BEYOND BLACK AND WHITE: COMPUTER VISION MODEL FOR AUTOMATIC IMAGE COLORIZATION

Yukai Wang, Chengrui Hu, Guojiao Zhao
Shanghai Jiao Tong University
800 Dongchuan Road, Minhang, Shanghai

Abstract

Automatic image colorization is a significant challenge in computer vision, with existing methods based on CNNs and GANs often suffering from desaturation and instability. To address these limitations, we propose a novel two-stage framework that enhances the pre-trained SIGGRAPH17 model with an advanced refinement module. Our core contribution is an Enhanced U-Net that uniquely integrates a diffusion-inspired single-step residual correction mechanism, a global color prior vector to ensure color coherence, and a self-attention block in the decoder to capture fine-grained details. Quantitative experiments demonstrate that our model achieves a state-of-the-art Mean Squared Error (MSE) of 63.812, Peak Signal-to-Noise Ratio (PSNR) of 30.500, and Structural Similarity Index (SSIM) of 0.983, marking a 19.03% MSE reduction compared to the powerful SIGGRAPH17 baseline. Our approach yields more vivid and semantically consistent colorizations, providing an effective solution for high-quality automatic image colorization.

1. Introduction

Color plays a crucial role in human visual perception, offering intuitive cues about object identity, material properties, and scene context. But many valuable images–such as historical portraits, grayscale medical scans, or black-and-white artistic works–exist without color information, which limits both their interpretability and visual appeal. Automatically restoring color to such grayscale images is a long-standing and inherently ambiguous challenge in computer vision, as multiple plausible colorization can correspond to the same input.

The specific problem we investigate in this project is: how to build effective deep learning models that can automatically colorize black-and-white images, especially grayscale facial portraits, using only the visual information present in the image itself. This task is interesting not only due to its technical difficulty but also because of its wide practical impact in fields such as digital heritage restoration, archival enhancement, and content creation. It requires models to capture subtle textures, shapes, and semantic patterns without relying on external hints or labels.

While traditional colorization methods often depended on hand-crafted features or required manual input, modern learning-based approaches–especially those using convolutional neural networks (CNNs)–have shown great potential in mapping grayscale to color. However, CNN-based models trained with pixel-wise losses tend to produce desaturated or overly smooth outputs, as they prioritize average correctness over perceptual realism. For another thing, generative adversarial networks (GANs) had been introduced to encourage more vivid and photorealistic colorization via adversarial training, but it bring the problem of unstable colorization and high computational costs.

In this paper, in order to implement and evaluate deep learning models that infer plausible color distributions from grayscale inputs, we design a novel and comprehensive framework for image colorization enhancement, which builds upon a pre-trained CNN-based model SIGRAPH17. Our model is trained using a public dataset **Flickr-Faces-**

HQ[2],which was first introduced alongside the StyleGAN architecture and featured with huamn frontal faces. 70000 pictures in total are splitted into 60000 for train set, 5000 for validation set, and 5000 for test set.

Our key contributions are:

- We designed a Enhanced U-Net embedded with a Self-Attention module as a correction network, significantly improving the model's perception of global color associations.
- We introduced a four-dimensional color prior, which serves as an effective global condition to guide color generation.
- Inspired by diffusion models, we developed an efficient single-step perturbation-correction scheme, complemented by a composite loss function designed with multi-dimensional supervision from perceptual, style, and attention metrics.

Leveraging these innovations, our model achieves a substantial performance breakthrough. Experimental results verify a 19.03% reduction in the Mean Squared Error (MSE) against the powerful SIGGRAPH17 baseline, clearly demonstrating the superiority of our approach.

2. Related Works

In recent years, many methods have been proposed to improve the performance of colorization models, and most commonly used architectures contain Convolutional Neural Networks(CNNs), Generative Adversarial Networks (GAN).

2.1. Convolutional Neural Networks (CNNs)

Automatic image colorization methods based on Convolutional Neural Networks (CNNs) have made significant progress. Early models treated colorization as a regression task, employing a Mean Squared Error (MSE) loss, but this often led to results with desaturated and grayish colors.

