Historical Image Colorization

Patrick Halim, Grace Boettger, Jamie Zhou
University of Michigan
500 S State St, Ann Arbor, MI 48109
pnhalim@umich.edu, graceboe@umich.edu, jamizh@umich.edu

December 12, 2025

1 Introduction

Image colorization of grayscale or black-and-white images can be used to revitalize old photos and films, enhance visualizations, and create realistic and artistic color representations. Converting black-and-white images to colored ones provides a more immersive, unambiguous experience, for educational, historical, or sentimental purposes.

This is a particularly challenging problem because it requires a machine learning model to add information to the existing black-and-white representation via prediction, and it is very easy for humans to see inaccuracies in the model's output. In our project, we aimed to create realistically colored representations of black-and-white images on a less computationally heavy scale. Given the L (lightness) channel of a black-and-white image, our model predicts the corresponding a (red-green component) and b (yellow-blue component) color channels in the CIE Lab color space. In order to test our accuracy, we asked survey participants to differentiate between ground-truth colored images and corresponding computer-generated colorizations.

2 Related Work

L2 loss (MSE loss) is commonly used in traditional image colorization to predict the a and b color channels by minimizing the squared differences between the predicted and ground truth values for each pixel [2]. This approach ensures precise pixel-wise color matching but struggles with ambiguity, as grayscale images often correspond to multiple valid colorizations. L2 loss tends to average out these possibilities, resulting in desaturated outputs. While simple and effective for straightforward tasks, L2 loss is less suitable for creative or diverse colorization challenges.

Zhang et al. describe a different method for colorization in their paper "Colorful Image Colorization" [1]. Unlike traditional approaches that use Euclidean loss and produce desaturated results, this method treats color prediction as a multinomial classification problem, predicting a probability distribution over 313 a-b bins that cover plausible colors in the Lab color space. When comparing an a-b mapping against the ground truth, they used a soft-encoding scheme to categorize and weigh ground-truth values for 5 out of 313 closest valid a-b bins.

Multinomial Cross Entropy Loss provides more flexibility and generalizes better across diverse image content. To address the natural image bias toward less saturated colors, a class rebalancing strategy emphasizes rarer, vibrant colors, and soft-encoding with nearby color bins improves accuracy. This classification-based approach with rebalancing produces more accurate and vibrant results than a regression loss or a classification loss without rebalancing.

For final color prediction, their model uses an "annealed mean" approach, adjusting the softmax temperature to balance vibrancy and spatial consistency. Taking the "annealed mean" allows for them to converge toward an optimum and output realistic, vibrant colors in a single feed-forward pass, achieving lifelike results that outperform previous methods. Additionally, this method can be applied to various grayscale images, regardless of the sub-

ject, making it more versatile than traditional methods.

While the article's method involves a large-scale training process, we adapted it to work on a smaller scale, using fewer images, less training time, and fewer resources. Despite these limitations, the core concepts, such as multinomial classification, soft-encoding, and class rebalancing, are still implemented. This adaptation explores the effectiveness of these techniques within the constraints of limited resources and a slightly modified loss function.

3 Method

We converted RGB images to LAB colorspace to extract L (lightness), a (red-green), and b (yellow-blue) channels. The L channel is the input and a-b channels are supervisory signals. We split the images into training, validation, and testing sets. In response to the constraints of limited GPU, we chose to limit our training dataset to only 10 images. In addition, we constrained images to only be images of garbage trucks.

During training, we tested with both MSE loss and cross-entropy loss with the model architecture shown in Figure 1. Our output is the predicted color distribution for all pixels, based on all valid a-b color bins. Instead of using all B = 313 valid color bins like in [1], we calculated the number of 10-by-10 a-b color bins present in our training dataset and used that as the number of valid color bins. Our reasoning was that due to the small training set, the number of valid bins would likely be less than 313, so we did not want extraneous mappings to color classes that were not present in training.

