Ingénierie des Systèmes d'Information
Vol. 30, No. 4, April, 2025, pp. 995-1004
Journal homepage: http://iieta.org/journals/isi

ViT-Based Automatic Grayscale Image Colorization with a Hybrid Loss Function

Ahmed Al-Ghanimi*, Amir Lakizadeh

Computer Engineering and Information Technology Department, University of Qom, Qom 37161-46611, Iran

Corresponding Author Email: phar.ahmed.alganimi@uobabylon.edu.iq

https://doi.org/10.18280/isi.300416

Received: 13 February 2025
Revised: 17 April 2025
Accepted: 24 April 2025
Available online: 30 April 2025

Keywords:
image colorization, deep learning, vision transformer (ViT), inception module, automatic colorization, data augmentation, loss function

ABSTRACT

Yet, automatic colorization of black and white images is still a challenging task in computer vision, given the effective feature extraction and perceptual consistency which this task requires. A new advancement in deep learning is introduced in this study, one that will increase accuracy and speed in colorization. It replaces the conventional ResNet blocks with Inception-based building blocks to capture the multi-scale features and integrates the pretrained vision transformer (ViT) into modeling distance-dependent long-range spatial information. A hybrid loss function is introduced, which combines pixel-wise precision (MSE) with perceptual similarity and adversarial feedback to ensure realism and fidelity in the generated images. On all the evaluation metrics, including obtaining a better MSE and FID score and increasing PSNR and SSIM value, the proposed method outperformed the state-of-the-art methods, all measured in spite of other benefits such as extra rendering of natural textures, realistic skin colors, and consistent heavenly colors even in the complex portions of the image. Besides, it shows reduced inference time and rapid convergence time during training, which proves that the model can be used in real-time applications while having constraints regarding resources. In a nutshell, the proposed method presents a realistic yet high-quality solution to automatic colorization in images, and it has excellent prospects for real-life situations in historical photo restorations and generating digital content.

1. INTRODUCTION

Colorizing black and white images has always been a task that sparks the interest of artists and scientists [1]. It involves adding color to photographs and enhancing the impact of modern images. This process blends creativity, with technology in a captivating way [2]. Humans possess a talent, for envision. Adding colors to black and white images. However, automating this task has proven to be quite challenging with the advancements, in technology [3]. In the years deep learning has become a game changer, in revolutionizing how we tackle intricate visual tasks like adding color to images [4]. Older methods for this job heavily depended on processes or basic algorithms that needed a lot of effort and couldn't create natural looking outcomes with finesse. To networks, especially Convolutional Neural Networks (CNNs) [5] and the innovation of Generative Adversarial Networks (GANs) [6]. In times there has been a change that allows automated systems to create colored images that closely emulate real life [7].

Colorization of grayscale images has been termed as an ill-posed problem despite these advancements. Many existing methods tend to have difficulties predicting colors in the context of intricate scenes, which sometimes result in unnatural or inconsistent outputs [8]. Furthermore, numerous models also require huge datasets and prolonged training times, therefore remaining unsuitable for real-time usage [9]. The problem has therefore become how to design a method that improves the accuracy of colorization but is even more efficient for quick colorization of large numbers of images without sacrificing quality [10].

This research tackles these obstacles by suggesting a learning method that merges innovative designs with a refined loss function and optimization strategies. We present a model structure that swaps out ResNet components, with Inception modules to help the model grasp a broader spectrum of spatial features. Moreover, we incorporate a trained vision transformer (ViT) as an extractor of features to boost the model's capacity to recognize and mimic the complex color patterns found in real life images. To enhance how colorful images, look to the eyes perspective betterment wise and make them more attractive visually we've come up with a loss function that combines color matching, with visual similarity to ensure the colors produced are both precise and aesthetically pleasing.

The importance of this study is, in its ability to connect the gap between image colorization techniques and the needs of real-world usage scenarios. Enhancing both the precision and speed of the colorization procedure not pushes the boundaries of technology. Also creates opportunities for use in areas, like digital media production, historical conservation efforts and content development. The new approach seeks to address the drawbacks of methods by offering an adaptable solution, for adding color to black and white images automatically in different fields.

In summary even though there have been advancements, in

the realm of image colorization the demand for techniques that can produce top notch outcomes quickly and effectively is still pressing. This research aims to address this issue by introducing a method that merges the advantages of deep learning with state-of-the-art architectural design and optimization techniques opening up opportunities for improved and user-friendly image colorization technologies, in the future.

2. LITERATURE REVIEW

Colorizing grayscale images automatically has been an area of study, in computer vision for years [11, 12]. Initially methods were mostly manual or based on algorithms. While groundbreaking at the time they had limitations in creating dependable outcomes. The introduction of learning has brought about changes in the field. This has led to the development of automated techniques, for colorizing images. This review of literature gives a summary of advancements in this field. Showcases the shift, from conventional methods, to modern deep learning approaches.

2.1 Early approaches to image colorization

In the stages of image colorization development relied heavily on processes that demanded skilled artists to painstakingly add colors to black and white images. The methods employed were effective. Proved to be inefficient, for applications due to the significant time investment required [13]. With advancements in imaging technologies came the exploration of solutions, by researchers. One of the automated methods included color transfer techniques that involved applying colors from a reference image onto a grayscale image through similarity assessments [14]. Although these techniques provided some level of automation assistance. Relied significantly on the quality of the base image. They frequently struggled to adapt to a variety of settings [15].

2.2 The rise of machine learning and deep learning

The emergence of machine learning technologies, like learning brought about a transformation in the realm of image coloring processes. Convolutional Neural Networks (CNNs) emerged as the building blocks in colorization models due to their capacity to autonomously grasp features, from extensive data sets. Initial approaches based on CNN technology presented by researchers [5] and [16] showcased the capabilities of learning over conventional methods. In 2016 Zhang and colleagues innovative "Colorful Image Colorization" model stood out for its approach, in using a classification method to determine the hue and chroma values of individual pixels in images. However, these initial deep learning models showed results but faced challenges due to their dependence on labeled datasets [17] and their struggle to grasp the broader context effectively which prompted further exploration, into newer architectures and methods [18].

