February 2024

Master's Degree Thesis

A Study on Image Colorization using Transformers

Graduate School of Chosun University

Department of Information and Communication Engineering

Shafiq Muhammad Hamza

A Study on Image Colorization using Transformers

트랜스포머를 이용한 이미지 컬러화에 관한 연구

February 23, 2024

Graduate School of Chosun University

Department of Information and Communication Engineering

Shafiq Muhammad Hamza

A Study on Image Colorization using Transformers

Advisor: Prof. Bumshik Lee

This thesis is submitted to Chosun University in partial fulfillment of the requirements for a Master of Engineering degree

October 2023

Graduate School of Chosun University

Department of Information and Communication Engineering

Shafiq Muhammad Hamza

This is to certify that the master's thesis of
Shafiq Muhammad Hamza

has been approved by the examining committee for the thesis requirement for the master's degree in engineering.

Committee Chairperson: Prof. Goo-Rak Kwon G. R. Kwon

Committee Member: Prof. Jae-Young Pyun

Committee Member: Prof. Bumshik Lee

December 2023

Graduate School of Chosun University

Table of Contents

List of Figures

List of Tables

Abstract

A Study on Image Colorization using Transformers

Shafiq Muhammad Hamza

Advisor: Prof. Bumshik Lee

Department of Information and Communication Engineering

Graduate School

Chosun University

Automatic image colorization has attracted the attention of researchers due to its diverse applications. The colorization neural network colorizes given grayscale images by estimating the color values of the respective image. Conventional methods mostly use semantics or clues to colorize the image. However, color bleeding and unsaturated results are the problems of automatic colorization methods. Recently, transformer-based approaches have shown promising results in various image generation and processing tasks. In this thesis, an adversarial approach is proposed to colorize given grayscale images using transformers. Specifically, convolution in the transformer layer of the generator is introduced to extract both local and global information. In the proposed colorization architecture, light-weight multi-head self-attention is proposed in the transformer layer to better colorize the image and reduce the complexity of the self-attention layer. Furthermore, information from the encoder is passed to decoder layers for better reconstruction of the image.

Moreover, perceptual loss in conjunction with an adversarial loss is used for better visual quality. Finally, experiment results show that the proposed method shows superior results over state-of-the-art methods. In addition, the impact of different model architectures and hyperparameters is analyzed on the performance of the proposed colorization architecture. Overall, the thesis presents a comprehensive study on image colorization using transformers and proposes a novel approach that achieves state-of-the-art performance. The proposed approach has the potential to be applied in various real-world applications, such as image restoration and colorization in the film industry.

한글요약

트랜스포머를 이용한 이미지 컬러화에 관한 연구

샤피크 무함마드 함자
지도교수: 이범식
정보통신공학과
조선대학교 대학원

이미지 컬러화는 그 다양한 응용으로 인해 연구자들의 만은 관심을 끌고 있다. 컬러화 신경망 네트워크는 각 이미지의 색상 값을 추정하여 그레이스케일 이미지를 컬러화하는 역할을 수행한다. 기존 방법은 대부분 의미론이나 단서를 사용하여 이미지를 색상화하는 방법을 사용하고 있으나 이러한 기존 컬러화 방법은 색상 블리딩 및 저채도 영상의 결과를 생성하는 문제점을 가지고 있다. 최근, 트랜스포머 기반 접근법은 다양한 이미지 생성 및 영상처리 분야에서 좋은 성능을 보여 주고 있다. 본 학위 논문에서는 트랜스포머를 사용하여 주어진 그레이스케일 이미지를 색상화하는 적대적 접근법을 제안한다. 특히, 제안 방법에서는 지역 및 전역 정보를 모두 추출하기 위해 생성기의 트랜스포머 레이어에 컨볼루션을 수행한다. 또한 경량 멀티헤드 셀프-어텐션이 트랜스포머 레이어에 적용되어 이미지를 컬러화를 보다 잘 수행하고 셀프-어텐션 레이어의 복잡성을 감소시킨다. 트랜스포머 레이어의 사용으로 인해

네트워크는 자연스럽고 생생한 컬러화 결과를 생성합니다. 또한 제안 방법에서는 더 나은 시각적 품질을 얻기 위해 적대적 손실과 함께 지각 손실을 사용한다. 실험 결과는 제안 방법이 최근 발표된 기존 방법보다 우수한 결과를 얻었다는 것을 보여주고 있다. 또한 실험에서 다양한 모델 아키텍처와 하이퍼파라미터가 제안하는 컬러화 신경망 네트워의 성능에 미치는 영향을 분석하였다. 본 학위 논문에서는 트랜스포머를 이용한 이미지 컬러화에 대한 포괄적인 연구 및 최첨단 성능을 달성하는 새로운 접근법을 제안한다. 제안된 접근 방식은 영화 산업의 이미지 복원 및 색상화와 같은 다양한 실제 응용 프로그램에 폭넓게 적용될 수 있다.

1. Introduction

Image colorization is the process of adding color to grayscale images. This is an important task in various fields such as film and video production, as well as for personal use in preserving and restoring old or legacy photos. With the advancements in deep learning, neural network-based approaches have become popular for image colorization due to their ability to learn complex features and patterns in the data.

One of the most significant applications of image colorization is in the restoration and preservation of legacy photos. Legacy photos as shown in Fig. 1-1, are often black-and-white or sepia-toned, and may have degraded over time due to factors such as fading, scratches, or discoloration. These photos can hold significant sentimental or historical value, and colorizing them can breathe new life into them, allowing viewers to see them in a new light.

Image colorization can be a challenging task, especially for legacy photos where the colors are not known. Deep learning-based approaches have shown significant improvements in image colorization accuracy and can produce realistic and vibrant colorizations. By applying these approaches to legacy photos, it is possible to bring out the colors and details that were not visible before, and create a more immersive and engaging viewing experience.

Migrant Mother (1936)

Betty Grable (1943)

Carrying Cranberries (1911)

Che Guevara (1960)

Abraham Lincoln (1860)

Winston Churchill (1941)

Dovima with Elephants

Couple in Raccoon Coats (1932)

The Iconic V-J Day (1945)

Figure 1-1. Famous grayscale legacy images

In addition to personal use, image colorization can also have practical applications in fields such as film and video production. Colorization can be used to add color to old movies or TV shows, providing a more modern and enjoyable viewing experience for audiences. It can also be used to create realistic visual effects, such as adding color to a grayscale image of a product or object, which can help enhance the overall visual appeal.

Overall, image colorization is a valuable technique for a wide range of applications, including the restoration and preservation of legacy photos. With the advancements in deep learning, the accuracy and quality of colorization techniques have significantly improved, allowing for more realistic and engaging colorizations.

In this chapter, Section 1.1 presents the overview motivation for proposed research in image colorization, Section 1.2 presents the research objective and major contributions of the thesis. Section 1.3 explains the outline of the thesis.

1.1 Overview and Motivation

Image colorization aims to restore fully colored images from the given black-and-white images. Colorization has a wide area of applications as there are various legacy black-and-white images from past times. Colorization can be used for cartoon colorization [1-2], restoring old photos [3], fake color detection [4], and even assisting in classification and segmentation [5]. However, object colors can vary a lot, i.e., the same object can have different colors, like leaves can be red or green. Hence, colorization is a highly ill-posed problem as it is difficult to assign colors based on intensity values only. Assigning a proper color to each object in an image is an open research problem. The problems in automatic image colorization are color bleeding, object intervention, and semantic confusion, etc.

