors, illumination, and other factors in the video. While a few studies have used more than one reference image, these images were manually selected and did not correspond proportionally to the number of scenes in the video. In this paper, we propose a method for automatically determining the number of scenes in a video and, based on this count, suggesting multiple reference images. The flowchart of the proposed model is shown in Figure 3.

The proposed method for counting the number of scenes in a video consists of several steps:

1. Frame Extraction: The input video is split into individual frames based on the video's size and format
2. SSIM Calculation: Calculate the Structural Similarity Index (SSIM) between each successive pair of frames.
3. SSIM Deviation (SSDev): Calculate the SSIM deviation (SD) from the ideal SSIM value of 1, as follows:

$$SSDev = 1 - SSIM\ (previous\ frame, current\ frame) \quad (7)$$

Storing SSDev Values: Store the SD values in a one-dimensional matrix (M).

4. Repeat Process: Repeat the above steps for all frames in the video.
5. Mean Calculation: Compute the mean (Mean_M) of the matrix M. In this case, Mean is the average of all the SD values in the matrix M. Compute the standard deviation (Std_M) of matrix M. Std_M is calculated by the following formula.

$$Std_M = \frac{\sum_{i=1}^{N}(SSDev_i - Mean\_M)}{N} \quad (8)$$

Where N is the matrix M size.

6. Threshold Calculation: Use the mean and standard deviation to calculate the threshold using the following suggested Equation (9):

$$Threshold = Mean\_M + 2.5 * Std\_M \quad (9)$$

7. Scene Detection: Compare each SSDev value in matrix M with the threshold. If the SSDev is less than the threshold, consider the corresponding frame as a new scene, and save the frame image to a specific file (S) with its index in matrix M. Otherwise, the frame is ignored. The index of the matrix M represents the frame number in the video frame sequence when SSDev was calculated. (Note: the first frame in the video is saved in the file (S) initially as the first scene).

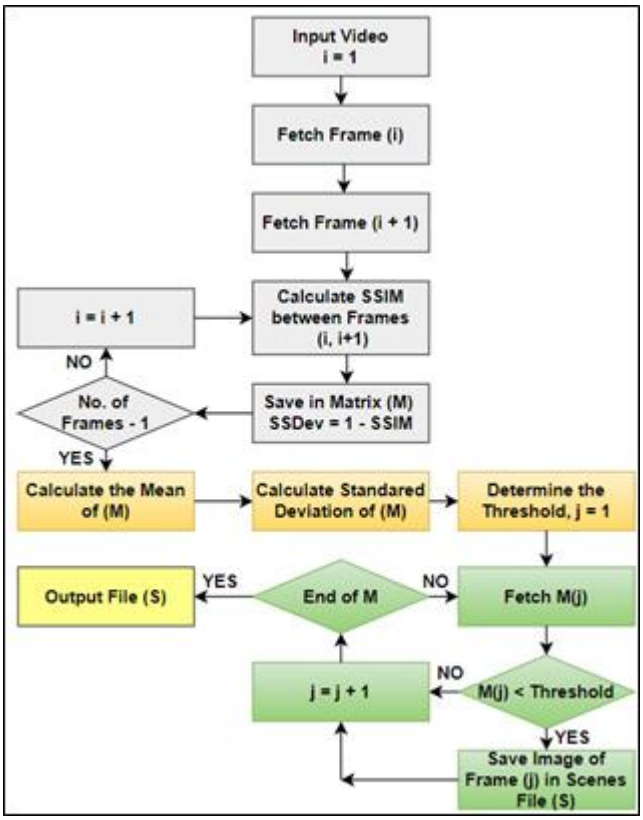

Figure 3: Flowchart to determine the number of senses.

Phase 2: Selecting the best reference images

In this phase, the trained ResNet50-GeM model is employed to select the most similar reference image from the dataset to a particular scene based on the features extracted during the

training phase. The input to this phase consists of the frames (images) selected in the previous phase. For each input image, the model identifies a reference image with the closest matching features. This selection process involves multiple stages, ensuring that the chosen image is highly similar to the input in terms of feature representation. The overall flowchart of this proposal is illustrated in Figure 4.

