User Guided Image Colorization

Sujeeth Bhavanam
Columbia University
New York, NY
sb4839@columbia.edu

Kushaan Gowda
Columbia University
New York, NY
kg3081@columbia.edu

Abstract

This paper introduces an approach to user-guided image colorization using a UNet architecture combined with a custom class rebalancing classification cross-entropy loss framework. Unlike traditional methods that treat colorization as a regression problem—often resulting in unsatisfactory, brown-dominated images—our approach treats it as a classification challenge. This shift enhances the vibrancy and accuracy of the resultant colorized images. By integrating user input directly into the colorization process, our system provides precise control over the color output, allowing users to selectively apply specified colors to designated pixels in grayscale images. The effectiveness of our approach is validated through extensive experiments on the MS COCO dataset, achieving an accuracy of over 60%*. This highlights how a sophisticated loss function, coupled with effective dataset preprocessing, can enhance simpler architectures to yield superior results in image colorization. Our implementation is available on GitHub* [1]

1. Introduction

Traditionally, image colorization has depended heavily on user input or relied on data-driven automatic methods, each with notable limitations—either requiring substantial user interaction or grappling with the inherent ambiguity of color perception, often yielding unrealistic outcomes.

Our study introduces an innovative user-guided image colorization framework that utilizes a UNet architecture, renowned for its efficacy in various image-to-image translation tasks, combined with a class rebalancing classification cross-entropy loss. This methodology redefines the colorization process as a classification challenge, rather than a regression problem, which has traditionally resulted in desaturated or uniformly brown images. By approaching colorization as a classification task, our framework effectively captures the multimodal nature of color perception, where objects may realistically appear in multiple colors.

The system is specifically designed to integrate sparse user inputs—color hints at particular pixels—to steer the colorization process. This interactive feature allows users to significantly influence the final outcome, ensuring that the colorization adheres to the semantics and textures of the original image while aligning with user expectations and artistic intentions. Our method not only facilitates more precise colorizations but also educates users about effective color choices, enhancing their engagement and satisfaction with the results. This approach not only propels the field of image colorization forward but also broadens the scope for user interaction in image editing tasks. The synergy of a robust neural architecture with a user-centric interaction model marks a substantial advancement in sophisticated image colorization, making it accessible to both novices and experts.

The contributions of this study are outlined as follows:

1. We explore various modifications of the UNet architecture, including the incorporation of general convolution layers, ResNet blocks, and transformer blocks.

2. We introduce a custom class rebalancing cross-entropy loss function that accounts for the significance of non-dominant colors and incorporates user-provided hints.

3. We develop a new dataset derived from the existing MS COCO dataset, tailored for training image colorization models.

4. We conduct comprehensive experiments on this dataset to validate our proposed approach's effectiveness through qualitative and quantitative analyses.

2. Related Work

The task of image colorization has garnered significant attention within the field of computer vision, focusing on converting grayscale images into their colorized counterparts. Early approaches [5, 8] primarily relied on man-

[1] https://github.com/kushaangowda/User_Guided_Image_Colorization

ual colorization techniques or simplistic automated methods based on predefined color palettes. However, recent advances have leveraged deep learning to enhance the automation and accuracy of this process. Zhang et al. [9] transformed the landscape by introducing a deep learning framework that predicts colorization as a classification problem using class-rebalancing to address the issue of color rarity and dataset bias

Building on these foundations, subsequent studies [1, 4] have explored various enhancements, such as integrating attention mechanisms to better capture spatial relationships and utilizing generative adversarial networks (GANs) to refine the aesthetics of the colorized outputs. Moreover, user-guided colorization methods [10] have been developed to allow user inputs to directly influence the colorization results, offering a blend of automation and personalization. These advancements highlight the ongoing evolution of image colorization techniques, moving from basic color prediction to complex models that better understand and recreate the nuances of human color perception.

3. Problem Statement

Given a grayscale image $G \in \mathbb{R}^{\text{HxWx1}}$ and user-provided hints $U \in \mathbb{R}^{\text{HxWx2}}$, our objective is to predict $P \in \mathbb{R}^{\text{HxWx2}}$, representing the A and B channels in the LAB color space.

The LAB color space is advantageous [9] as it simplifies the prediction task to only two channels—A and B—unlike the RGB color space, which involves three components. This reduction is possible because the L channel, representing lightness, corresponds directly to the grayscale image G that serves as the input to our model.