Recently, multiple methods have been proposed to address the problem of image colorization. These methods can be divided into two categories: 1. User-guided colorization 2. Automatic colorization. The user-guided colorization requires user intervention to assign colors to objects in an image. It is labor intensive task requiring human intervention to assign colors. In user-guided colorization, assigned colors depend upon user selection, making this method less prone to errors. There are two types of user hints, which are scribble and reference images. Scribble-based methods use user hints in kind of color scribes to colorize objects according to that color. Example-based methods use reference images similar to the input grayscale image and transfer color to the input image. Both methods require human intervention to colorize. On the other hand, automatic colorization methods do not require user intervention and can learn to generate a color image using a deep learning model. These methods learn the end-to-end mapping of the grayscale image to the color image. Deep learning methods for colorization have become popular recently due to their efficacy and a large number of datasets publicly available, such as ImageNet [6] and Places [7] which have 1.3 million and 1.8 million images, respectively.

Although automatic colorization systems achieve better results, the problems of unnatural color and color bleeding still exist. To address these problems, semantic clues and segmentation information are added to the colorization network. It is not easy to be applied in every situation even though it solves the problem to some extent, and network complexity is still very high.

The motivation for this thesis comes from the need for a more accurate approach to image colorization. To this end, a transformer-based image colorization network, named ColorFormer is proposed.

Overall, this thesis aims to propose and evaluate a novel approach to image colorization using transformers, which provides more accurate and natural colorization results while achieving state-of-the-art performance on benchmark datasets.

In striving for progress, efforts are being made to enhance techniques for image colorization. The goal is to contribute to the development of more accurate and effective methods, with potential applications in fields such as film and video production, personal use, and beyond.

1.2 Research Objective

To resolve the above-mentioned problems, a transformer based image colorization method named ColorFormer is proposed. The proposed network is designed based on the combination of convolution and transformer layers to learn local as well as global information. It is a kind of Conditional Wasserstein Generative Adversarial Network (CWGAN) based approach that uses features of the Wasserstein Generative Adversarial Network (WGAN) [8] and Conditional Generative Adversarial Network (CGAN) [9] to achieve better image colorization. The main advantage of the proposed ColorFormer is that it can capture local as well as global information. The feed-forward network consists of convolution as well as linear layers. Therefore, it can capture long-range dependencies as well as local information. Another feature of the proposed network is that a window-based method is used inside the transformer layer to reduce the complexity of the network.

The main contributions of this thesis are as follows

1) ColorFormer, a novel colorization method based on the transformer architecture and Conditional Wasserstein Generative Adversarial

Network (CWGAN) is proposed, addressing the limitations of colorization techniques.

2) The proposed ColorFormer Block integrates the convolution layers with the transformer architecture which enables the effective extraction of both local and global information, resulting in higher-quality and visually appealing colorized images.

3) Lightweight multi-head self-attention is proposed to enhance the colorization process while significantly reducing the complexity of self-attention in the ColorFormer Block.

4) Extensive experiments and evaluations demonstrate that the proposed ColorFormer approach outperforms existing state-of-the-art colorization methods, showcasing its ability to generate high-quality, visually appealing colorized images from grayscale inputs.

5) The thesis provides a thorough analysis of the proposed model, including comparisons to existing techniques, ablation studies, and discussions on the impact of different components on the overall performance. This in-depth investigation contributes to a deeper understanding of the factors that contribute to the success of ColorFormer.

1.3 Thesis Layout

This thesis is organized into six chapters, each focusing on a specific aspect of the research. The structure of the thesis is as follows:

Chapter 1: Introduction - This chapter provides an overview of the research background, motivation, problem statement, and objectives. It highlights the main contributions of the thesis and outlines the proposed ColorFormer approach to image colorization.

Chapter 2: Related Works - This chapter presents a comprehensive review of existing image colorization techniques, including deep learning-based methods, Transformer-based architectures, and Generative Adversarial Networks (GANs).

Chapter 3: Proposed Method - This chapter delves into the details of the proposed ColorFormer, explaining each component, including data preprocessing, ColorFormer, the Convolutional Wasserstein Generative Adversarial Network (CWGAN), and post-processing. The proposed window mechanism and its integration with the self-attention mechanism are also described in this chapter.

Chapter 4: Experimental Results - This chapter presents the experimental setup, dataset, and performance evaluation of the proposed approach. It covers both quantitative and qualitative results, as well as comparisons with state-of-the-art methods.

Chapter 5: Ablation Studies - This chapter conducts a series of ablation studies to analyze the impact of various components of the proposed approach on its performance. It includes discussions on the role of the Transformer architecture and CWGAN, sensitivity analysis of model hyperparameters.

Chapter 6: Conclusion - The final chapter summarizes the main findings of the research, emphasizing the contributions to the field of automatic image colorization. It also outlines future research directions and potential improvements to further enhance the performance and applicability of the proposed approach.

2. Related Works

2.1 User-guided Colorization

Early works in colorization were mostly user-guided, in which the user provides a hint to colorize pixels, and they can be further categorized into scribble- or example-based approaches. In the scribble-based method, a user provides high-level scribbles. Then colors propagate based on low-level similarity matrices. For example, the early method assigns a similar color to pixels with the same luminance [11]. This is a very labor extensive task and requires accurate scribbles for good colorization. In [12], the colorization using the edge detection technique was proposed. Luminance-weighted chrominance bleeding was proposed in [13] for fast colorization. Zhang et al. [24] proposed a method with additional deep prior from CNN to ensure colorization without giving scribbles. However, these methods might result in color bleeding as pixels with similar intensity produce similar colors. In the example-based methods, an image is provided as a reference to colorize the grayscale image. In [14], it was proposed to extract luminance values and match them with the grayscale image to transfer color. In [15], the reference source image is segmented, and color information from the segmented image is used as a scribble in [11] to transfer colors. Moreover, an automatic reference image retrieval method was proposed to reduce the effort of selecting a reference image [17]. However, these methods highly depend upon the reference image and provide unnatural results if the semantics of the reference image do not match with the input grayscale image. In [18] and [19], colorization methods were proposed to match the semantics for the reference image and target image. In addition, authors use different types of references like words [20-21] and sentences [22]. Although these methods improved over the years, it still required user interference, and results depend on the provided information.

2.2 Fully Automatic Colorization

Fully automatic methods generally use deep learning-based structures to learn semantic information for colorizing grayscale images. In [23], using CNN to color images fully automatic manner was first tried, where they used patches to colorize images with a simple model architecture. A class rebalancing scheme to resolve the inherent ambiguity and multimodal nature of colorizing grayscale images was proposed in [24], where the VGG-style network is adopted to colorize images. In [25], the network is trained jointly for classification and colorization using a labeled dataset. The VGG network is used to augment input grayscale images in [16] and pass-through CNN networks. However, these methods still have problems such as color bleeding and semantic confusion. In [26-27], additional semantic information was used to resolve these problems. They achieved notable results in generating more contextually accurate colorizations by leveraging semantic information.

However, despite their achievements, these approaches still faced limitations such as color bleeding and unsaturated results, which led to less visually appealing colorized images. Another disadvantage is their reliance on semantic information, which might not always be available or accurately estimated. This dependency can limit the applicability of these methods.

Moreover, generative adversarial networks (GAN) [28] have recently become popular. These generative models help in multimodal colorization. An image-to-image translation model is proposed using conditional GAN in [29], where a U-Net-based generator was used, resulting in more vivid colorized resultant images owing to adversarial training. The model is generalized to high-resolution images in [30]. The input noise is sampled at various times for diverse colorization [31]. Grayscale images are mapped to GMM using a mixture density network [32]. Moreover, class distribution is added to the WGAN model as proposed in [33]. In [35], more focuses on colorization using generative priors given the spatial structure of an image have been made.