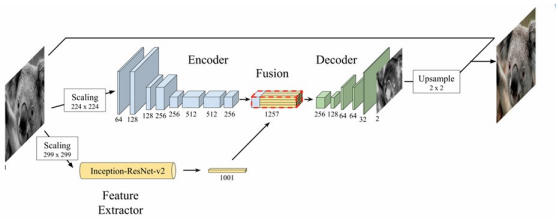

Figure 1. CNN Architecture

To address this issue, Zhang et al. (2016) [3] pioneered the reframing of colorization as a classification task, generating more vivid and vibrant colors by predicting the probability distribution of pixels in a quantized color space and incorporating a class-rebalancing strategy. Building on this foundation, subsequent research has explored various network architectures. For example, Jwa and Kang (2021) adapted and applied the FusionNet encoder-decoder architecture for colorization, also using a classification loss to avoid the problems caused by MSE. Meanwhile, Baldassarre et al. (2017) [1] proposed a hybrid method that fuses a CNN trained from scratch with high-level semantic features extracted from a pre-trained Inception-ResNet-v2 network to enhance the model's understanding of image content.

Furthermore, to merge the advantages of user control with deep learning, Zhang et al. (2017) developed a real-time interactive colorization system. Its CNN framework can directly process a grayscale image along with sparse user-provided hints (such as local color points or global statistics) and intelligently propagate the user's intent throughout the entire image. In summary, CNN-based image colorization technology has evolved from fully automatic classification frameworks to comprehensive systems that integrate multi-model features with real-time user interaction.

2.2. Generative Adversarial Networks (GAN)

Generative Adversarial Networks (GANs) have emerged as a powerful framework for image generation tasks, including automatic colorization. Unlike early colorization methods that optimized pixel-wise reconstruction losses such as L2 or L1, which often led to blurry and desaturated results, GAN-based approaches aim to generate perceptually realistic images by learning a data-driven loss function. A seminal work by Isola et al. (2017) introduced the conditional GAN (cGAN) framework for general-purpose image-to-image translation, including grayscale-to-color tasks. Their method employed a U-Net-based generator with skip connections to preserve spatial structure and a PatchGAN discriminator to focus on local realism. By combining adversarial loss with an L1 reconstruction term, their model produced outputs with sharper edges and more vivid colors than traditional regression-based techniques.

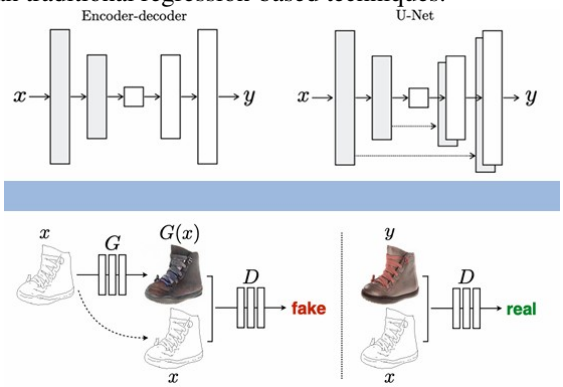

Figure 3. GAN Architecture

Subsequent research has extended this foundation by integrating semantic guidance, perceptual losses, or multi-scale discriminators to improve the quality and consistency of generated colors. Additionally, some models incorporate