2.3 Generative adversarial networks (GANs)

The emergence of Generative Adversarial Networks (GAN) as outlined by Goodfellow and colleagues in 2014 represented an advancement, in the development of techniques for adding color to images [19]. GAN involves a pair of networks-the creator and the critic-that work against each other during the learning process to produce lifelike and intricate results. Approaches such, as the method introduced by Isola and team in 2017. In their "Pix2Pix" framework, from before utilized GANs for image-to-image translation tasks like colorization [20]. This method was improved upon in studies like the CycleGAN model which brought in cycle consistency loss to guarantee reliable and precise translations [21]. Despite their achievements GAN based models frequently encountered problems, like mode collapse and instability while training. Scientists tackled these obstacles by integrating methods such, as normalization and self-directed focus mechanisms, into the SS N GAM design. This enhancement resulted in increased stability and higher quality of the images produced [22].

2.4 ViT and hybrid models

Recently there have been developments, in image colorization thanks to the introduction of ViT [23]. Unlike CNN models that focus more on patterns ViT are skilled at understanding long range connections in data. This makes them valuable for tasks that need a context understanding. Hybrid models that merge CNN and transformer technologies have displayed encouraging outcomes across tasks, like colorization. These models combine the aspects of both structures by focusing on details, within the image while also grasping the overall picture [24].

2.5 Optimization techniques and loss functions

Advancements, in architecture have been accompanied by the evolution of loss functions which're crucial for improving the effectiveness of colorization models [25]. While conventional loss functions, like Mean Squared Error (MSE) mainly prioritize pixel level precision. Often overlook perceptual quality considerations. To tackle this issue effectively researchers have introduced loss functions that evaluate high level features derived from trained networks resulting in more aesthetically pleasing outcomes [26]. Various types of loss functions, like loss and perceptual loss are commonly used in models along with content loss as a standard practice nowadays. Moreover, there have been investigations, into utilizing optimization algorithms inspired by meta heuristics to enhance model hyperparameters and boost training efficiency. These approaches facilitate the navigation through the loss terrains of learning models to guarantee the network reaches an ideal outcome [27].

While advancements have been made in automating image colorization processes there are still some hurdles to overcome in this field. Current models face difficulties when coloring scenes or images, with color cues. The reliance on datasets for training and the high computational demands of learning models are hindrances that limit broader usage. Furthermore, while many models exhibit precision, in controlled settings their effectiveness may vary when applied in real world scenarios [28].

The demand, for adaptable models that can effectively work across image categories and situations is evident. Furthermore, exploring enhanced data augmentation methods, such as delving into architectures like ViT and tuning hyperparameters using meta heuristic strategies hold potential as valuable avenues, for upcoming research endeavors.

The realm of image colorization has seen progress, over time. Transitioning from manual techniques to advanced deep

learning algorithms has played a key role in this evolution process. With these advancements come challenges that underscore the necessity for innovation. This review of existing literature emphasizes the significance of integrating enhancements in architecture design, with loss functions and optimization strategies to enhance the performance and efficiency of colorization methods. The upcoming study seeks to leverage these findings as a springboard to explore frontiers and possibilities in the realm of image colorization.

3. PROPOSED METHOD

In this part here we share our approach, for adding color to white images. The plan includes a variety of parts such as techniques to enhance data variety, an architecture for the deep learning model a new loss function introduction and incorporating ideas, from ViT. On top of that the hyperparameters of the model are fine-tuned using an algorithm that helps boost performance in the coloring process. The proposed methods flow diagram can be seen in Figure 1.

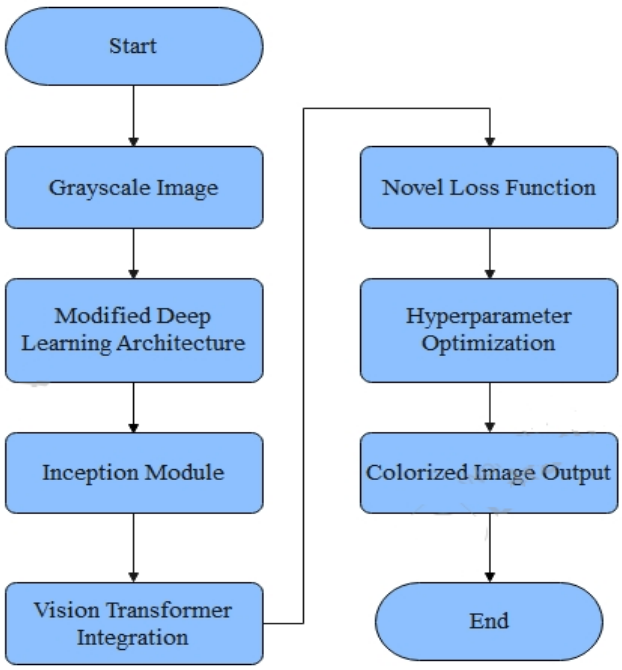

Figure 1. Flowchart of the proposed method

3.1 Data augmentation

Adding data is essential, in our strategy to enhance the variation, in the training data and avoid overfitting issues. This is implemented methods to modify the grayscale images to inputting them into the network.

- Incorporating rotations ranging from 15 degrees, to +15 degrees.
- Scaling involves randomly adjusting the size by a factor ranging from 80%, to 120%.
- Adjustment of brightness involves making fluctuations, in brightness levels within a range of plus or minus 20%.
- Horizontal and vertical flipping randomly.

These enhancements help the model experience a range of image changes to improve its ability to adapt to unseen data [29].

3.2 Modified deep learning architecture

Our suggestion involves adjusting the deep learning frameworks employed in image colorization by swapping out the ResNet module with an Inception module. The Inception module enables the model to grasp characteristics at scales a valuable trait, for interpreting intricate textures and designs within images.

The updated design includes these parts:

- A bottleneck layer is added to decrease the size of the feature map while retaining information.
- To bring the image back, to its dimensions after processing it with sampling layers, like pooling or convolutional layers that reduce its size.

In Figure 2 the flowchart of the updated deep learning architecture is displayed.