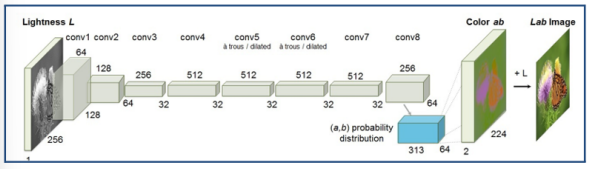

Fig. 1. Each convolutional layer contained 2 to 3 Conv2d and LeakyReLU layers, with no pool layers. Conv5 and Conv6 use dilated convolutions. All layers end with BatchNorm2d besides Conv8. We replaced *B* = 313 with the number of *a-b* bins present in our training dataset.

For MSE loss, our model contained a final convolutional layer mapping B color bins to 2 channels, a or b, using a softmax function followed by a convolution. Whereas for multinomial cross-entropy loss, our model ended with Conv8, and we implemented a custom function converting predictions to images based on the a-b bin with highest confidence.

To calculate the weights for multinomial cross-entropy loss, we calculated the proportion of pixels in each a-b bin from our training set, calculated the unnormalized weights, and then normalized the weights using the expected value equation.

$$\mathbf{w} \propto \left((1-\lambda)\tilde{\mathbf{p}} + \frac{\lambda}{Q}\right)^{-1}, \quad \mathbb{E}[\mathbf{w}] = \sum_q \tilde{p}_q w_q = 1 \tag{1}$$

We defined a custom soft-encoding scheme, where ground truth Z was defined using the 9-nearest neighbors' weights proportional to their distance from the ground-truth a-b pair using a Gaussian kernel with σ = 5. Bins that were not in the array are given the probability of zero. From there, we defined V as the weight of the a-b pair with the highest probability.

$$v(\mathbf{Z}_{h,w}) = w_{q^*}, \quad \text{where } q^* = \arg\max_q Z_{h,w,q} \tag{2}$$

We initially attempted to use this soft-encoding scheme to compare ground-truth weights with our predicted weights using the loss function defined in (3). However, we struggled to define a differentiable custom loss function following this soft-encoding scheme. Thus, we instead defined ground truth as the index of the valid a-b bin.

$$\mathcal{L}_{\text{cl}}(\hat{\mathbf{Z}}, \mathbf{Z}) = -\sum_{h,w} v(\mathbf{Z}_{h,w}) \sum_q Z_{h,w,q} \log(\hat{Z}_{h,w,q}) \tag{3}$$

4 Experiments

We trained two different models on images from ImageNet database. Our MSE loss used a learning rate of 0.01 and 100 epochs, achieving a final loss of 3.400. Our multinomial cross-entropy loss model used a learning rate of 0.01 and 100 epochs, achieving a final loss of 4.275. Figure 2 shows our results. Our MSE loss model often generated fairly realistic results; however, the output colors were often more dull than the ground truth image, even on the train set. Our cross-entropy loss model generated more vibrant colorizations on the train set, matching

Fig. 2. Ground truth images are displayed above. Images colorized by our model are displayed below.

very close to the ground truth image; however, the results on the test set were mostly grayscale images with a few colorized regions. Overall, we found that the MSE loss did a better job generating realistic output images given the limited training set, even though the images were more dull.

We conducted a survey involving 19 participants to evaluate the effectiveness of our image colorization technique. The survey presented six sets of images, each consisting of an original image and its colorized counterpart side by side. Participants were asked to identify the original image in each set. The results showed that the correct original image was selected 65.8% of the time, which is close to the 50% threshold that would indicate perfect colorization, suggesting that our colorization method is quite convincing.

5 Conclusion

Our model demonstrates the potential for recoloring historical photos using two methods: multinomial cross-entropy and regression approach. A significant limitation to our results is the small dataset used for training, validation, and testing. Additionally, our model was run on a simplified soft-encoding scheme and insufficient epochs due to GPU limitations, hindering the model's ability to converge to optimal performance. Future work could address these weaknesses by fine-tuning the loss function and image pool, so that the model will better handle nuances of actual pictures (i.e., challenging cases such as worn, faded, sepia-toned pictures), and significantly enhance the model's accuracy and reliability for broad applications.

References

[1] 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] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In European conference on computer vision, pages 649–666. Springer, 2016.

[3] Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. European Conference on Computer Vision (2016)