4. Methodology

In this section, we discuss the dataset construction, detail the various model architectures explored, and discuss the loss function utilized.

4.1. Dataset

We utilized images from the MS COCO dataset [6] to generate grayscale and color pairs. The images were initially converted to LAB color space. The original dataset consists of approximately 118,000 images, a significant proportion of which are unsaturated, leading to a predominance of pale-colored outputs in initial tests. To address this, we excluded images where more than 70% of pixels exhibit colors near gray (A and B values close to zero). This filtering process resulted in a refined dataset of 20,000 images for training and testing our model.

Each selected image was resized to 256x256 pixels. In the LAB conversion, the L channel—identical to the grayscale image—was used as input. This was concatenated with user-provided hints, which were integrated by randomly sampling 3x3 patches of the A and B channels, ensuring the total unmasked area was approximately 0.5% of the image.

Furthermore, the image colorization task is framed as a classification problem by discretizing the continuous A and B channel values into categorical bins. After experimenting with various bin sizes, it was determined that using 32 bins for both the A and B channels yielded the most effective results. Opting for classification over regression [9] is strategic, as regression methods are typically characterized by images dominated by brown tones. The classification framework allows for more precise colorization, avoiding the muted outputs commonly associated with regression-based approaches.

4.2. Model Architecture

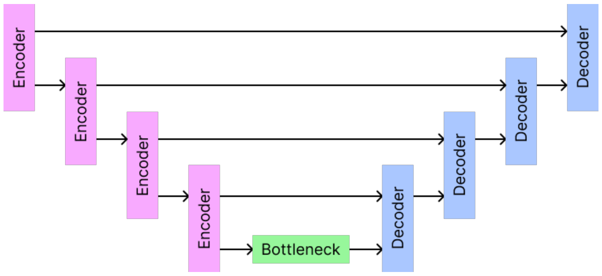


Figure 1. UNet Architecture

U-Net [7] is a convolutional neural network initially developed for biomedical image segmentation. The network features a distinct U-shaped architecture, comprising a contracting path that captures context and a symmetric expanding path that facilitates precise localization. Its architecture, enriched with deep, multi-scale contextual layers, allows for an effective interpretation and reconstruction of colorized images, ensuring the final output maintains the structural integrity and the subtle color nuances of the original scenes. In this study, we developed and evaluated three different variants of U-Net, to predict pixel-level color details from grayscale images. Specifically, we created 4 encoder and decoder layers comprising different components as described in the following sections.

4.2.1 Simple UNet

Each encoder layer consists of a convolutional layer followed by a max pooling operation, which helps in downsampling the image while capturing essential features at various granularities. Conversely, the decoder part of the network utilizes convolutional layers paired with upsampling operations to progressively recover the spatial dimensions. At the bottleneck, two 1x1 convolutional layers act as feature selectors, compressing the feature maps to retain only the most critical information before the decoding process begins. This allows for efficient learning, ensuring that the restored images maintain the integrity and vibrancy of their original color spectrums.

4.2.2 ResUNet

We modified the U-Net architecture by incorporating elements of ResNet [3], specifically through the integration of ResNet blocks within both the encoder and decoder pathways, to enhance feature extraction capabilities and improve gradient flow during training. The encoder consists of a ResNet block followed by a max pooling step, while the decoder includes a ResNet block followed by an upsampling operation. This architecture leverages the strengths of deep residual learning, facilitating the training of deeper networks by alleviating the vanishing gradient problem, making it highly effective for the task of accurate image colorization.

4.2.3 TransUNet

We combined ResNet and Vision Transformer [2] (ViT) blocks to enhance the model's capacity for both local and global feature extraction, crucial for detailed image colorization. The encoder consists of a ResNet block for robust feature extraction followed by a ViT block to capture complex, global dependencies within the image. This sequence is concluded with a max pooling step to reduce spatial dimensions while preserving essential information. The decoder consists of a ResNet block and a ViT block, followed by upsampling. This configuration harnesses the strengths of both convolutional and transformer architectures, ensuring comprehensive feature learning.