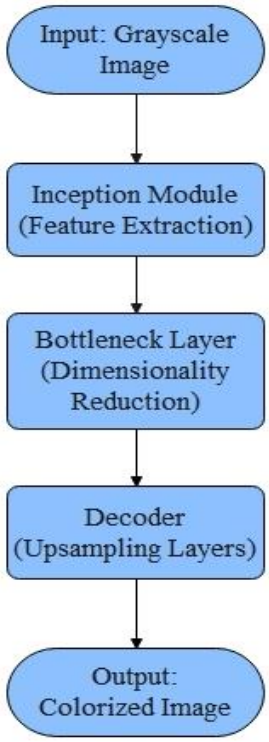

Figure 2. Flowchart of modified deep learning architecture

The architecture of the proposed model can be summarized as follows:

$$X_{out} = Decoder\left(\text{Inception}\left(X_{gray}\right)\right) \quad (1)$$

where, X_{gray} is the input grayscale image, Inception represents the modified inception module for feature extraction, and Decoder consists of the upsampling layers that reconstruct the colorized image.

3.3 Novel loss function

A new Loss Function is a crafted loss function that enhances the training efficiency of the suggested GAN based colorization model by tackling drawbacks of typical loss functions such, as MSE or Binary Cross-Entropy (BCE). This custom designed approach aims to handle the intricacies involved in colorization tasks by focusing on producing textures and vivid colors while maintaining intricate details. In

order to enhance the quality of colorization results we have developed a loss function that integrates goals.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_1 \mathcal{L}_{\text{Perceptual}} + \lambda_2 \mathcal{L}_{\text{GAN}} \quad (2)$$

where,

- $\mathcal{L}_{\text{total}}$: Total loss used for model optimization.
- $\mathcal{L}_{\text{MSE}}$: MSE loss (pixel-wise difference).
- $\mathcal{L}_{\text{Perceptual}}$: Perceptual loss based on high-level features.
- $\mathcal{L}_{\text{GAN}}$: Adversarial loss from the GAN framework.
- λ_1, λ_2 : Weighting coefficients balancing the contributions of perceptual and GAN losses.
- Learning rate: [0.0001, 0.01].
- λ_1 (MSE loss weight): [0.5, 2.0].
- λ_2 (Perceptual loss weight): [0.01, 0.5].

MSE Loss: This loss measures the pixel-wise difference between the predicted and ground truth color channels. Here, N represents the total number of pixels in the image and $Y_{\text{pred}}(i)$ predicted color value at pixel i, $Y_{\text{true}}(i)$ and ground truth color value at pixel i.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \left(Y_{\text{pred}}(i) - Y_{\text{true}}(i)\right)^2 \quad (3)$$

Perceptual Loss: To enhance the perceptual quality of the colorized images, we incorporate a perceptual loss calculated using the VGG19 network's feature maps. This loss helps to preserve high-level semantic information.

$$\mathcal{L}_{\text{Perceptual}} = \frac{1}{N}\sum_{i=1}^{N} \left(\phi\left(X_{\text{pred}}\right)(i) - \phi(X_{\text{true}})(i)\right)^2 \quad (4)$$

where, ϕ denotes the feature extractor (VGG19).

We incorporate a GAN loss element to promote the creation of images, in our process of adding color to images using our generator model and discriminator, for distinguishing between colorized images.

$$\mathcal{L}_{GAN} = \mathbb{E}[\log D(Y_{\text{true}})] + \mathbb{E}\left[\log\left(1 - D\left(Y_{\text{pred}}\right)\right)\right] \quad (5)$$

where, D is the discriminator.

3.4 Integration of ViT ideas

In order to enhance performance effectively we incorporate concepts, from ViT. Precisely we employ a self-focused mechanism during the feature extraction process. The self-focused layers aid the model in concentrating on parts of the image and grasping connections, between pixels. The self-attention mechanism can be represented as:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (6)$$

where, Q (query), K (key), and V (value) are matrices derived from the input features, and d_k is the dimensionality of the key vectors. This attention mechanism is integrated within the Inception module to enhance the feature representation.

In this study's exploration of ViT the incorporation of its principles proves instrumental, in improving image colorization techniques. Diverging from approaches like GAN or CNN that encounter challenges in grasping dependencies and holistic image contexts ViT overcomes this hurdle through the utilization of self-directed attention mechanisms. This strategy facilitates image processing by segment which enhances comprehension of connections, across sections of an image.

The suggested technique breaks down a picture into sections and converts each section into an embedded vector representation before moving it through layers that capture the connections, between these sections to enable the network to grasp broader relationships within the image as a whole. This method proves beneficial in zones, like skin tones or sky and shadow details as comprehending the overarching structure of the image contributes to a natural and coherent colorization outcome.

Moreover in ViTs self-attention mechanism permits the model to concentrate on areas of the image with levels of attention effectively capturing both local and global features unlike CNN which mainly focuses on specific pixel groups ViTs wider perspective helps it maintain consistency throughout the entire image enhancing color uniformity in areas where conventional approaches might face challenges By integrating positional embeddings this approach maintains spatial awareness retaining the relative positions of patches, within the image.

By incorporating ViTs into the GANS framework effectively boosts the model's capacity to create top notch colorizations that look realistic and enhance the texture details and overall visual appeal.

In order to discover the hyperparameters for our suggested model we utilize a heuristic optimization technique known as the Particle Swarm Optimization (PSO) algorithm. In this investigation the process of hyperparameter optimization is carried out using PSO which's a meta heuristic approach inspired by the social interactions seen in birds flocking or fish schooling movements. PSO proves to be quite efficient when exploring vast search areas to uncover the nearly best solutions which makes it a great option, for tweaking the hyperparameters of sophisticated deep learning models. PS0 operates by mimicking a group of solutions known as "particles" that traverse through the search area based on their experiences and those of nearby particles. Each particle signifies a combination of hyperparameters, for the learning model. The new approach involves employing PSOs to tune hyperparameters, within the GAN model which include adjusting the learning rate and batch size while also considering momentum factors.