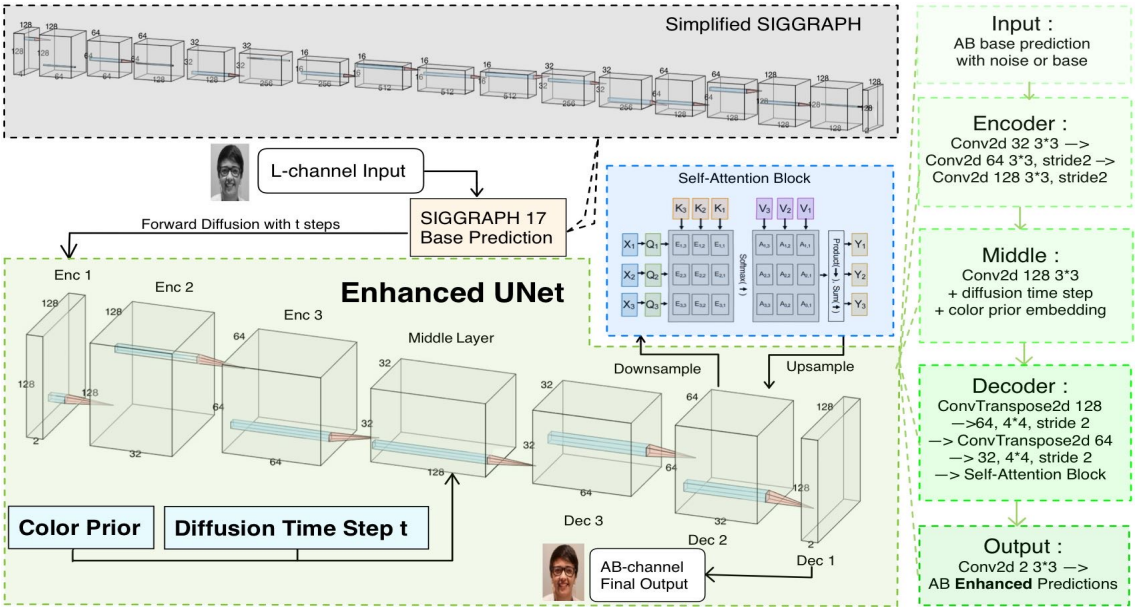

Figure 2. Our Network Architecture. The whole model is composed of a simplified SIGGRAPH architecture, an Enhanced UNet with a Self-Attention Block. The rightmost procedure presents a detailed workflow of Enhanced UNet.

user interaction, enabling sparse color hints to guide the generator while maintaining the learned prior from large-scale data. These GAN-based approaches represent a significant shift in image colorization, focusing on matching the perceptual distribution of real images rather than minimizing simple pixel-wise errors, thus aligning better with human visual preferences.

3. Method

We propose a two-stage colorization framework that builds upon SIGGRAPH17 and introduces a residual refinement mechanism guided by an enhanced UNet module. The overall architecture is shown in Figure 2. In our model, the first stage is SIGGRAPH17 Base Prediction, and the second stage is Enhanced UNet Refinement.

3.1. SIGGRAPH17 Base Prediction

In our model, given a grayscale input image, an initial color prediction will be made by SIGGRAPH17(the base model). This base model captures low- and mid-level features and outputs a rough estimation of the AB channels.

3.2. Enhanced UNet Refinement

We design an enhanced UNet module for residual refinement, whose structure is shown on the right side of Figure 2. This enhanced UNet module is composed of a encoder, a middle bottleneck featuring color prior and denoising, as well as a decoder with self-attention block. This enhanced UNet module takes the base prediction result with diffusion noise added as the input. It outputs the final prediction result, the colorized image. Our enhanced UNet module is similar to traditional UNet module, but we improve it by adding color prior feature vector, diffusion mechanism, and self-attention block, which contributes a lot to the model improvement compared with other colorization methods.

3.2.1. Color Prior Feature Vector

To further enhance color correctness and semantic understanding, a color prior feature vector based on the base prediction result is introduced, which provides contextual color distributions as guidance for subsequent refinement.

Given the base prediction AB color tensor $AB \in \mathbb{R}^{2\times H\times W}$, the color prior vector $P \in \mathbb{R}^4$ is computed as:

$$P = \left[\mu_a, \mu_b, \sigma_a + \sigma_b, \frac{\|AB\|_2}{H \cdot W}\right]$$

where:

- μ_a and μ_b represent mean color values,
- σ_a and σ_b represent standard deviations of A and B channels,
- $\frac{\|AB\|_2}{H\cdot W}$ represents normalized color magnitude.

This vector provides prior knowledge about color distributions to the model. It can help the model infer more realistic and semantically consistent colors by reducing ambiguity and improving global color coherence. The color prior also help encourage color diversity beyond the base prediction.

3.2.2. Diffusion Mechanism

After the base prediction result is outputted by the base model SIGGRAPH17, it is then perturbed by noise with t steps, where t is an argument that can be set freely.