4.3. Loss Function

We employed a pixel-wise categorical cross-entropy loss function, incorporating category-based weights to enhance the model's accuracy in predicting less frequent color values. The weights were calculated as follows:

$$\tilde{f}_i = \frac{1}{f_i}, w_i = \frac{\tilde{f}_i}{\sum_j \tilde{f}_j} \tag{1}$$

where f_i is the frequency of the i^{th} color bin in the dataset. This weighting scheme helps to mitigate the dominance of more common colors and promotes a balanced color distribution in the colorized images. The loss function is as follows:

$$\text{loss}_1 = \text{crossentropy}(\text{True}_{\text{A,B}}, \text{Pred}_{\text{A,B}}) \tag{2}$$

$$\text{loss}_2 = \lambda_{\text{patch}} * \text{crossentropy}(\text{True}^{\text{patch}}_{\text{A,B}}, \text{Pred}^{\text{patch}}_{\text{A,B}}) \tag{3}$$

$$\text{loss} = \text{loss}_1 + \text{loss}_2 \tag{4}$$

In this context, $patch$ denotes the unmasked pixels used as user-guided hints. Assigning higher weights to these predictions is essential, as they serve as critical benchmarks that the model must accurately predict. This emphasis ensures that the model prioritizes these areas during training, thereby reducing the potential for misclassification of neighboring areas and enhancing the overall reliability of the colorization results.

5. Experimental Results

Here, we discuss the model setup and provide quantitative and qualitative analysis.

5.1. Setup

For training, our model was configured to optimize performance using a batch size of 32 and was trained over 200 epochs. The learning rate was set at 1×10^{-2} with a weight decay of 1×10^{-3}, utilizing the Adam optimizer. The training was conducted on a distributed setup involving 4 Nvidia T1 GPUs, allowing for efficient parallel processing and significantly reducing training time without compromising the depth of learning.

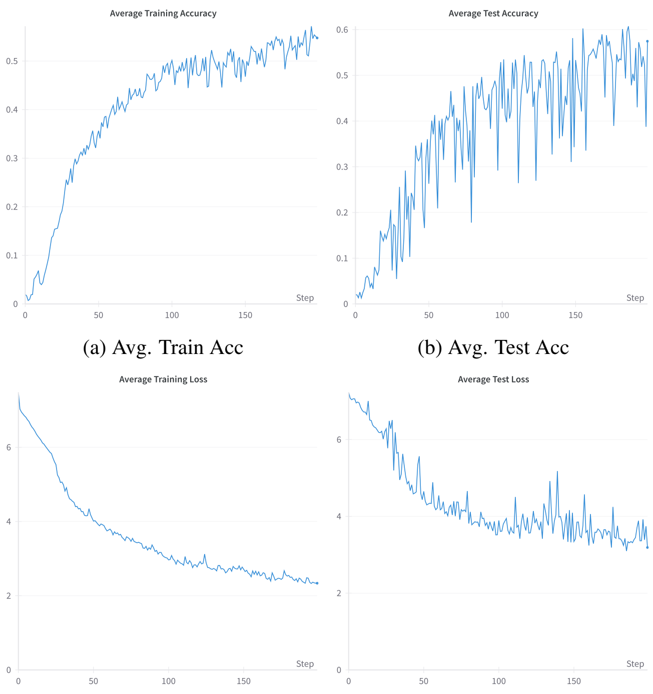
5.2. Quantitative Analysis

Variant	$\text{Acc}_{\text{Train}}$	Acc_{Test}	$\text{Loss}_{\text{Train}}$	$\text{Loss}_{\text{Test}}$
Simple UNet	0.4716	0.4883	3.0272	3.5171
ResUNet	**0.5505**	**0.6074**	**2.3735**	**3.1884**
TransUNet	0.3412	0.3564	4.1615	4.6785

Table 1. Results comparison

From table 1 ResUNet shows the best performance in terms of accuracy, with a training accuracy of 0.5505 and a testing accuracy of 0.6074. This suggests that the residual connections help in training deeper networks effectively by mitigating the vanishing gradient problem, thereby learning more complex features that are beneficial for the colorization task. The Simple U-Net performs moderately well, with a training accuracy of 0.4716 and a testing accuracy of 0.4883. The lack of residual connections may limit the depth to which the network can learn effectively without suffering from training difficulties. TransUNet exhibits the lowest accuracy. Transformers typically excel with larger datasets, as their self-attention mechanisms require substantial data to learn effectively. In this case, it is plausible that the Transformer blocks did not have enough data to leverage their capacity for capturing long-range dependencies within the images.