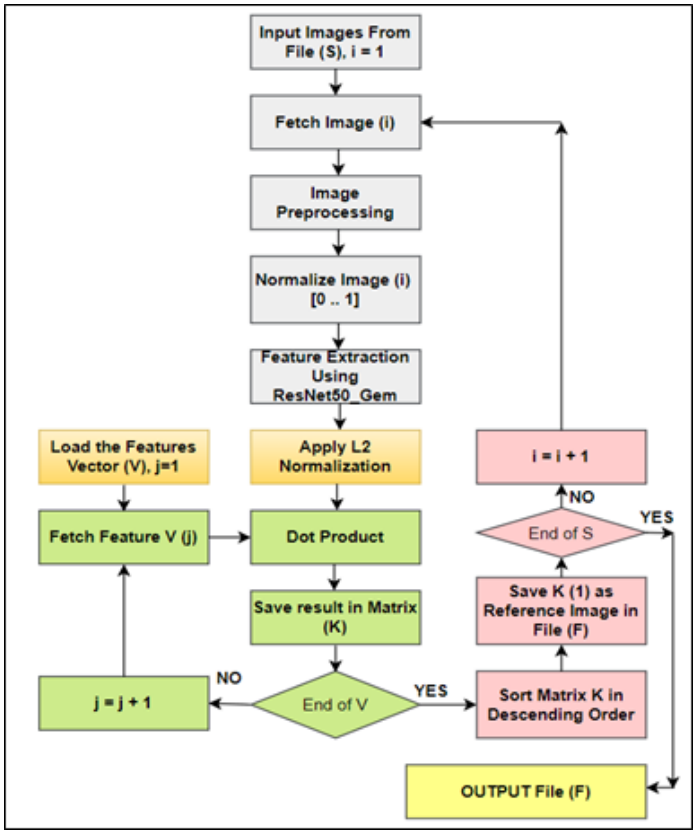

Figure 4: Selecting the best reference image Phase.

Stage 5.2.1: Image preprocessing: in this step, the input image is resized into 224×224 pixels, and then normalized to be in the standard range of [0..1]. Also, the image is denoising using the mean filter.

Stage 5.2.2: The processed image input to the trained ResNet50-GeM model for feature extraction. The extracted features are normalized using L2 normalization.

Stage 5.2.3: Calculate Similarity Scores: apply the dot product between the extracted features from the stage (5.2.2) with each feature stored in the extracted features vector (V) generated in the training phase. The result of each dot product process is one value, saved in one-dimension matrix (K).

Stage 5.2.4: The values in matrix K are sorted in descending order.

Stage 5.2.5: The image corresponding to the first rank in the matrix K is selected as an ideal reference image for the input video frame.

Stage 5.2.6: The previous stages repeated for all the frames that counted as new scenes in phase one.

6 Proposed Method

To achieve optimal results, the proposed network must be thoroughly trained on a sufficiently large dataset with an appropriate number of epochs. In this proposal, a public video dataset (the YouTube-8M dataset) was selected for training. The dataset included 1,569 different videos that provide a high number of images (frames) with a wide variety of scenes and content. The dataset was divided into two main groups, one for training which included 1269 videos (1,219 videos (7,993,132 frames) for training and 50 videos (321,306 frames) for validation). This approach ensures a large pool of candidate reference images for the selection process. The other group is for testing which includes 300 videos ((1,928,572 frames)). For the training phase, we used 1,219 videos (7,993,132 frames).

The network was trained with different numbers of epochs, and both accuracy and loss were measured, as shown in Figure 5. From the figure, we observe that the model's accuracy improves with an increasing number of epochs. However, after 100 epochs, the accuracy plateaus, indicating no further significant improvement. Therefore, we concluded that the optimal number of epochs for training is 100 where the accuracy is up to 99%.