Transformers [48] have been getting attention in the computer vision domain. The transformer architecture was first introduced by Vaswani et al. [48]. Later, a new architecture for image classification was proposed using the transformers, called Vision Transformers (ViT) [49]. The ViT architecture divides the input picture into a grid of patches, which are subsequently processed by the transformer network to perform classification tasks. In addition to image classification, the transformers have also been applied to other image processing tasks, such as object detection, segmentation, image super-resolution, denoising, and colorization. Furthermore, the transformers show promising results in image-to-image translation, such as ColTran [34], demonstrating their effectiveness in this area.

Current CNN-based and Transformer-based colorization networks face several limitations, including color bleeding, unsaturated results, and difficulties capturing both local and global features. To address these problems, ColorFormer, a novel method that combines a CWGAN framework with transformers and convolutional layers is proposed for improved color consistency and stable training. Lightweight multi-head self-attention is introduced to enhance colorization. The proposed approach ultimately achieves superior-quality, visually pleasing colorized images that outperform existing methods.

3. Proposed Method

Given a grayscale image $x^{g} \in \mathbb{R}^{H\times W\times 1}$, the goal is to predict the other two color channels a and b, representing the chrominance in CIELAB color space (also called LAB) [51], which is a color space designed to represent all colors visible to the human eye. The $L*$ channel represents the lightness or brightness of the image, and the $a*$ and $b*$ channels represent the color information. The a* channel ranges from green to red, and the $b*$ channel ranges from blue to yellow. $x^{Lab} \in \mathbb{R}^{H\times W\times 3}$ is the original color image with L and ab , which are the luminance and the chrominance channels, respectively. The use of the CIELAB color space in colorization gives more precise control over the colorization process and a more accurate representation of the color. The objective of the colorization network is to generate natural and realistic colors for input grayscale images. Fig. 3-1. shows an overview of the proposed ColorFormer architecture. As shown in Fig. 3-1, given an input image $x^{in} \in \mathbb{R}^{H\times W\times 3}$, the image is first converted to CIELAB color space and split into L and ab channels. L channel image $x^{L} \in \mathbb{R}^{H\times W\times 1}$ is passed through the generator, and the generator outputs ab channel image $y^{ab} \in \mathbb{R}^{H\times W\times 2}$, having color information. This image is passed through the discriminator along with the real image, and the discriminator evaluates the quality of the generated image. The model is trained in an adversarial manner.

The ColorFormer structure is based on the GAN architecture that uses transformer layers in the generator for natural and diverse colorization. The network consists of a generator and a discriminator. The generator is based on ColorFormer Blocks, which consist of self-attention, feed-forward network, normalization layer, and convolution layers. The discriminator is based on the Markovian discriminator architecture (PatchGAN) [29], which focuses more on capturing high-frequency components using local patches.

The GAN architecture with the ColorFormer network is designed based on two key concepts; Conditional GAN (CGAN) architecture and Wasserstein GAN (WGAN)., which is called Conditional Wasserstein GAN (CWGAN). The motivation for using CWGAN is to improve colorization accuracy and stability. Conditional GAN incorporates additional information, such as grayscale images, resulting in more accurate and visually appealing colorization results. However, Wasserstein GAN addresses some common issues in traditional GAN, such as instability and mode collapse. Overall, CWGAN produces more realistic and high-quality colorizations by leveraging the strengths of both Conditional and Wasserstein GAN. Instead of conventional adversarial loss, earth mover distance-based objective function (Wasserstein distance) [8] for GAN is used, which improves the overall stability and convergence of the model.

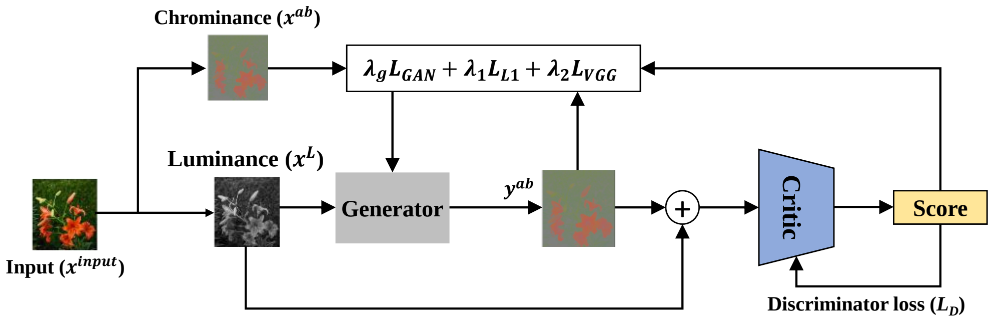


Figure 3-1. Overall architecture of the proposed colorization network

The overall architecture of the proposed method is based on CWGAN. In this section, generator and critic architecture is introduced and the objective function used for training the generative adversarial network is presented.

3.1 Proposed Generator Architecture:

Generator architecture is divided into three parts: encoder, decoder, and latent space. Generator architecture is based upon transformer layers which use the windows mechanism to reduce complexity. The generator part consists of ColorFormer Blocks and convolutional layers. Convolutional layers are used to downsample the features of the image. Firstly, the input image is passed through the input projection layer, which consists of convolution and activation. ReLU [57] is used as activation in the input projection layer. Then the image is flattened and passed through the ColorFormer layer, which consists of depth-wise convolution, layer normalization (LN), lightweight multi-head self-attention (LW-MHSA), and color feed-forward network

(CFFN). Depth-wise convolution (DWS) is used to extract local information, which is very helpful for colorization as the color of neighboring pixels is dependent on each other. ColorFormer Block uses lightweight multi-head self-attention. The window mechanism is used to reduce the complexity of the model. In lightweight multi-head self-attention, the heads are split between the shifted windows. In a lightweight multi-head self-attention layer with a cyclic shift size of 1, the input can be divided into two window configurations, with one using the original window and the other using a cyclically shifted window. This allows N/2 attention heads to use the original window and the remaining N/2 heads to use the shifted window. Since the shifted window returns to its original position after one shift, this approach enables the model to capture both local and global information. The final output of the self-attention layer is obtained by adding the results, as shown in Fig. 3-3. This reduces the complexity of the network by getting better long-range dependencies. Finally, the input of the ColorFormer Block is added to the output using the residual connection to compensate for the missing information.