The noisy color prediction is then passed into the enhanced U-Net module. The enhanced U-Net is tasked with predicting the residual or noise required to refine the base prediction.

In the middle bottleneck, color prior vector is embedded through a small MLP and fused with the timestep embedding (Figure 2). The color prior embedding provides semantic color guidance, while the timestep embedding controls the denoising process at each stage. Together, they condition the diffusion model to produce coherent and realistic colorization.

3.2.3. Self-Attention Block

A self-attention block is added to the decoder of the enhanced UNet to improve the model performance on lips and eyes. These regions' color is typically very different from natural skin, and colorizing them with vivid color is challenging in previous models. Therefore, a self-attention block is necessary.

We intended to add this block in the middle bottleneck, but adding it in the decoder performs much better than the initial design. Therefore, the final model structure is settled.

3.3. Loss Function

We design a composite loss function, which provides a significant advantage by holistically training the model to balance multiple, often competing, aspects of image quality. It synergistically combines pixel-level losses (L-residual, L-output) for fundamental accuracy with perceptual and style losses to ensure the results are visually realistic and stylistically consistent. Furthermore, the diversity and attention losses actively encourage the generation of more vibrant colors while guiding the model to focus on the most critical image regions, leading to a perceptually superior final output. The total loss is computed as a weighted sum of multiple components:

$$\begin{aligned}\mathcal{L}_{\text{total}} = 0.35 \cdot \mathcal{L}_{\text{residual}} + 0.25 \cdot \mathcal{L}_{\text{output}} + 0.10 \cdot \mathcal{L}_{\text{perceptual}} \\ + 0.10 \cdot \mathcal{L}_{\text{diversity}} + 0.10 \cdot \mathcal{L}_{\text{style}} + 0.10 \cdot \mathcal{L}_{\text{attention}} \quad (1)\end{aligned}$$

Where each loss component is defined as in Table 1.

4. Experiment

Based on the architecture we designed, we begin our training step by step. Firstly, we train the backbone model (SIGGRAPH17) we mentioned in previous session, to do the basic colorization with 30 epochs and 8 batch size on our dataset. In this stage of output, we could detect the model has the fundamental recognition and colorization in the background and human faces.

In the next step, we freeze the parameter we get from the backbone model training. Depending on base prediction result, we train the diffusion model with our well-designed UNet module. To avoid our model losing the focus on the whole image, we dismiss the attention structure at first. We trained in direct diffusion with UNet and without attention. 5 warm-up epochs and 15 normal epochs are trained. In the warm-up epochs, we temporarily forbid noise addition, to preserve the model and prevent it from groundless guess. Subsequently we add the attention module, and trained for 18 epochs with 5 warm-up epochs in 10,000 images. Finally, we extend the dataset to 24,000 images to train for 9 more epochs to ensure the diversity of our training and further polish our result.

5. Results

5.1. Quantitative Evaluation

5.1.1. Loss

To evaluate the effectiveness of our training procedure, we analyze the loss curves and validation performance across different training stages.

Figure 4 shows the training loss evolution during the diffusion training without attention. After 5 warm-up epochs (where no noise is added), the model transitions into the normal training stage, where residual and output losses are optimized. The total loss shows a clear downward trend, indicating the model is gradually learning to refine the base colorization output.

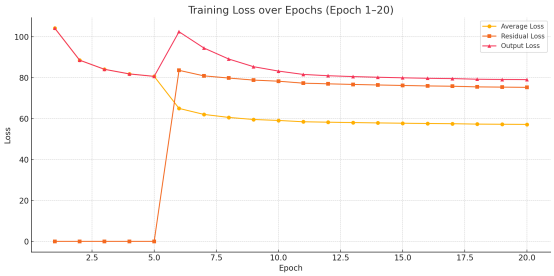

Figure 4. Loss without Attention

Figure 5 presents the training loss after incorporating the attention mechanism. We observe that the model converges to lower losses in both residual and output components, demonstrating that the attention module enhances the model's capacity to capture fine-grained spatial features.

Table 1. Loss components used in 1-step diffusion mode.