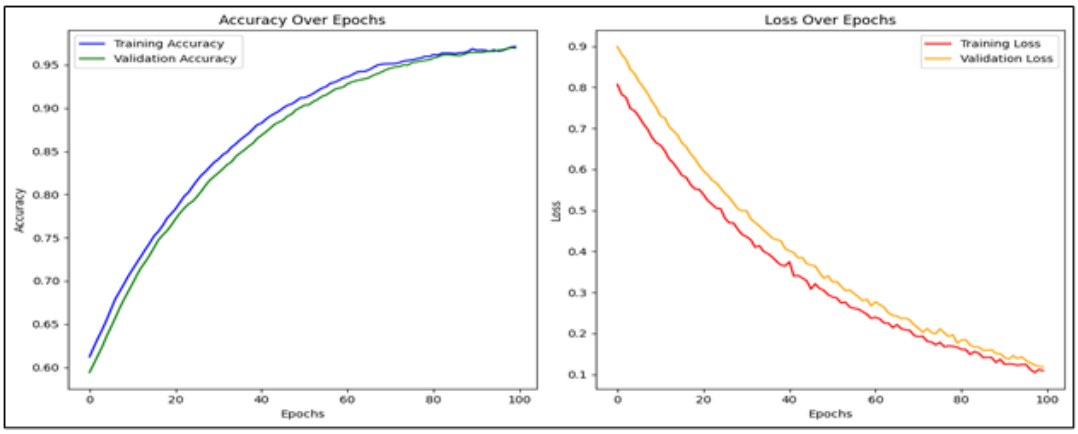

Figure 5: Training Process Flow of ResNet50-Gem Model

To further evaluate the efficiency of the proposed method in counting the number of scenes, we tested it on 10 different videos, each containing a unique number of scenes. The performance results for predicting the number of scenes are presented in Table 1. Based on the results in Table I, we conclude that the proposed model accurately and efficiently predicts the number of scenes, demonstrating high accuracy across various video samples.

Table 1: Performance Indices Reflecting Image Search Reliability Across Cases When Different Numbers of Images Are Retrieved.

No Image Retrieval	Mean Precision	Mean Recall	Mean F1-Score	Mean Average Precision (MAP)
1	0.9960	0.9850	0.9905	0.9952
2	0.9945	0.9870	0.9907	0.9940
3	0.9920	0.9895	0.9907	0.9930
4	0.9915	0.9910	0.9912	0.9925
5	0.9900	0.9925	0.9912	0.9918
6	0.9890	0.9940	0.9915	0.9910
7	0.9875	0.9955	0.9915	0.9905
8	0.9865	0.9960	0.9912	0.9900
9	0.9858	0.9965	0.9911	0.9898
10	0.9850	0.9968	0.9909	0.9895

We also evaluated the performance of the proposed method in selecting the best reference image for each scene, with the results shown in Figure 6. The method effectively identifies the frame number for each new scene and selects the best-matching colored image to use as a reference. This reference image is then used to transfer colors to the black-and-white video frames, ensuring accurate colorization for each scene.