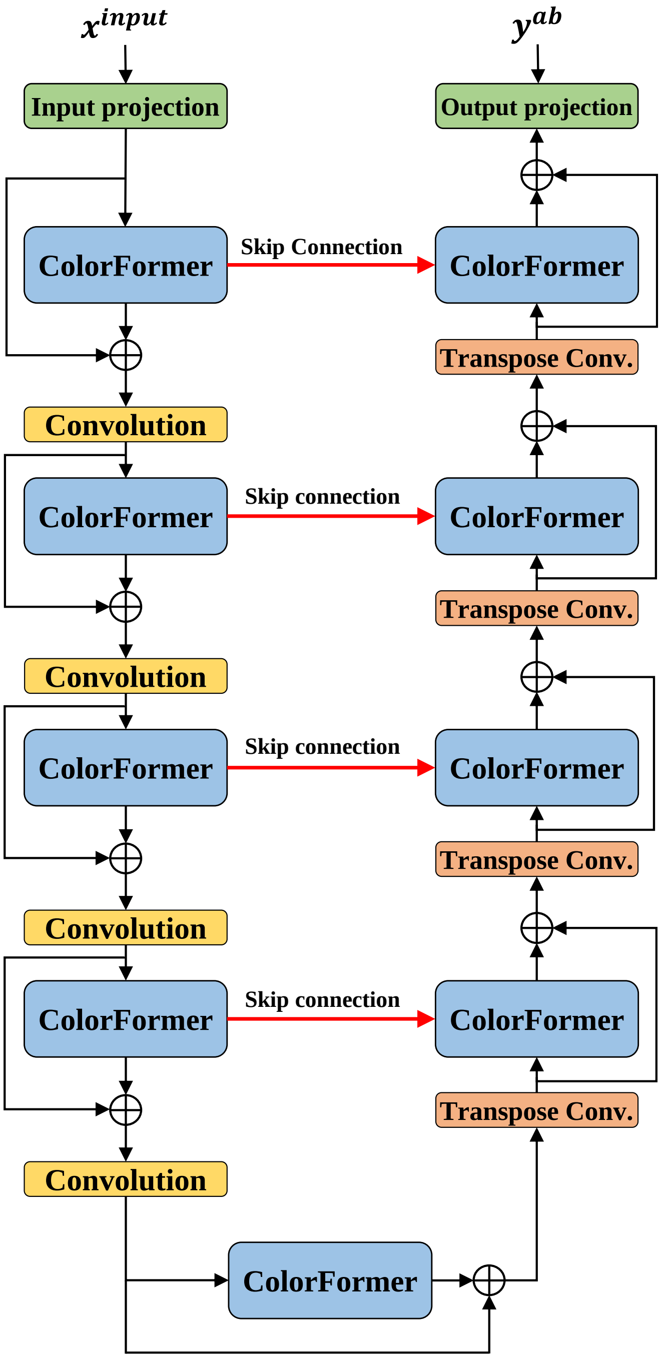


Figure 3-2. Generator architecture in the proposed colorization network

3.1.1 Proposed ColorFormer Block

Self-attention in vision transformers captures long-range dependencies. The standard transformer calculates the relationship of each token with all other tokens in the feature map. Since this results in quadratic computation complexity, it is not suitable for applying global self-attention on high-resolution feature maps. Moreover, although colorization requires local and global information due to the fact that the color of a pixel in an image can depend on both its local context, such as the colors of adjacent pixels, and its global context, such as the overall color distribution of the image, transformers are more biased toward long-range dependencies.

To resolve these issues, ColorFormer Block is proposed that can capture both local and global information with less computation complexity, where the proposed ColorFormer Block includes self-attention to capture global information and convolutions for local contextual information. Fig. 3-2 shows the architecture of the proposed ColorFormer Block. As shown in Fig. 3-2, the ColorFormer Block consists of three components: 1) depth-wise convolution (DWC), 2) lightweight multi-head self-attention (LW-MHSA), 3) color feed-forward network (CFFN). The input is passed through the DWS which is used to capture local information. The output of the DWS is added element-wise to the input features and then passes to layer normalization (LN) before self-attention. The output from LN is passed through the

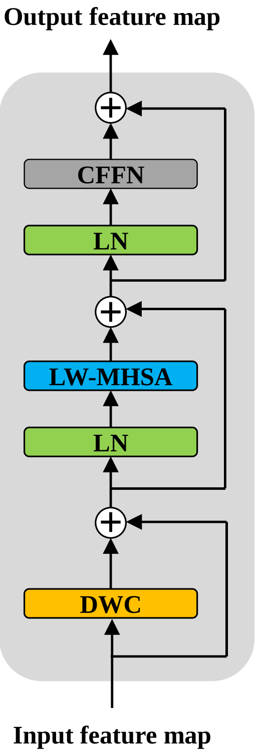


Figure 3-3. Proposed ColorFormer block

lightweight multi-head attention and then performed element-wise addition with input features for LN. Finally, the features are first passed through an LN step before being processed by the Color Feed Forward (CFFN). The output of the CFFN is then added element-wise to the input features. Overall, the ColorFormer block is designed to process input features in a hierarchical manner, where each layer processes the intermediate output from the previous layer to capture increasingly complex representations of the input data. The use of DWC, LWMHSA, and CFFN allows the network to capture both local and global information in the input features. The overall process can be expressed by (1), (2) and (3).

$$x^{l-1} = DWC(x^{l-1}) + x^{l-1} \tag{1}$$

$$x^{l} = LWMHSA\Big(LN(x^{l-1})\Big) + x^{l-1} \tag{2}$$

$$x^{l} = CFFN\Big(LN(x^{l})\Big) + x^{l}\ , \tag{3}$$

where *x* is the input feature map to the ColorFormer Block, DWC is a function of the depth-wise convolution layer, LWMHSA is the lightweight multi-head self-attention, LN is the layer normalization, and CFFN represents the color feed-forward network.

Lightweight multi-head self-attention: Given a feature map **X**, it is split into non-overlapping patches of size $M \times M$.

$$X = \{X^1, X^2, \dots, X^n\}\ , \tag{4}$$

where *n* is the number of patches.

Then split each patch into windows of size $W \times W$ and flattened the windows to pass through self-attention. To capture dependencies outside the window, shifted window mechanism is used. Windows is shifted in a cyclic manner and split between the number of heads of self-attention. If the shift size is 1 i.e., the window returns to the same position after the 2-nd shift, then there are two window configurations. The *N*/2 heads of self-attention are given with one window configuration and the other with a different window configuration. Fig. 3-4 shows the windows mechanism of the proposed LW-MHSA. As shown in Fig. 3-4, half the heads are given with one window configuration and another half with shifted window, and after attention, both are added to get the final output of self-attention. The overall process of proposed self-attention can be defined as follows

$$\{X_s^n, X_{s+1}^n, \dots, X_{s+w}^n\}, \tag{5}$$

where $X_s^n$is the window configuration for the *n*-th patch, and *w* is the number of shifted window configuration for the *n*-th patch.

$$Y_j^i = ATN_k\left(X_s^i W_j^Q, X_s^i W_j^K, X_s^i W_j^V\right) + \\ ATN_k\left(X_{s+1}^i W_j^Q, X_{s+1}^i W_j^K, X_{s+1}^i W_j^V\right) + \cdots, \\ i, j \in \{1, 2, \dots, N\}, \tag{6}$$

where $Y_j^i$is the output of LW-MHSA for the *i*-th patch, $W^Q, W^K, and\ W^V$are the projection matrices of the query, key and value for the single head, respectively, *k* (= $floor(N/(w+1))$ is the number of attention heads that can use a single window configuration is equal to the total number of attention heads (*N*) divided by the number of windows (*w*) that can be created from the input sequence. i.e., if the number of windows (*w*) is 2 and the number of heads (*N*) is 16, then *k* will be 8 (8 heads for each window configuration), *s*+1 is the shifted window