Loss Component	Expression	Description
$\mathcal{L}_{\text{residual}}$	$\|\hat{R} - (AB - AB_{\text{base}})\|_2^2$	Residual prediction error between predicted residual and target residual.
$\mathcal{L}_{\text{output}}$	$\|\hat{AB}_{\text{final}} - AB_{\text{target}}\|_2^2$	Final output color error (after adding residual to the base prediction).
$\mathcal{L}_{\text{perceptual}}$	$\|\phi(R\hat{G}B) - \phi(RGB_{\text{target}})\|_2^2$	Perceptual loss using VGG features to ensure perceptual similarity in appearance.
$\mathcal{L}_{\text{diversity}}$	$\|\hat{AB}_{\text{final}} - AB_{\text{base}}\|_2^2$	Color diversity loss to encourage more vibrant outputs than the base prediction.
$\mathcal{L}_{\text{style}}$	$\|G(R\hat{G}B) - G(RGB_{\text{base}})\|_2^2$	Style loss using Gram matrix to preserve stylistic consistency with the base.
$\mathcal{L}_{\text{attention}}$	$\|\text{AttnMap} - \text{ColorMag}\|_2^2$	Attention loss to guide the attention map to focus on regions with high color variation.

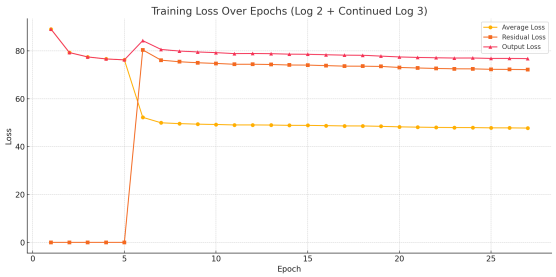

Figure 5. Loss with Attention

5.1.2. Validation

To quantitatively evaluate the performance of our proposed model, we employed three standard and widely-recognized image quality assessment metrics: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and the Structural Similarity Index (SSIM).

- **Mean Squared Error (MSE)** measures the average of the squares of the pixel-wise differences between the generated image and the ground truth image. It provides a direct measure of the reconstruction error, where a **lower value indicates better performance**.
- **Peak Signal-to-Noise Ratio (PSNR)** is the ratio between the maximum possible power of a signal (i.e., the maximum pixel value) and the power of the distorting noise, which is measured by the MSE. PSNR is expressed in decibels (dB), and a **higher value signifies a higher-quality reconstruction**.
- **Structural Similarity Index (SSIM)** is a perceptual metric that assesses image quality based on human visual perception. It compares three key features: luminance, contrast, and structure. The SSIM value ranges from -1 to 1, where a score closer to 1 indicates that the generated image is more similar in structure and perception to the original. A **higher value is better**.

We conducted a comparative analysis of our model against three prominent methods: a standard CNN, a GAN-based model, and the established SIGGRAPH17 model. The results of this evaluation are summarized in Table 2.

Table 2. Comparison Metrics' Results

	MSE ↓	PSNR ↑	SSIM ↑
CNN	337.021	24.673	0.926
GAN	368.143	22.471	0.921
SIGGRAPH17	78.807	29.634	0.982
Ours	**63.812***	**30.500***	**0.983***

As demonstrated by the data, our proposed model achieves state-of-the-art performance, outperforming all other methods across all three evaluation metrics. Our model obtains the lowest MSE score of **63.812***, indicating the smallest pixel-level error. Consequently, it achieves the highest PSNR value of **30.500*** dB, signifying the best signal fidelity. Furthermore, our model scores a superior **0.983*** on the SSIM metric, which confirms that its output is not only mathematically more accurate but also structurally and perceptually closest to the ground truth. Compared with our backbone model, our model reaches a reduction of **19.03%** in MSE, and an improvement of **2.92%** in PSNR, **0.101%** in SSIM. These comprehensive results validate the superior effectiveness and robustness of our proposed approach.

5.2. Visualization Comparison

Visually speaking, we're also easy to detect positive change of our result compared to base model. For example, in Figure 6, we find the boy's eyes and lips significantly enhance.