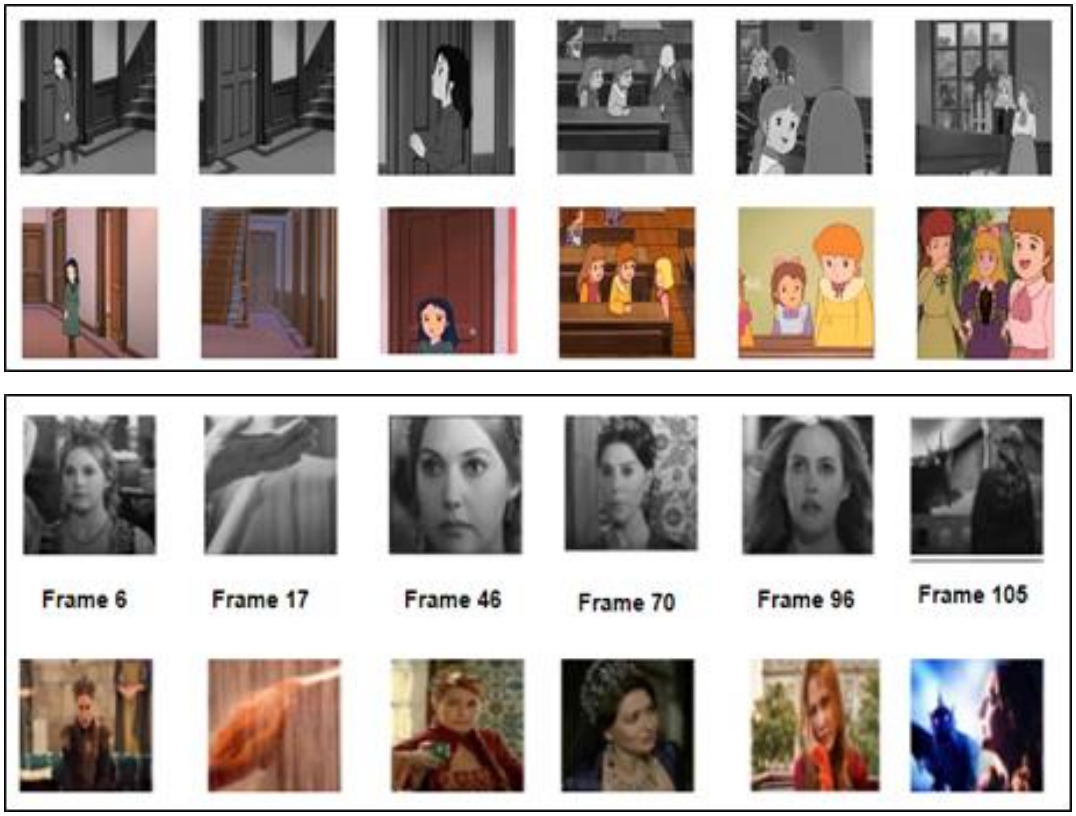

Figure 6: Selecting reference images for gray frames in video across multiple scenes.

In addition to these accuracy metrics, we also evaluated the perceptual quality of the colorized videos using Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). These metrics offer a complementary perspective on visual fidelity, confirming the quality and consistency of the colorized frames. Our method achieved an average SSIM of 0.92 and a PSNR of 31.7 dB across the test samples, indicating high visual fidelity and quality in the colorization results.

The current method was compared with previous works, as illustrated in Table 2. The results demonstrate that the proposed

method outperforms the other methods, offering superior performance in terms of both accuracy and efficiency.

Reference	Mean Precision	Mean Recall	Mean F1-Score	Mean Average Precision (MAP)	Retrieval Accuracy
[12]	0.92	0.95	0.94	0.86	97%
[9]	0.88	0.92	0.90	0.82	96%
[10]	0.85	0.90	0.88	0.80	95%
[11]	0.82	0.88	0.85	0.78	94%
[13]	0.80	0.86	0.83	0.76	93%
[8]	0.80	0.84	0.81	0.74	92%
Our	0.9920	0.9919	0.9920	0.9920	99%