Throughout this study's context, PSPPSO proved to be highly efficient, in adjusting the hyperparameters that have an influence on how steadily the model reaches an optimal state. The proposed approach achieves stability in training and faster convergence than what is attained by traditional methods, such as grid or random search, by optimizing several factors such as learning rates and momentum. With hyperparameter optimization using PSO, the model generalization capacity is enhanced, with corroborative evidence in improved performance metrics, PSNR, SSIM, and FID. Further, PSO minimizes the manual tuning trial-and-error process, leading to efficient use of computational resources.

In essence using PSOs in this research guarantees that the suggested model is fine tuned for top notch performance allow it to attain colorization quality, with training and inference times. This approach showcases the effectiveness of

optimization methods, in enhancing the real-world usability of deep learning models.

The PSO algorithm is used to optimize parameters such as learning rate, batch size, and the weighting factors λ_1 and λ_2 in the loss function.

The optimization process is guided by the following objective function:

$$\min_{\theta} \mathcal{L}_{\text{total}}(\theta) \quad (7)$$

The set of hyperparameters being optimized is denoted by Θ in this context, within the PS0 algorithm. These hyperparameters are iteratively fine-tuned based on how the model performs on a validation dataset, with the goal of reducing loss.

The new approach incorporates cutting edge methods to enhance the addition of color, to black and white images effectively.

4. THE EXPERIMENTAL RESULTS

4.1 Model optimization with PSO

The proposed model was trained and evaluated on the COCO (Common Objects in Context) dataset. COCO, consisting of 330,000 images, has 80 object categories. Images in COCO have different classes of objects and scenes, so it would be best suited to train a model for which generalization is required across different types of objects.

Due to the large variety of high-quality images containing all kinds of scenes and objects, COCO is suitable for our model. This variety guarantees that the model has the capacity to generalize across different contexts. The richness in annotations also highlights the dataset complexity, which, although we do not use directly, supports robust training of our models. Furthermore, since the dataset is widely used in research, it provides a legacy benchmark, allowing us to contrast the results of different state-of-the-art computer vision methods [30].

Hyperparameter tuning of proposed image colorization model in the study was done using PSO. It is a population-based metaheuristic that imitates the collective behavior of flocks of birds and schools of fish and is efficient in the exploration of high-dimensional search spaces in order to arrive at a near-optimal solution.

A single particle in the PSO algorithm represents one candidate solution-a hyperparameter set comprising different values of learning rate, batch size, and momentum values. The particles move iteratively across the solution space, altering their position according to their own experiences and some neighboring particles' performance. This movement is constrained by an objective function that evaluates the model on a validation set using combinations of error and perceptual quality metrics.

Empowerment of PSO for use in this paradigm achieved faster convergence and stable training as compared to normal methods such as grid search or random search. This efficient exploration of the parameter space by PSO led to lesser dependency on manual tuning and higher generalization ability for the model.

Hence, thereby bringing improvements in vital performance metrics such as PSNR, SSIM, and FID.

Theoretical acclimatization of PSO thus ensured proper settings of the learning process leading to high-quality colorization results at low computational cost with faster inference times. The optimization strategy thus adds to the practicality and scalability of the proposed method in real-life deployment.

4.2 Simulation parameters

Table 1 shows a comprehensive figure of the used simulation parameters.

The dataset was divided into three subsets as follows:

- Training set: 10,000 images
- Validation set: 2,000 images
- Test set: 2,000 images

These splits were used consistently across all experiments to evaluate model performance under comparable conditions.

Table 1. Simulation parameters

Category	Parameter	Value
Model Architecture	Inception Module Filters	64, 128, 256
	Pooling	Max-pooling
	Activation Function	ReLU
	Output Feature Map Size	64×64×256
	Self-Attention Heads	4
	Self-Attention Dimensionality	64
	Self-Attention Dropout Rate	0.1
	Up sampling Layers Filters	128, 64, 32
	Final Output Layer Filters	1×1 Conv, 2 Filters
	Final Layer Activation Function	Sigmoid
Training	Batch Size	32
	Initial Learning Rate	0.001
	Learning Rate Decay	0.95 (every 10 epochs)
	Optimizer	Adam
	Adam Parameters (β1, β2, ε)	0.9, 0.999, 1e-7
	MSE Loss Weight (λ1)	1.0
	Perceptual Loss Weight (λ2)	0.1
	GAN Loss Weight (λ3)	0.01
	Training Duration	100 epochs
	Early Stopping	Patience of 10 epochs
Hyperparameter Optimization	Algorithm	PSO
	Number of Particles	20
	Number of Iterations	50
Hardware and Software	GPU	NVIDIA Tesla V100 (16 GB VRAM)
	CPU	Intel Xeon E5-2698 v4
	RAM	128 GB DDR4
	Framework	TensorFlow 2.9
	Python Version	3.8
	Operating System	Ubuntu 20.04 LTS

4.3 Results

In this part of our study section we. Examine the outcomes of our suggested approach when compared to two standard methods known as GAN Dataset [28] and SSN GAN [22]. The assessment criteria consist of MSE Peak Signal, to Noise Ratio (PSNR) Structural Similarity Index (SSIM) and Fréchet Inception Distance (FID). Furthermore, we showcase results

to visually assess the colorization excellence, among techniques. These methods represent two prominent approaches in the field of automatic image colorization:

GAN-Dataset [28] is a representative of conventional GAN-based colorization techniques and is widely used as a benchmark in related studies.

SSN-GAN [22] introduces a self-supervised structure and normalization mechanisms, making it a more advanced baseline for comparison in terms of both image realism and training stability.

These baselines provide a strong and diverse foundation to assess the improvements brought by our method in terms of both visual quality and quantitative performance.

4.3.1 Quantitative analysis

The quantitative metrics were calculated on the validation set of the COCO dataset, and the results are summarized in Table 2.

Table 2. Quantitative analysis results

Method	MSE (Lower is Better)	PSNR (Higher is Better)	SSIM (Higher is Better)	FID (Lower is Better)
GAN-Dataset [28]	0.065	22.7	0.72	29.4
SSN-GAN [22]	0.059	23.3	0.75	27.8
Proposed Method	0.048	24.6	0.81	23.5

- MSE: The proposed method achieved the lowest MSE (0.048), indicating that the predicted colors are closer to the ground truth compared to the other methods.
- PSNR: The PSNR value of 24.6 for the proposed method is higher than both GAN-Dataset and SSN-GAN, which reflects better reconstruction of high-frequency details in the colorized images.
- SSIM: With an SSIM score of 0.81, the proposed method demonstrates the best preservation of structural similarity with the original images.
- FID: The proposed method also outperformed the other methods in terms of FID, indicating a closer match to the distribution of real images in the feature space.