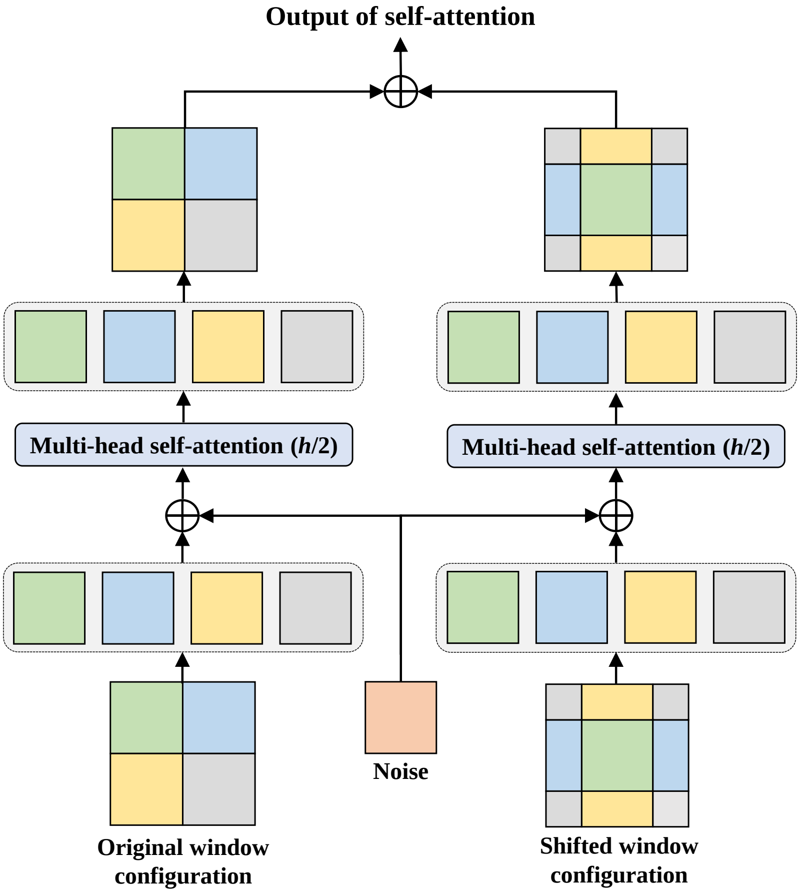

Output of self-attention
Multi-head self-attention (*h*/2)
Multi-head self-attention (*h*/2)
Noise
Original window configuration
Shifted window configuration

Figure 3-4. Proposed Window Mechanism in LW-MHSA

and ATN_k is the function of the attention module for the *k*-th number of heads.

$$\hat{X}_j \; = \{Y_j^1, Y_j^2, \ldots, Y_j^n\}, \tag{7}$$

where $\hat{X}_j$ is the output of LW-MHSA after combining all patches.

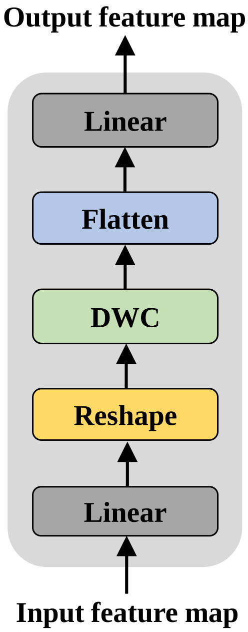


Figure 3-5. Proposed Color Feed-Forward Network

The total number of heads of self-attention is split between windows configurations. The relative position encoding is added in the attention module inspired by previous works [36-37].

$$Atn\ (Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V, \tag{8}$$

where *B* is the relative position bias term and *Q*, *K*, and *V* are the query, key, and values, respectively, and Softmax is a softmax function [56].

Color feed-forward network (CFFN): CFFN block consists of linear and convolution layers. According to [38-39], the feed-forward network of transformers suffers from capturing local contextual information.

In colorization, neighboring pixel information is crucially important. To address this issue, convolution layers is added in the feed-forward network as in [39-42]. In CFFN, the first linear layer is applied, and reshape the features to apply convolution. Depth-wise convolution enables the CFFN to capture local features. After convolution layers, features are flattened and passed through the linear layer. GeLU [43] activation function is used in each layer.

Convolution layers are applied in between ColorFormer Blocks to downsample and upsample the image. Moreover, the decoder part recovers the spatial information, and skip connections are used from the encoder to pass information for better recovery of spatial information. The shape of the overall generator architecture is similar to U-Net [44]. The output of the generator is two channel image $Y_G \in \mathbb{R}^{H \times W \times 2}$ with only color information. This image is passed to the discriminator for predicting the score.

3.2 Discriminator Architecture

The Critic of the proposed method is based on Markovian discriminator architecture (PatchGAN) [29], where the PatchGAN discriminator keeps track of the high-frequency structure of the image generated by the generator. PatchGAN discriminator uses local patches in the image to determine generated image is real or fake. Critic produces high scores for input and ground truth (GT) pair and low scores for input and generated pair. Discriminator is conditioned on a grayscale image and gets both luminance and chrominance channels as input.

3.3 Objective Function

The objective function of the proposed architecture is defined as

$$L_{total} = \lambda_g L_{GAN} + \lambda_1 L_{L1} + \lambda_2 L_{VGG}, \tag{9}$$

where L_{GAN} , L_{L1} and L_{VGG} represents the GAN, L1 and VGG losses, respectively and λ_g, λ_1 and λ_2 are the hyperparameters. The first term $\lambda_g L_{GAN}$ of (9) denotes the adversarial loss for GAN training. The WGAN loss function is used and defined as

$$L_D = E_y[\,D(y,x)] - E_{\tilde{y}}[\,D(\tilde{y},\,x)] + \lambda\,\times GP \tag{10}$$

$$L_G = -\,E_{\tilde{y}}[\,D(\tilde{y},x)]\,, \tag{11}$$

where L_D and L_G represent the loss for discriminator and generator of WGAN [8], respectively. The x, y and $\tilde{y}$ are the input grayscale image, GT and generated chrominance image, respectively. The GP is the gradient penalty used for stable training of GAN. It is known that WGAN loss offers better properties as compared to other GAN losses and resolves the problem of vanishing gradient and achieves stable training of GAN [8].

L_{L1} is the pixel-wise loss function which is defined as

$$L_{L1} = ||y - \tilde{y}||_1, \tag{12}$$

The VGG loss function is also used for better perceptual quality of generated images. VGG loss function in (9) is defined by the rectified linear unit activation layer of the pretrained VGG network.

$$L_{VGG} = ||\varphi_k(y) - \varphi_k(\tilde{y})||_2^2 \ , \tag{13}$$

where φ_k is the features of the k-th layer of the pretrained VGG network. VGG loss is used in measuring the semantic similarity of generated and ground truth images. Hence, total losses for the generator and discriminator can be expressed as (14) and (15), respectively, by rewriting (10) and (11).

$$L_G = -\,E_{\tilde{y}}[\,D(\tilde{y})] + \lambda_1 L_{L1} + \lambda_2 L_{VGG} \tag{14}$$

$$L_D = E_y[\,D(y)] - E_{\tilde{y}}[\,D(\tilde{y})] + \lambda\ \times GP\ , \tag{15}$$

4. Experimental Results

4.1 Implementation Details

The publicly available PASCAL-VOC dataset with 17,125 images is used for the experiments. Images are resized to 256×256 using bilinear interpolation. In the experiment, it was observed that resizing is better than random cropping, as cropping can result in a negative effect on learning colors. Images are normalized to the range [-1, 1]. Images are divided into l and ab channels and used as input and output during training.

An ADAM optimizer is used with learning rates of 1×10^{-4} and 2×10^{-4} for the generator and discriminator, respectively. The size of the embedding dimension was set to 16, and exponential decay rates β_1 and β_2 values in ADAM optimizer [52] were set to 0.5 and 0.999, respectively. The generator and discriminator of ColorFormer are trained until the network converges. The hyperparameters λ_g, λ_1 and λ_2 in (9) were empirically set to 0.5, 100, and 1000, respectively. The network was implemented in the Pytorch framework with GeForce RTX3090 GPU.