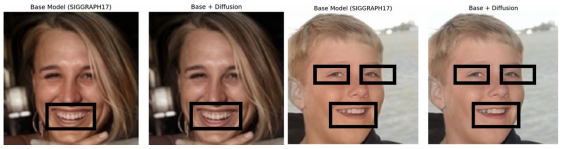

Figure 6. Base Output vs Optimized Output

Figure 7 shows some result of our model's output. We could find that our output has been very close to the input. Some generated images are even more vivid than the original one. The attention map visualization also demonstrate the positive impact that the attention layer brings, where the bright area means more attention.

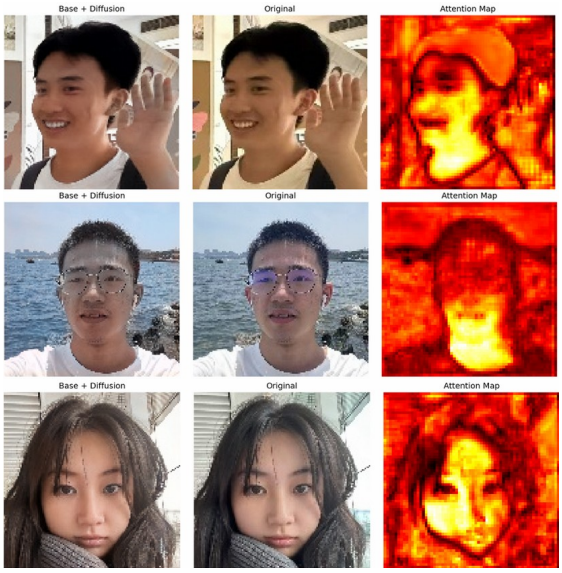

Figure 7. Output vs Origin vs Attention

Figure 8 illustrate the effect of our model on the background and multi-face image. The output prove our model has the basic ability to recognize the common surroundings and also could extend to images with multiple human faces colorization.

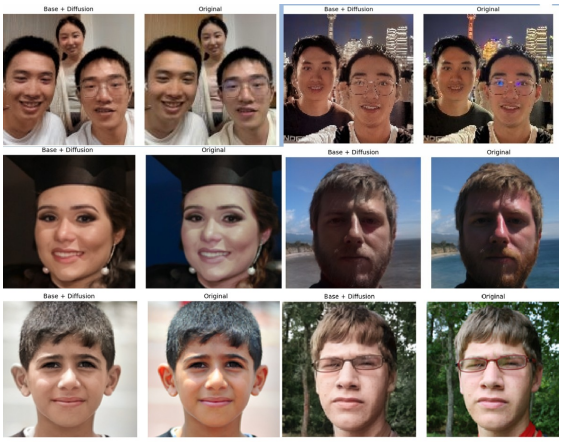

Figure 8. Output vs Origin

6. Limitation and Discussion

Although our model has been shown a good recognition and colorization on the face image. The ability to colorize cool colors is still very limited. And the model could only only learn some common surroundings such as sky, ocean and plants. If it encounter some unknown background without any prompt, such as clothes, the model will give it a color similar to its neighbor or leave it gray.

Figure 9. Ineffective result

Considering these shortcomings, we also propose some directions to improve. For example, we could tag the pictures as warm or cool tones. And we trained two different model targeted at different color style with separate data. It may significantly improve the expressive force of the output image, in response to human demand. And setting the multi-step diffusion is also a potential optimization plan. Because the GPU and other resources limitation, we try to only use one-step diffusion in this project. There is no doubt that the result could be more vivid if we have more resources to perform multi-step diffusion.

In conclusion, our project based on the prior model, develop enhanced UNet with diffusion, color prior, etc., which has been proven to be effective and progressive. Despite drawbacks existing in our project, it still shows a strong improvement and promising for further optimization.

References

[1] Federico Baldassarre, Diego González Morín, and Luis Rodés-Guirao. Deep koalarization: Image colorization using cnns and inception-resnet-v2. arXiv preprint arXiv:1712.03400, 2017. 2

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2

[3] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, pages 649–666. Springer, Cham, 2016. 2