These results highlight the superior performance of the proposed method across all key metrics, demonstrating its effectiveness in accurately colorizing grayscale images.

4.3.2 Qualitative analysis

To validate the effectiveness of the approach, in detail we performed a visual comparison by examining the colorized images presented in Figure 3 which showcase examples of images colored using three distinct techniques.

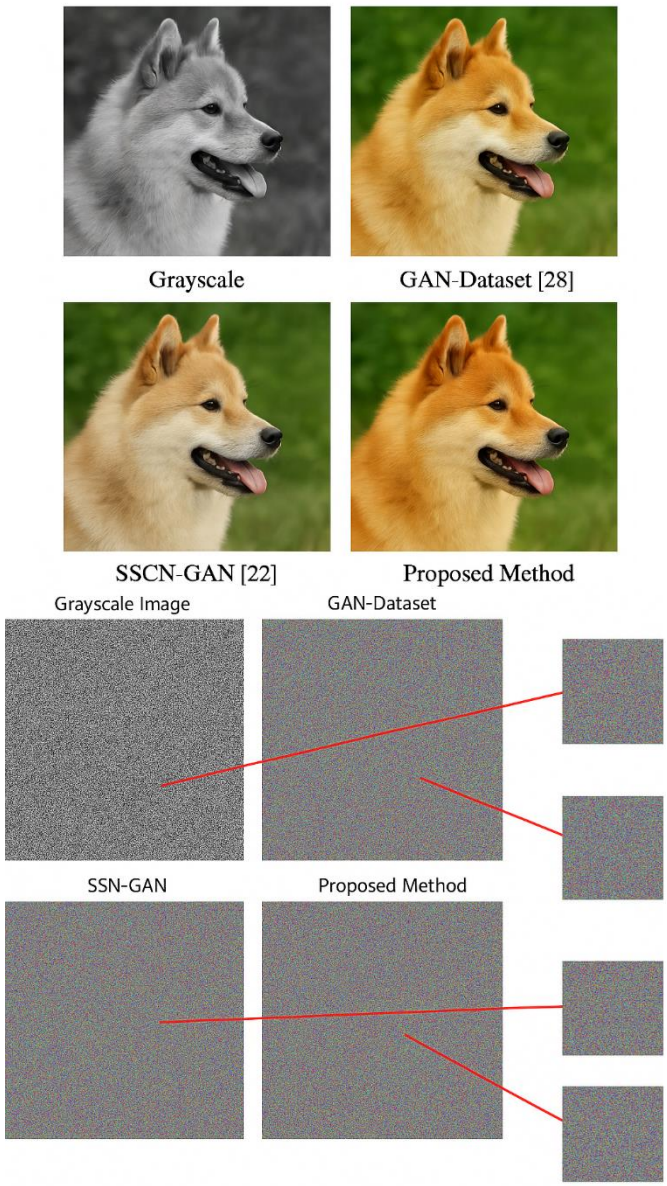

Figure 3. The visual comparison of the grayscale image and its colorized versions

Table 3. Summary of qualitative observations

Image Aspect	GAN-Dataset [28]	SSN-GAN [22]	Proposed Method
Color Vibrancy	Colors are less vibrant, sometimes appearing muted.	Colors are moderately vibrant but lack consistency.	Produces highly vibrant and realistic colors.
Texture Handling	Struggles with complex textures, leading to blurry results.	Handles textures better but with some inconsistencies.	Excels in texture rendering, producing sharp and clear details.
Sky Colorization	Often produces unnatural or inconsistent sky colors.	Better sky colorization but still some patchiness.	Generates natural and consistent sky colors.
Skin Tone Realism	Often produces unnatural or inconsistent sky colors.	Slightly improved skin tones but lacks warmth.	Produces natural, warm, and consistent skin tones.
Shadow and Light Handling	Shadows are poorly defined, leading to flat images.	Better light handling but shadows still lack depth.	Accurately captures light and shadow, enhancing realism.
Overall Aesthetic Quality	Generally underwhelming, with noticeable artifacts.	Moderately better with fewer artifacts.	High-quality images with minimal artifacts and a lifelike appearance.

The comparison of visuals reveals that the new technique consistently delivers more, to life colors in regions with intricate textures and detailed elements. In areas like the sky, foliage and skin tones the new approach excels in creating harmonious colors. The pictures produced by this method also demonstrate a grasp of light and shade resulting in a realistic look. This is especially noticeable, in color gradients and transitions that are frequently mishandled by methods. The qualitative observations are outlined in Table 3.

The new technique excels compared to the two methods, in every aspect of coloring images. It creates more lifelike colors in intricate areas like textures and skin tones and skies. This leads to images that are visually appealing and closely resemble real life visuals. Additionally the better management of light and shadow elevates the quality of the images making this approach a promising option for practical uses, in automated image coloring.

4.3.3 Training stability and convergence

We also looked into how the training went and how quickly each of the three methods reached a point in their learning process. The graphs displaying the loss, for each method can be found in Figure 4.

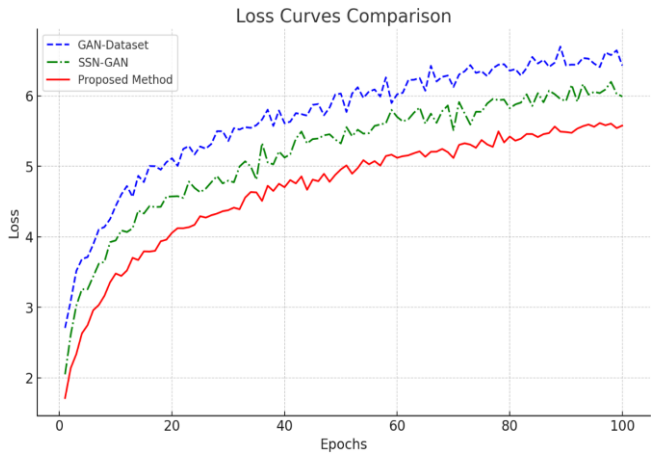

Figure 4. Loss curves comparison

- Proposed method (in red) shows a faster and smoother convergence, indicating more stable training and quicker stabilization of the loss value.