Fig. 2 showcases the ResUNet model's learning efficacy in the image colorization domain, as evidenced by the upward trends in both Average Training Accuracy (Fig. 2a) and Average Test Accuracy (Fig. 2b). While exhibiting



(a) Avg. Train Acc

(b) Avg. Test Acc

(c) Avg. Train Loss

(d) Avg. Test Loss

Figure 2. Performance Metrics

slight variations, the test accuracy maintains similar levels as train accuracy, underscoring the model's capability to learn and adapt to new data. The loss metrics mirror this positive outcome, with Average Training and Testing Loss (Fig. 2c, 2d) showing a pronounced decrease, affirming the model's optimization prowess.

Variant	**Acc$_{\text{Test}}$**	**Loss$_{\text{Test}}$**
ResUNet	**0.6074**	**3.1884**
ResUNet w/o user hints	0.3006	4.4610

Table 2. Ablation studies for user-guided hints

In Table 2, we present ablation studies to evaluate the impact of user-guided hints on ResUNet's performance. The variant lacking user hints demonstrates a notable decline in performance, achieving a test accuracy of 0.3006. This result underscores the substantial role that user hints play in improving the model's colorization accuracy, thus confirming the effectiveness of user interaction in steering the colorization process in our proposed framework.

5.3. Qualitative Analysis

The ResUNet model closely mirrors the ground truth images with remarkable accuracy (Fig. 3). The nuanced understanding of color spaces and contexts allows for the reproduction of vivid and natural-looking images that are often indistinguishable from their original colored counterparts. However, it is observed that the model occasionally exhibits slight smudging in the colorization of extremely small objects within the scene. This minor artifact likely stems from the model's convolutional nature, which can sometimes merge fine details in areas of minimal spatial extent. Despite this, the predictions indicate a sophisticated level of image comprehension and an impressive ability to replicate the rich, diverse palette of real-world colors.

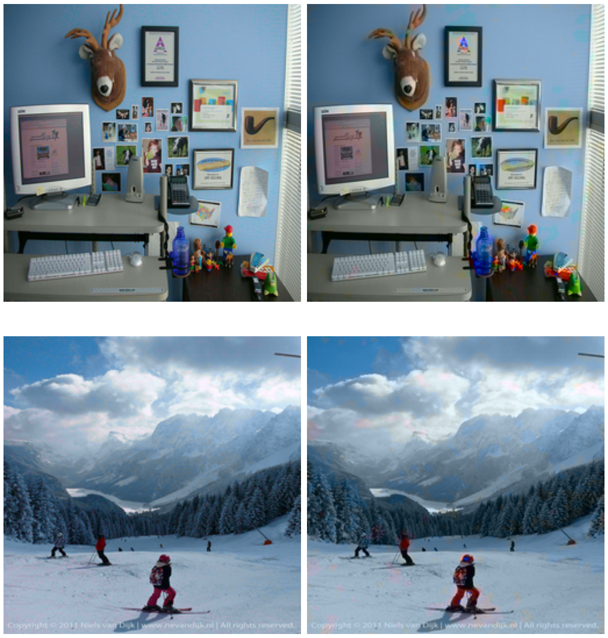


Figure 3. Groundtruth (Left), Predictions (Right)

6. Conclusion

In conclusion, while ResUNet shows promising results in both training and testing, suggesting a good balance between complexity and performance for this task, TransUNet's underperformance might be attributed to the limited size of the dataset which is often crucial for training transformer models. The Simple U-Net serves as a baseline, performing better than TransUNet, but it is outclassed by the enhanced feature learning capability of ResUNet. Future work will prioritize the expansion of training datasets and the exploration of deeper neural architectures to address fine-detail smudging observed in small objects. Leveraging transfer learning, particularly from models proficient in tasks like image segmentation, could further refine the colorization accuracy, providing a robust foundation for enhancing the model's precision and generalization capabilities.

7. Contributions

- Kushaan Gowda: Literature review, dataset creation, user-guided hints implementation, model training, hyperparameter search, qualitative analysis, report writing
- Sujeeth Bhavanam: Literature review, exploration and implementation of several models, model training, hyperparameter search, ablation studies, report writing

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[5] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*, pages 689–694. 2004. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[8] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. Transferring color to greyscale images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 277–280, 2002. 1

[9] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 2

[10] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017. 2