4.2 Quantitative Metrics and Comparisons

The Peak Signal-to-Noise Ratio (PSNR) [53] and Structural Similarity Measure (SSIM) [54] are used to measure the quality of colorized images with respect to GT Images. PSNR and SSIM are popular metrics for evaluating the performance of colorization. The PSNR measures the difference between the original image and the processed image in terms of the ratio of the maximum possible power of a signal to the power of corrupting noise. SSIM, on the other hand, measures the structural similarity between two images by considering the luminance, contrast, and structure of the images. Higher PSNR and SSIM values generally indicate better image quality and closer similarity to the original image. The colorfulness metric [55], which evaluates the amount of color variation in an image, is used to evaluate the quality of the colorized images. The colorfulness metric [55] is based on the standard deviation of the chrominance channels (a* and b*) of the image and used to quantify the overall amount of color in the image. The standard deviation represents the degree to which the chrominance values in the image vary from their average value. A higher standard deviation indicates a greater range of chrominance values and, thus, a more colorful image. △Colorfulness measures the difference in colorfulness values between colorized and GT images. The proposed method is also compared with fully automatic state-of-the-art colorization methods, including CIC[24], Deoldify[47], BigColor[35], InstColor[46], ChromaGAN[33], ColTran[34] and CT2[50]. Table 4-1. shows the quantitative results and comparisons. As shown in Table 4-1, the proposed method achieves significantly higher PSNR and SSIM values than the state-of-the-art methods. Fig. 4-1 shows the visual results of the colorization for the proposed and other methods. As shown in Fig. 4-1, the proposed method shows more natural colorization than others. For instance, the horse color (column 2) of CIC [24] looks unnatural and reddish compared to the results of ColorFormer. In addition, the proposed method does not produce rare colors, and output images are close to ground truth. While BigColor [35] demonstrates more saturated results in image colorization, the method exhibits a notable drawback in generating unnatural outputs that deviate from the true colors of the ground truth (GT) images. Coltran [34], a transformer-based colorization network, shows desaturated results with bleeding artifacts. On the other hands, the result images of the proposed method might not achieve higher colorfulness values since it is likely to encourage rare colors. However, the proposed method shows the colored images more similar to GT with higher PSNR and SSIM values, which indicates that the proposed method is successful in reproducing the original colors accurately. Although Coltran[34], BigColor[35], and CT2[50] produce rare colors, they show bleeding artifacts and unnatural colorization. The △colorfulness values obtained in the experiments indicate that the proposed method produces color variations closely aligned with the ground truth (GT) images. This demonstrates the effectiveness of the proposed approach in generating more accurate and natural

Table 4-1. Quantitative comparisons

Models	**PSNR (dB)**	**SSIM**	**Colorfulness**	**△Colorfulness**
CIC [24]	21.000	0.925	30.43	2.55
Deoldify [47]	22.972	0.911	16.60	16.38
BigColor [35]	21.473	0.883	35.71	2.73
InstCol [46]	22.911	0.910	22.21	10.77
ChromaGAN [33]	23.636	0.882	21.89	11.09
ColTran [34]	23.839	0.868	35.74	2.76
CT2 [50]	19.304	0.912	**36.04**	3.06
ColorFormer	**24.518**	**0.943**	31.38	**1.60**

colorizations, thereby enhancing the overall quality and visual appeal of the colorized images. Overall, the proposed network achieves more natural and consistent results.

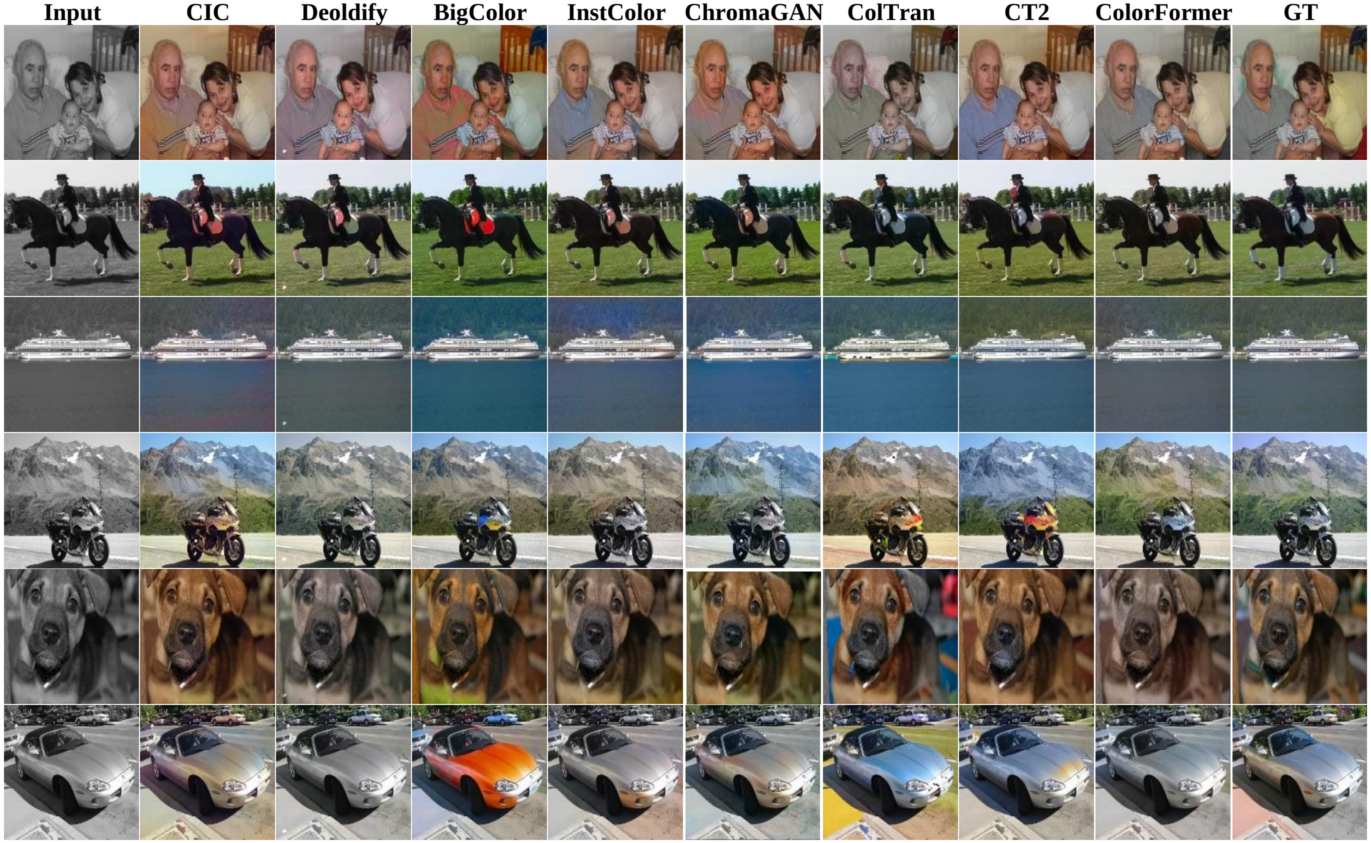


Figure 4-1. Visual comparisons for the colorized output images

4.3 Ablation Studies

Extensive ablation studies were performed to validate the effectiveness of each module of the proposed method. The ablation study to demonstrate the effectiveness of ColorFormer Block, residual connections, and adversarial learning is set up. The setup of ablation studies is shown in Table 5-1 and results are in Fig. 4-2, Table 4-3, Table 4-4, and Table 4-5.

4.3.1 Convolution:

To test how the transformer block contributes to the overall performance, the ColorFormer Block is replaced with convolution layers. The transformer plays an important role in capturing long-range dependencies. Experimental results show that the convolution results in inconsistent and unnatural color to the image, compared to the proposed ColorFormer Block. Finally, ColorFormer leads to a gain of approximately 1.6 dB in PSNR.

4.3.2 Residual Connection:

The residual connections are added to every ColorFormer Block to compensate for missing information in the proposed network. Colors are dull and unsaturated if the residual connections are disabled, as shown in Fig. 4-3-(a) and –(b). The results indicate that residual connections alleviate the problem of vanishing gradient and help to improve the flow of information. Residual connections show promising results in both PSNR and SSIM values, as shown in Table 4-3.