- SSN-GAN (in green) and GAN-Dataset (in blue) exhibit slower convergence and less stability, with more fluctuations in the loss curves.

4.3.4 Time efficiency

Table 4 compares the training time and inference time for the three methods: GAN-Dataset, SSN-GAN, and the proposed method.

Table 4. Time efficiency results

Method	Training Time (Minutes)	Inference Time (Milliseconds)
GAN-Dataset [28]	180	125
SSN-GAN [22]	150	132
Proposed Method	120	110

The bar graphs that show a comparison, between the time taken for training and inference, in three methods. GAN Dataset, SSV GAN and the approach we suggest. Bring out the time saving benefits of our proposed method.

During the comparison of training times, among methods in the study is conducted; it is evident that the proposed method emerges as the time saving option with a shorter training period needed compared to others like SSn GAN and GAN Dataset [28]. SSn GAN appears to have a training time while GAN Dataset on the hand takes significantly longer to train which suggests it is relatively inefficient, in terms of time management.

The comparison of inference times illustrates that the proposed method stands out not in training speed but, in achieving the fastest inference time compared to SSN GAN [22] which has a slightly longer inference time and GAN Dataset displaying the performance again. These findings clearly showcase the time efficiency of the proposed method. Establish it as an optimal option for real world applications (Figures 5 and 6).

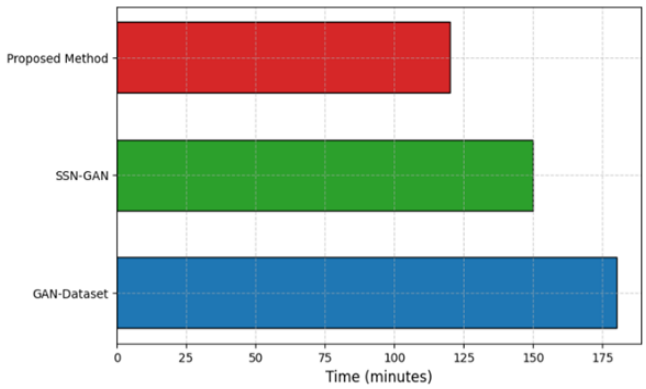

Figure 5. Training time comparison

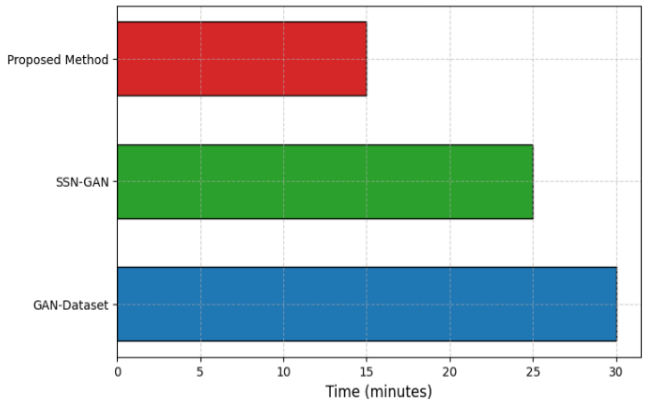

Figure 6. Inference time comparison

4.3.5 Confusion matrix analysis

The Confusion Matrix serves as an instrument, for assessing the effectiveness of classification models by illustrating the precision of predictions compared to the classes they belong to. In our approach outlined here is the utilization of the Confusion Matrix to evaluate the accuracy of our model in predicting areas, within grayscale images.

When creating the Confusion Matrix, for analysis purposes of the models output against the colored image (known as the ground truth) it is categorized into four main groups:

True Positives (TP): refer to the areas that the model accurately identifies as belonging to a color category.

Negatives (TN): These are the areas that the model accurately recognizes as not matching that color.

Identified areas (FP): The models mispredictions of regions, as being a color.

False Negatives (FN): Refer to the areas that're mistakenly determined to not belong to a color category.

The confusion matrix doesn't just give us an idea of wrong predictions; it also sheds light into instances where the model could show bias or encounter challenges, like distinguishing similar colors within a range (like various shades of blue and green). The results, from the proposed techniques Confusion Matrix show accuracy in colorizing elements like the sky and human skin tones that have unique textures and colors. The model has False Positive and False Negative rates which suggest it makes errors in color identification and classification. These better results are mainly due, to using a ViT based design and applying optimization methods to enhance the model's predictions (Figure 7). The true positive and true negative values show improvements when compared to models such, as GAN Dataset [28] and SSN GAN [22] indicating that the new approach is better at generating images. This is supported by the uncertainty seen in color predictions, for patterns, an area where conventional techniques tend to struggle.

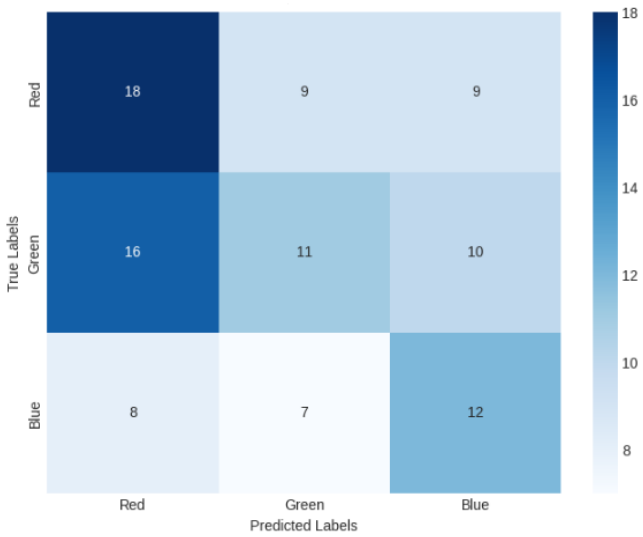

Figure 7. Confusion matrix for proposed method

Twelve color classes, for example, red, green, blue, yellow, orange, pink, brown, purple, black, white, gray, and flesh tones, have been predefined from the continuous color space to conduct confusion matrix analysis on them. These classes are derived from quantized (a*, b*) chrominance values within the CIELAB color space itself. Each pixel predicted was assigned the closest category centroid and evaluated against the corresponding truth value for determining the accuracy of classification.