7 Future Work

- **Support for Dynamic Scenes:** While the proposed method demonstrates strong performance in videos with relatively stable scenes, its effectiveness may diminish in highly dynamic environments characterized by rapid scene transitions or significant camera motion. Future work could focus on developing adaptive reference update mechanisms that adjust dynamically to the temporal complexity of the video content. Furthermore, integrating object tracking and motion segmentation techniques may enhance the model's robustness across diverse temporal contexts.
- **Dataset Diversity:** Although the training and evaluation were conducted using the large-scale YouTube-8M dataset, future work should consider incorporating a more diverse set of video types. These could include animated content, historical footage, and cinematic material, allowing for a broader evaluation of the generalizability of the proposed method acrossvarious domains and stylistic variations.
- **Computational Complexity and Execution Requirements:** The proposed method was implemented using an NVIDIA Tesla V100 GPU, with an average processing time of approximately 5.2 minutes for a 30-second video at 24 frames per second. While this is acceptable for offline applications, optimizing inference time and reducing model complexity will be crucial for real-world deployment. Future work could explore model distillation, pruning strategies, or other optimizations to accelerate the performance of the system.
- **Real-Time Colorization:** The current framework is not optimized for real-time video colorization, as it prioritizes accuracy and temporal consistency over speed. However, by leveraging lightweight backbone architectures and employing temporal skipping strategies, it is feasible to develop a real-time version of the proposed pipeline. Such a system would be suitable for streaming or live video applications, enabling faster processing while maintaining highquality.
- **User Interaction Integration:** While the reference image selection process is fully automated in the current system, incorporating optional user interaction could increase flexibility in applications requiring creative or subjective color choices. For example, allowing users to manually select or refine reference frames could provide greater control over the aesthetic output, making it particularly beneficial for artistic or archival restoration tasks.

8 Conclusion

This paper proposes a novel method for auto-color transfer across video frames, leveraging critical video features to select the optimal reference images. By utilizing the dynamic nature of video data instead of static images, this method outperforms previous techniques, achieving impressive accuracy up to 99% and an F1-Score of up to 99.2%. The proposed approach reduces overfitting and enhances model robustness, eliminating the need for an augmentation process, thanks to the inherent diversity within video frames.

A key contribution of this method is the fusion of ResNet50 and GeM architectures, which play a crucial role in scene detection and generating specific reference images for each video scene. This combination provides a versatile foundation for color transfer, ensuring the preservation of subtle visual elements across diverse scenes. By integrating ResNet50 with GeM pooling, the method captures both multi-scale and detailed features from the video frames. ResNet50's deep layers excel at robust feature extraction, while GeM pooling focuses selectively on relevant image regions, enhancing the model’s performance and adaptability.

The synergy between ResNet50 and GeM pooling enables the accurate transfer of colors across frames, maintaining consistency in the visual content of video scenes. This approach also improves visual quality and adaptability, particularly for tasks requiring adjustments of scene-specific colors. Furthermore, this method demonstrates its capacity to handle large, diverse datasets, including high-resolution videos, making it suitable for a wide range of applications.

In future work, it is recommended that this method be extended to handle dynamic scenes with frequent transitions. By developing adaptive reference update mechanisms and incorporating motion tracking, the system could be further optimized to perform effectively under these more complex conditions.

Additionally, automated video colorization, especially when applied to archival or historical footage, raises important ethical considerations. Altering the color content of such materials can unintentionally influence viewers’ perceptions of historical

accuracy. To address this concern, we advocate for transparency in the use of colorization models, encouraging users to clearly disclose when content has been modified through automated techniques. This ensures that the integrity of historical representations is maintained and that the impact of colorization on viewer perception is properly acknowledged.

References

[1] Kouzouglidis, P., Sfikas, G. & Nikou, C. Automatic video colorization using 3D conditional generative adversarial networks. *Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14*, Springer, 209–218 , doi: 10.1007/978-3-030-33720-9_16 (2019).

[2] Zhang, R., Isola, P. & Efros, A. A. Colorful image colorization. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, Springer, 649–666, doi:10.1007/978-3-319-46487-9_40 (2016).

[3] Kang, X., et al. NTIRE 2023 video colorization challenge. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1570–1581, doi : 10.1109/CVPRW59228.2023.00159 (2023).

[4] Shi, M., Zhang, J.-Q., Chen, S.-Y., Gao, L., Lai, Y.-K. & Zhang, F.-L. Reference-based deep line art video colorization. *IEEE Trans. Vis. Comput. Graph.* **29**, 2965–2979, doi: 10.1109/TVCG.2022.3146000 (2023).

[5] Liu, Y., et al. Temporally consistent video colorization with deep feature propagation and self-regularization learning. *Comput. Vis. Media* **10**, 375–395 (2024).