Table 4-2. Setup for ablation studies

Test items	**Transformer block**	**GAN architecture**	**CFFN**	**Residual connection**
Convolution without residual	✗	✓	✗	✗
Convolution with residual	✗	✓	✗	✓
Convolution w/o discriminator	✗	✗	✗	✓
ColorFormer without residual	✓	✓	✓	✗
ColorFormer w/o discriminator	✓	✗	✓	✓
Feedforward (MLP)	✓	✓	✗	✓
ColorFormer	✓	✓	✓	✓

Table 4-3. Ablation study on the effect of convolution and residual connections

Variations	**PSNR (dB)**	**SSIM**	**Colorfulness**	**△Colorfulness**
Convolution without residual	22.806	0.936	26.51	6.47
Convolution with residual	22.965	0.937	25.32	7.66
ColorFormer without residual	24.414	0.930	29.45	3.53
ColorFormer	**24.517**	**0.943**	**31.38**	**1.60**

Table 4-4. Ablation study on the effect of adversarial learning

Variations	**PSNR (dB)**	**SSIM**	**Colorfulness**	**△Colorfulness**
Convolution without discriminator	22.7928	0.935	26.73	6.26
ColorFormer without discriminator	24.3615	0.952	17.48	15.50
ColorFormer	**24.517**	**0.943**	**31.38**	**1.60**

4.3.3 Adversarial Learning:

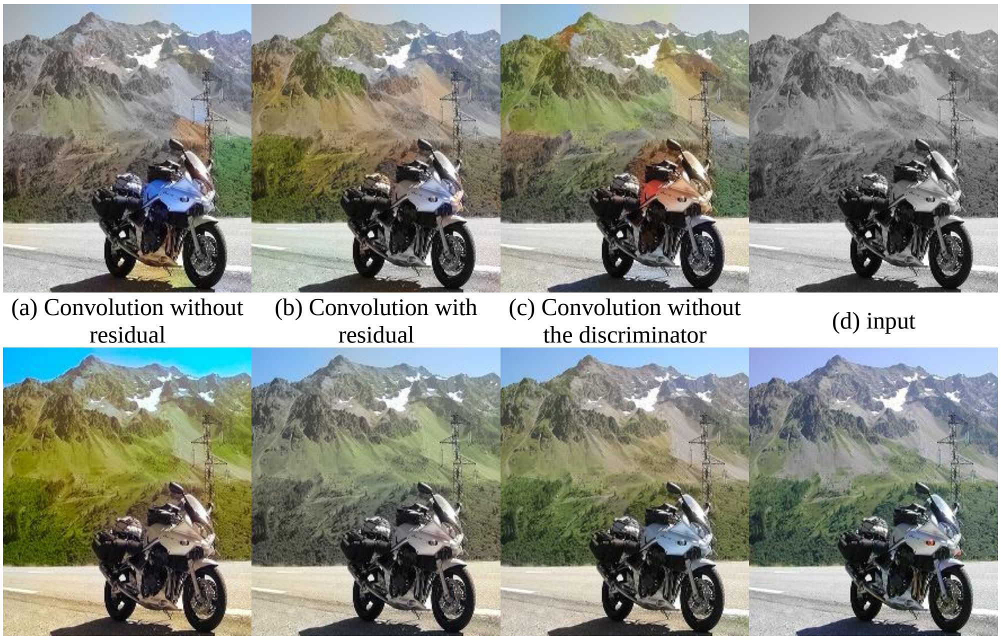
The effect of the discriminator and adversarial loss for colorization is investigated. The discriminator is removed, and the generator is trained with perceptual loss. In particular, adversarial learning can have a substantial effect on colorization by enabling the generative model to generate colorized images that are more visually realistic and consistent with real color images. Adversarial learning helps overcome the limitations of conventional colorization algorithms that rely on heuristics and color distribution assumptions. Table 4-4. shows how adversarial learning contributes to the overall performance of the network. Non-adversarial learning approaches tend to generate less realistic colorization results, as demonstrated in Fig. 4-3-(c) and -(f), where the absence of adversarial learning leads to suboptimal colorization quality.

Table 4-5. Ablation study on the effect of parameters

Variations	PSNR (dB)	SSIM	Colorfulness	△Colorfulness
Feed-forward (MLP)	24.291	**0.944**	19.44	13.54
ColorFormer	**24.517**	0.943	**31.38**	**1.60**

4.3.4 Feed-forward Network:

The feed-forward network (FFN) within the transformer layer plays a crucial role in introducing non-linear transformations, enabling the model to learn more complex features and relationships, thereby enhancing its expressive power and representation capability. Incorporating convolution within the FFN offers additional benefits, such as efficient local feature extraction and capturing spatial relationships in images, which helps the model better understand the structure and context of the input data. To test the effect of convolution in the feed-forward network, a conventional feed-forward network is used in this ablation study. As a result, PSNR values get higher, and output images look more natural when the convolution layer is used in the feed-forward network, as shown in Fig. 4-3-(g). Colorfulness and △ Colorfulness values in Table 4-5. show that convolution in the feed-forward network significantly enhances the colorization results.



(a) Convolution without residual

(b) Convolution with residual

(c) Convolution without the discriminator

(d) input

(e) ColorFormer without residual

(f) ColorFormer without the discriminator

(g) ColorFormer

(h) GroundTruth

Figure 4-2. Visual comparisons of the ablation study for the proposed network

5. Conclusion

In this thesis, a novel and robust image colorization technique is proposed that leverages the power of a conditional Wasserstein Generative Adversarial Network (CWGAN) and the ColorFormer Block, which employs a window-based multi-head self-attention mechanism for lightweight and efficient processing. By integrating convolution layers into the ColorFormer Block, the proposed approach can effectively capture both local and global features, leading to superior colorization results compared to existing state-of-the-art methods. Through extensive ablation studies, the effectiveness of the proposed method is demonstrated, showcasing its ability to generate visually appealing and realistic colorized images. The proposed adversarial training methodology ensures that the generated colorizations are plausible and visually compelling. The proposed method holds significant potential for various applications, including art restoration, video colorization, and enhancing low-quality or historical images. In conclusion, the proposed ColorFormer architecture demonstrates the effectiveness of leveraging both GANs and transformer blocks in the image colorization domain, paving the way for further advancements and research in this area.

Acknowledgment

I express my deepest gratitude to Allah Almighty, the Most Merciful and Compassionate, for granting me the strength, guidance, and blessings throughout the journey of completing this thesis.

I extend my sincere and deepest appreciation to Prof. Bumshik Lee for his invaluable mentorship, unwavering support, and scholarly insights that have significantly enriched the quality of this research. I am truly grateful for the knowledge and skills I have gained under his supervision.

My heartfelt thanks go to my family for their enduring love, encouragement, and understanding. Their constant support has been my pillar of strength.

Finally, I am also grateful to my friends and fellow lab mates for their support and encouragement during challenging times. Their friendship has been a source of motivation and joy.

May Allah's grace continue to guide and bless us all.

References

[1] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, “Two-stage sketch colorization,” ACM Trans. on Graphics, vol. 37, no. 6, pp. 1–14, 2018.

[2] Y. Qu, T.-T. Wong, and P.-A. Heng, “Manga colorization,” ACM Trans. on Graphics, vol. 25, no. 3, pp. 1214–1220, 2006.

[3] Y. Chen, Y. Luo, Y. Ding, and B. Yu, “Automatic colorization of images from chinese black and white films based on cnn,” in Proc. ICALIP, 2018, pp. 97–102.

[4] Y. Guo, X. Cao, W. Zhang, and R. Wang, “Fake colorized image detection,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 8, pp. 1932–1944, 2018.