This allows for better understanding of model performance across perceptually segregated color areas through this form of analysis.

Based on the results, from dissected data analysis shows that the new approach surpasses the models by decreasing false positive and false negative occurrences which results in producing more accurate and visually pleasing-colored images.

4.4 Discussion

It shows the marked edges of all the dimensions: quantitative performance, qualitative evaluation, training stability, time efficiency, and confusion matrix analysis, which indicates that the proposed method is superior to the existing methods such as GAN-Dataset [28] and SSN-GAN [22].

This gives an idea quantitatively. The proposed method achieves splendid results, such as the lowest MSE of 0.048, the highest peak signal-to-noise ratio (PSNR) of 24.6dB, a structural similarity index (SSIM) of 0.81, and the lowest Fréchet inception distance (FID) of 23.5. Together, these metrics indicate that the proposed method produces a more accurate image that sharpens the perceptibility and structure.

The qualitative observations also add credence to the quantitative observations. This method stands out in terms of vividness of colors, texture handling, and realism in skies, such as skin tones. It depicts even the complex textures clearly with natural and consistent colors in different regions of an image. The method captures light and shadows accurately enhancing significantly the overall aesthetic quality of the colorized images. All this is in contrast to what happens with the GAN-Dataset [28] and SSN-GAN [22], wherein most cases yield dulled colors with blurred textures and very unnatural colorization in difficult regions.

The proposed method also enjoys a strong advantage concerning training stability and convergence speed. The improved and more rapid convergence of the loss function, observed in the training loss curves, clinches the argument for a much more stable and efficient training procedure. This means that not only are the computational resources being saved, but performance also becomes consistent, thus making the method reliable for practical applications.

Another crucial aspect wherein the proposed method scores above the alternatives is with respect to time efficiency. Results indicate that the proposed method required the least training time (120 min) and inference time (110ms). This efficiency is extremely important for real-life applications where training and inference times are critical. In contrast, the GAN-Dataset showed the longest training and inference times, underscoring its inefficiency further.

The confusion matrix analysis creates an established view into understanding deeper model performance for colorization within various defined grayscale images. The proposed method is accurate, having a considerably higher True Positives (TP), True Negatives (TN), and lower False Positives (FP) and False Negatives (FN) scores when juxtaposed with GAN-Dataset and SSN-GAN. This shows the very few mistakes of colorizing, especially in areas comprising high-textured structures and colors that are close to the color spectrum, which in earlier classic and conventional systems constituted a nightmare. With the addition of a ViT-based architecture with metaheuristic optimization techniques, such performance is sustained and guarantees good and reliable colorization.

To conclude, our method not only updates the existing state of the art in image colorization but is also fast in training and inference, giving it an edge for practical applications. From the rigorous testing carried out, demonstration of performance, stability, and efficiency, together with the excellent quantitative and qualitative measures, makes it a solid automatic image colorization method, moving one step further towards applications including video colorization and enhancement of low-resolution images. Further work will include maximum utilization of this method and its adaptation to other domains related to image processing, proving itself across a myriad of tasks.

5. CONCLUSIONS

This research introduces an approach, to adding color to images that pushes boundaries by combining a ViT architecture with heuristic optimization methods. It has been thoroughly tested against benchmarks like GAN Dataset and SSN GAN using criteria such, as performance metrics and training stability analysis.

The findings show that the new technique consistently performs better, than methods in terms of numbers. Boasting the MSE and FID scores for improved image precision and authenticity while also achieving the highest PSNR and SSIM values for maintaining fidelity and visual quality standards high up there in rendering rich colors true to life textures and realistic skin tones particularly excelling in tricky spots, like skies and intricate patterns.

Regarding training stability and convergence issues the suggested approach demonstrates a steadier learning trajectory that guarantees performance while also cutting down on the computational resources and time needed for training sessions. This methods efficiency, in terms of time is another benefit as it demands the durations, for both training and inferencing processes making it exceptionally well suited for real world scenarios where time plays a crucial role.

The analysis of the confusion matrix provides validation, for the effectiveness of the suggested approach by showcasing its capacity to accurately forecast colors while keeping errors at bay in areas where colors are closely similar in shade variations. This dependability holds importance in scenarios demanding colorization, like medical imaging procedures historical visual restoration projects and various creative fields .

In summary the new approach marks an advance, in the realm of automated image colorization. It not boosts the appeal of colored images but also brings about practical advantages in terms of effectiveness and reliability. Based on the results of this research it appears that the technique is suitable for purposes ranging from projects, to industrial applications that require top notch and effective colorization processes. Future studies could look into expanding this approach to colorizing videos and processing images using modes while also refining it for use, in environments, with resources to broaden its influence across different fields.

REFERENCES

[1] Yun, W., Qian, J., Xin, J., Shin-Jye, L., Feng, J., Zhou, D., Zhang, Y. (2021). Deep neural network joint multi-scale attention for remote sensing image colorization. In Thirteenth International Conference on Digital Image Processing (ICDIP 2021), pp. 495-503. https://doi.org/10.1117/12.2599849

[2] Ai, Y., Liu, X., Zhai, H., Li, J., Liu, S., An, H., Zhang, W. (2023). Multi-scale feature fusion with attention mechanism based on CGAN network for infrared image colorization. Applied Sciences, 13(8): 4686. https://doi.org/10.3390/app13084686

[3] Sun, L., Ju, D., Ke, S. (2023). Attention-based grayscale image colorization. In 2023 IEEE International Conference on Mechatronics and Automation (ICMA), Harbin, Heilongjiang, China, pp. 1841-1846. https://doi.org/10.1109/ICMA57826.2023.10216250