[6] Polat, A. A., Şahın, M. F. & Karsligıl, M. E. Video colorizing with automatic reference image selection. *2021 29th Signal Processing and Communications Applications Conference (SIU)*, IEEE, 1–4, doi: 10.1109/SIU53274.2021.9477833 (2021).

[7] Yu, B., Zhang, S.-Y., Chen, M., Wang, X., Zhang, C., Zhu, X. & Li, X. Temporal consistent automatic video colorization via semantic correspondence. *arXiv.org*, doi:10.48550/arXiv.2305.07904 (2023).

[8] Vijetha, U. & Geetha, V. Superpixel based image colorization with automated reference image selection. *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, IEEE, 1–6, doi : 10.1109/SCEECS57921.2023.10061822 (2023).

[9] Yao, J., Li, Y., Yang, B. & Wang, C. Learning global image representation with generalized-mean pooling and smoothed average precision for large-scale CBIR. *IET Image Process.* **17**, 2748–2763, doi:10.1049/ipr2.12825 (2023).

[10] Gherbi, T., Zeggari, A., Ahmed Seghir, Z. & Hachouf, F. An evaluation metric for image retrieval systems, using entropy for grouped precision of relevant retrievals. *J. Intell. Fuzzy Syst.* **45**, 3665–3677, doi: 10.3233/JIFS-223623 (2023).

[11] Ali, O. & Ibrahim, S. (1 + 1)-evolutionary gradient strategy to evolve global term weights in information retrieval. *Advances in Computational Intelligence Systems. Contributions presented at the 16th UK Workshop on Computational Intelligence*, doi:10.1007/978-3-319-46562-3 (2022).

[12] Panghurian, F. P., Pranoto, H., Irwansyah, E. & Pramudya, F. S. Comparison of ResNet models in UNet classifier for mapping oil palm plantation area with semantic segmentation approach. *Int. J. Adv. Comput. Sci. Appl.* **15**, 1–8, doi: 10.14569/IJACSA.2024.0150741 (2024).

[13] Y. Yang, J. Pan, Z. Peng, X. Du, Z. Tao, and J. Tang. Bistnet: Semantic image prior guided bidirectional temporal feature fusion for deep exemplar-based video colorization. *IEEE Trans. Pattern Anal. Mach. Intell*., doi: 10.1109/TPAMI.2024.3370920 (2024).

[14] Jawad, M. A. & Khursheed, F. A novel approach for color-balanced reference image selection for breast histology image normalization. Biomed. Signal Process. Control **94**, 106299, doi: 10.1016/j.bspc.2024.106299 (2024).

[15] Gautam, G. & Khanna, A. Content-based image retrieval system using CNN-based deep learning models. Procedia Comput. Sci. **235**, 3131–3141, doi: 10.1016/j.procs.2024.04.296 (2024).

[16] Alsaadi, E. M. T. A., Fayadh, S. M. & Alabaichi, A. A review on security challenges and approaches in the cloud computing. AIP Conf. Proc. 2290, doi:10.1063/5.0027460 (2020).

[17] Thabit, A. E. M. & El Abbadi, N. K. Auto animal detection and classification among (Fish, reptiles and amphibians categories) using deep learning. J. Adv. Res. Dyn. Control Syst. **11** (10 Special Issue), 726–736 (2019), doi:10.5373/JARDCS/V11SP10/20192863.

[18] Memon, F., Unar, M. A. & Memon, S. Image quality assessment for performance evaluation of focus measure operators. Mehran Univ. Res. J. Eng. Technol. **34** , 523–530 (2016), ISSN: 0254-7821.

[19] Bakurov, I., Buzzelli, M., Schettini, R., Castelli, M. & Vanneschi, L. Structural similarity index (SSIM) revisited: A data-driven approach. Expert Syst. Appl. **189**, 116087, doi:10.1016/j.eswa.2021.116087 (2022).