[5] G. Larsson, M. Maire, and G. Shakhnarovich, “Colorization as a proxy task for visual understanding,” in Proc. CVPR, 2017, pp. 6874–6883.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., “Imagenet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, 2015.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, 2017.

[8] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proc. ICML, 2017, pp. 214–223.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. CVPR, 2017, pp. 1125–1134.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. NeurIPS, 2014, pp. 2672–2680.

[11] A. Levin, D. Lischinski, and Y. Weiss, “Colorization using optimization,” in ACM Trans. on Graphics, vol. 23, no. 3, 2004, pp. 689–694.

[12] Y.-C. Huang, Y.-S. Tung, J.-C. Chen, S.-W. Wang, and J.-L. Wu, “An adaptive edge detection based colorization algorithm and its applications,” in Proc. ACM MM, 2005, pp. 351–354.

[13] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," IEEE Trans. Image Process., vol. 15, no. 5, pp. 1120–1129, 2006.

[14] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images" In ACM Transactions on Graphics (TOG), volume 21, pp. 277–280. ACM, 2002.

[15] R. Ironi, D. Cohen-Or, and D. Lischinski, "Colorization by example. In Rendering Techniques", pp. 201–210. Citeseer, 2005.

[16] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in Proc. ECCV, 2016, pp. 577–593.

[17] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images", In ACM Transactions on Graphics (TOG), volume 30, page 156. ACM, 2011.

[18] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplarbased colorization," ACM Trans. on Graphics, vol. 37, no. 4, p. 47, 2018.

[19] B. Zhang, M. He, J. Liao, P. V. Sander, L. Yuan, A. Bermak, and D. Chen, "Deep exemplar-based video colorization," in Proc. CVPR, 2019, pp. 8052–8061.

[20] H. Bahng, S. Yoo, W. Cho, D. Keetae Park, Z. Wu, X. Ma, and J. Choo, "Coloring with words: Guiding image colorization through text-based palette generation," in Proc. ECCV, 2018, pp. 431–447.

[21] H. Kim, H. Y. Jhoo, E. Park, and S. Yoo, "Tag2pix: Line art colorization using text tag with secat and changing loss," in Proc. ICCV, 2019, pp. 9056–9065.

[22] C. Zou, H. Mo, C. Gao, R. Du, and H. Fu, "Language-based colorization of scene sketches," ACM Trans. on Graphics, vol. 38, no. 6, pp. 1–16, 2019.

[23] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization", In Proceedings of the IEEE International Conference on Computer Vision, pp. 415–423, 2015.

[24] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization" In European conference on computer vision, pp. 649–666. Springer, 2016.

[25] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization

with simultaneous classification", ACM Transactions on Graphics (TOG), 35(4):110, 2016.

[26] J. Zhao, L. Liu, C. G. Snoek, J. Han, and L. Shao, "Pixel-level semantics guided image colorization," in Proc. BMVC, 2018.

[27] J. Zhao, J. Han, L. Shao, and C. G. Snoek, "Pixelated semantic colorization," Int. J. Comput. Vis., pp. 1–17, 2019.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proc. NeurIPS, 2014, pp. 2672–2680.

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proc. CVPR, 2017, pp. 1125–1134.

[30] K. Nazeri and E. Ng, "Image colorization with generative adversarial networks", arXiv preprint arXiv:1803.05400, 2018.

[31] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks", In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 151–166. Springer, 2017.

[32] A. Deshpande, J. Lu, M.-C. Yeh, M. Jin Chong, and D. Forsyth, "Learning diverse image colorization," in Proc. CVPR, 2017, pp. 6837– 6845.

[33] P. Vitoria, L. Raad, and C. Ballester, "Chromagan: Adversarial picture colorization with semantic class distribution," in Proc. WACV, 2020, pp. 2445–2454.

[34] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization Transformer", in International Conference on Learning Representations, 2021.

[35] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee ,Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, Sunghyun Cho, "BigColor: Colorization using a Generative Color Prior for Natural Images", in European Conference on Computer Vision (ECCV), 2022.

[36] Tobias Plotz and Stefan Roth, "Benchmarking Denoising Algorithms with Real Photographs", In CVPR, 2017.

[37] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee, "Single Image Defocus Deblurring Using KernelSharing Parallel Atrous Convolutions", In ICCV, 2021.

[38] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo, "Learning Texture Transformer Network for Image Super-Resolution", In CVPR, 2020.

[39] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao, "CycleISP: Real image restoration via improved data synthesis", In CVPR, 2020

[40] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "SwinIR: Image Restoration Using Swin Transformer", In ICCV Workshops, 2021.

[41] Jianping Shi, Li Xu, and Jiaya Jia, "Just Noticeable Defocus Blur Detection and Estimation", In CVPR, 2015.

[42] Z. Wang, X. Cun, J. Bao, and J. Liu, "Uformer: A General U-Shaped Transformer for Image Restoration", CoRR, vol. abs/2106.03106, 2021.

[43] Dan Hendrycks and Kevin Gimpel, "Gaussian Error Linear Units (GELUs)", arXiv preprint arXiv:1606.08415, 2016.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", In MICCAI. Springer, 2015.

[45] R. Zhang et al., "Real-Time User-Guided Image Colorization with Learned Deep Priors", ACM Trans. Graph., vol. 36, no. 4, Jul. 2017.

[46] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware Image Colorization", in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[47] "Deoldify," https://github.com/jantic/DeOldify.

[48] Vaswani, Ashish, et al. "Attention is all you need." Advances in Neural Information Processing Systems. 2017.

[49] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." International Conference on Computer Vision. 2021.

[50] S. Weng, J. Sun, Y. Li, S. Li, and B. Shi, "CT2: Colorization Transformer via Color Tokens", in ECCV, 2022.

[51] International Commission on Illumination (CIE), "Colorimetry," 3rd ed. Vienna, Austria: CIE, 2004.

[52] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7-9, 2015, pp. 1-13.

[53] D. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," IEEE Transactions on Information Theory, vol. IT-19, no. 4, pp. 471-480, Jul. 1973.

[54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[55] D. M. Chandler and S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," IEEE Transactions on Image Processing, vol. 16, no. 9, pp. 2284-2298, Sep. 2007.

[56] C. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006, pp. 205-210.

[57] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, June 21-24, 2010, pp. 807-814.

Publications

International Journal Paper:

1. H. Shafiq, T. Nguyen, and B. Lee, “ColorFormer: A Novel Colorization Method Based on a Transformer”, submitted in IEEE Transaction on Multimedia, 2023.

2. H. Shafiq and B. Lee, "Image Colorization Using Color-Features and Adversarial Learning," in IEEE Access, vol. 11, pp. 132811-132821, 2023, doi: 10.1109/ACCESS.2023.3335225

Domestic Conference Paper:

1. Hamza Shafiq and Bumshik Lee, “TransVivid: Reviving Visuals with Transformer-based Colorization”, 인공지능신호처리 학술대회, Sep 2023.

2. Hamza Shafiq and Bumshik Lee, “An image colorization method using a transformer”, 한국통신학회 학술대회논문집, June 2023.

3. Hamza Shafiq and Bumshik Lee, “ColorGAN: Generative Adversarial Network based Image Colorization”, in Proceedings of KIIS Autumn Conference, Oct 2022.

Patent:

1. 트랜스포머 블록을 포함하는 GAN 기반의 모델을 이용하여 흑백 이미지를 컬러링하기위한 전자 장치 및 그 동작.

2. 기반의 모델을 이용하여 흑백 이미지를 컬러링하기 위한 전자 장치 및 그 동