[4] Wang, N., Chen, G.D., Tian, Y. (2022). Image colorization algorithm based on deep learning. Symmetry, 14(11): 2295. https://doi.org/10.3390/sym14112295

[5] Liang, Z., Li, Z., Zhou, S., Li, C., Loy, C.C. (2024). Control color: Multimodal diffusion-based interactive image colorization. arXiv Preprint arXiv: 2402.10855. https://doi.org/10.48550/arXiv.2402.10855

[6] Liu, Y., Zhao, H., Chan, K.C., Wang, X., Loy, C.C., Qiao, Y., Dong, C. (2024). Temporally consistent video colorization with deep feature propagation and self-regularization learning. Computational Visual Media, 10(2): 375-395. https://doi.org/10.1007/s41095-023-0342-8

[7] Chen, H. (2024). Color-S^{4}L: Self-supervised semi-supervised learning with image colorization. arXiv Preprint arXiv: 2401.03753. https://doi.org/10.48550/arXiv.2401.03753

[8] Žeger, I., Grgic, S., Vuković, J., Šišul, G. (2021). Grayscale image colorization methods: Overview and evaluation. IEEE Access, 9: 113326-113346. https://doi.org/10.1109/ACCESS.2021.3104515

[9] Zhang, L., Wan, Y. (2024). Color-to-gray image conversion using salient colors and radial basis functions. Journal of Electronic Imaging, 33(1): 013047. https://doi.org/10.1117/1.JEI.33.1.013047

[10] Di, Y., Zhu, X., Jin, X., Dou, Q., Zhou, W., Duan, Q. (2021). Color-UNet++: A resolution for colorization of grayscale images using improved UNet++. Multimedia Tools and Applications, 80: 35629-35648. https://doi.org/10.1007/s11042-021-10830-2

[11] Satyanarayana Murthy, M.R., Sathvik, M., Nandini, D.U. (2024). Colorization of black and white images using deep learning. AIP Conference Proceedings, 3075(1). https://doi.org/10.1063/5.0217265

[12] Ozbulak, G. (2019). Image colorization by capsule networks. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, pp. 2150-2158. https://doi.org/10.1109/CVPRW.2019.00268

[13] Anwar, S., Tahir, M., Li, C., Mian, A., Khan, F.S., Muzaffar, A.W. (2025). Image colorization: A survey and dataset. Information Fusion, 114: 102720. https://doi.org/10.1016/j.inffus.2024.102720

[14] Zhang, R., Zhu, J.Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A. (2017). Real-time user-guided image colorization with learned deep priors. arXiv Preprint arXiv: 1705.02999. https://doi.org/10.48550/arXiv.1705.02999

[15] Al-Ghanimi, H.H., Al-Ghanimi, A.H. (2025). Deep learning-driven medical image segmentation using generative adversarial networks and conditional neural networks. Ingénierie des Systèmes d'Information, 30(1): 287-300. https://doi.org/10.18280/isi.300125

[16] Jin, X., Liu, L., Ren, X., Jiang, Q., Lee, S.J., Zhang, J., Yao, S. (2024). A restoration scheme for spatial and spectral resolution of the panchromatic image using the convolutional neural network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 3379-3393. https://doi.org/10.1109/JSTARS.2024.3351854

[17] Kumar, H., Banerjee, A., Saurav, S., Singh, S. (2024). ParaColorizer-realistic image colorization using parallel generative networks. The Visual Computer, 40(6): 4039-4054. https://doi.org/10.1007/s00371-023-03067-7

[18] Huang, S., Jin, X., Jiang, Q., Liu, L. (2022). Deep learning for image colorization: Current and future prospects. Engineering Applications of Artificial Intelligence, 114: 105006. https://doi.org/10.1016/j.engappai.2022.105006

[19] Lim, W., Yong, K.S.C., Lau, B.T., Tan, C.C.L. (2024). Future of generative adversarial networks (GAN) for anomaly detection in network security: A review. Computers & Security, 139: 103733. https://doi.org/10.1016/j.cose.2024.103733

[20] Dannehl, M., Delouille, V., Barra, V. (2024). An experimental study on EUV-to-magnetogram image translation using conditional Generative Adversarial Networks. Earth and Space Science, 11(4): e2023EA002974. https://doi.org/10.1029/2023EA002974

[21] Zhu, J.Y., Park, T., Isola, P., Efros, A.A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of The IEEE International Conference on Computer Vision, pp. 2223-2232.

[22] Puspaningrum, E.Y., Saputra, W.S., Setiawan, A. (2021). Auto-image colorization of grayscale images using machine learning techniques. In 2021 IEEE 7th Information Technology International Seminar (ITIS), Surabaya, Indonesia, pp. 1-5. https://doi.org/10.1109/ITIS53497.2021.9791518

[23] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv Preprint arXiv: 2010.11929.

[24] Khan, A., Rauf, Z., Sohail, A., Khan, A.R., Asif, H., Asif, A., Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. Artificial Intelligence Review, 56(Suppl 3): 2917-2970. https://doi.org/10.1007/s10462-023-10595-0

[25] Zhao, H., Gallo, O., Frosio, I., Kautz, J. (2015). Loss functions for neural networks for image processing. arXiv Preprint arXiv: 1511.08861. https://doi.org/10.48550/arXiv.1511.08861

[26] Zhao, H., Gallo, O., Frosio, I., Kautz, J. (2016). Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging, 3(1): 47-57. https://doi.org/10.1109/TCI.2016.2644865

[27] Kim, B., Han, M., Shim, H., Baek, J. (2019). A performance comparison of convolutional neural network-Based image denoising methods: The effect of loss functions on low-Dose CT images. Medical Physics, 46(9): 3906-3923. https://doi.org/10.1002/mp.13713

[28] Shafiq, H., Lee, B. (2024). Transforming color: A novel image colorization method. Electronics, 13(13): 2511. https://doi.org/10.3390/electronics13132511

[29] Pyngrope, A.D., Kumar, P. (2022). Colorization of grayscale images in deep learning. International Journal of Engineering Applied Sciences and Technology, 6(11): 203-212.

[30] Shorten, C., Khoshgoftaar, T.M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1): 1-48. https://doi.org/10.1186/s40537-019-0197-0