IMAGE COLOURIZATION VIA CONVOLUTIONAL NEURAL NETWORKS AND DEEP LEARNING

Youssef Fikry
Student# 1005678901
youssef.fikry@mail.utoronto.ca

Harkirpa Kaur
Student# 1011242479
harkirpa.kaur@mail.utoronto.ca

Peter Leong
Student# 1005678901
peter.leong@mail.utoronto.ca

Thulasi Thavarajah
Student# 10115358424
t.thavarajah@mail.utoronto.ca

ABSTRACT

This project addresses the challenge of automated colourization for 256×256 grayscale images using a dataset of 12,600 image pairs, balanced across human subjects, animals, and natural scenery. We frame colourization as a supervised learning problem in the CIELAB colour space, where a model predicts chrominance channels (a^*, b^*) from the luminance channel (L^*). A shallow convolutional neural network (CNN) provides the baseline performance, while our primary solution employs a deeper convolutional encoder-decoder architecture. This design captures high-level semantic features and spatial context, addressing limitations of shallow networks in perceptual realism. All source code, datasets, and results are publicly available here. —-Total Pages: 6

1 BRIEF PROJECT DESCRIPTION

The invention of photography provided the technology to capture a moment in time. However, for most of photographic history, the process of obtaining coloured images eluded photographers (Science & Museum, 2020). As a result, much of historic photography is grayscale; lacking the visual richness found in modern photography and is inaccessible to individuals with vision impairments (Vogelsang et al., 2024). This project aims to leverage deep learning to automate the colorization of grayscale images, thereby aiding in the revitalization of grayscale photographs. Furthermore, image colorization technology assists archivists and museums in restoring lost visual information and has applications in the media, medical and geospatial industries.

Traditional image colorization methods involve manual labour that is costly and time-consuming (Farella et al., 2022). On the other hand, deep learning approaches automate the colorization process and supplant traditional techniques (Farella et al., 2022). In machine learning, colorization is defined as a model taking a black-and-white image as an input, and outputting its coloured counterpart (Olah & Yang, 2022). From a machine learning perspective, deep convolutional neural networks are best suited for this task as they can extract and learn features such as colors, patterns and shapes in images and affiliate them with object classes (Hwang & Zhou, 2016). These characteristics aid CNNs in excelling at object classificition tasks, as well as image colorization (Hwang & Zhou, 2016).

//insert image here.

2 Individual Contributions & Responsibilities

3 Notable Contribution

3.1 Data Processing

The project's image dataset was compiled using publicly available datasets on Kaggle.com, an online platform that provides access to real-world datasets and a community for data scientists (Kaggle). To train a model that can generalize to a broad range of images, the final dataset for this project includes three categories: human, animal, and scenic. All selected datasets are licensed for public domain use.

3.1.1 Repurposing Online Datasets

The human image dataset, originally intended for human detection, contains a diverse range of 17,300 images of people in different environments (Elmenshawii, 2023). Furthermore, the animal image dataset, initially developed for image classification contains 5,400 images of 90 different animals (Banerjee, 2022). For this project's purpose, this dataset is ideal as it encompasses a diverse set of images with an equal distribution of each animal. Additionally, the scenic image dataset from Rougetet (2020) contains 4,319 images of a variety of landscapes spanning a large breadth of colour palettes, potentially influencing the robustness of the final model.

3.1.2 Cleaning Up The Datasets

The team extracted all the images from each dataset and relocated them into folders corresponding to their category (human/animal/scenic). Due to the disparity in the size of the three datasets, each dataset was reduced to exactly 4200 images using Python's `random.sample` function with the seed set to 42. The images were then renamed in accordance to their respective categories (ex. human_0001.jpg).

3.1.3 Formatting the Data

This project requires a unique dataset with black-and-white images paired with their coloured counterparts as the ground truths. To format the cleaned up data and create this dataset, PyTorch's `torchvision.transforms` library was utiltized. The team first converted the original images into 256 x 256 pixel ground truth images using the `Resize` transform and sorted them according to their category. Following this, `Grayscale` transform with a output channel of 1 was used to convert the 256 x 256 pixel colour images to 256 x 256 pixel grayscale images to feed into the model.

3.1.4 Splitting Dataset Into Training, Validation and Test Sets

To create the training, validation and test sets, a ratio of 70:15:15 was chosen. The first 2940 images of each of dataset categories (human/animal/scene) were selected for the training set. The remaining images were split evenly to create the validation and testing datasets.

3.1.5 The Final Dataset

The final dataset is composed of 12,600 pairs of colour and their corresponding grayscale images. Each category (human, animal, and scenic) contains 4,200 image pairs and an even distribution of each category was maintained in the training, validation and test sets.

b

3.2 Baseline Model

To establish a performance benchmark for our colorization task, we implemented a shallow convolutional neural network called `ShallowColorizationNet`. This model uses a simple encoder-bottleneck-decoder architecture and does not incorporate adversarial training. Its purpose is to test

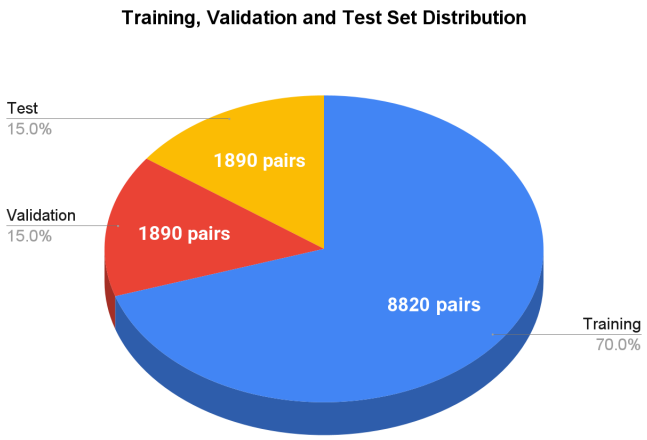

Figure 1: Training, Validation and Test Set Distribution of 12600 Image Pairs

human_0004.jpg bw_human_0004.jpg

Figure 2: Sample Data Pair from Training Set

the feasibility of learning a direct mapping from grayscale luminance (L) input to the chrominance (ab) channels in the CIELAB color space.

3.2.1 Model Architecture

The architecture of the baseline model is as follows:

- **Encoder:** Two convolutional layers with ReLU activation followed by max pooling, reducing spatial dimensions while increasing channel depth.
- **Bottleneck:** A single convolutional layer that expands the feature representation.
- **Decoder:** A transposed convolution to restore spatial resolution and a final convolutional layer outputting two channels (for a and b), with a $\tanh$ activation to constrain values between $[-1, 1]$.

3.2.2 Training and Comparison Approach

The baseline model was trained using the L1 loss between predicted and ground truth ab channels. This model was compared to our primary neural network—a conditional GAN model that includes a discriminator to encourage more realistic colorizations—based on both quantitative and qualitative results.

3.2.3 Quantitative and Qualitative Results

Quantitative: The baseline model was trained for 5 epochs using L1 loss. The generator's loss remained stable around **1.5647**, while the discriminator's loss converged to approximately **0.7074**. These consistent values suggest that the baseline network was able to learn a basic mapping from grayscale to chrominance, but likely lacked the capacity or incentive to produce high-fidelity or diverse outputs. These values serve as reference points for evaluating improvements made by our primary GAN-based model.

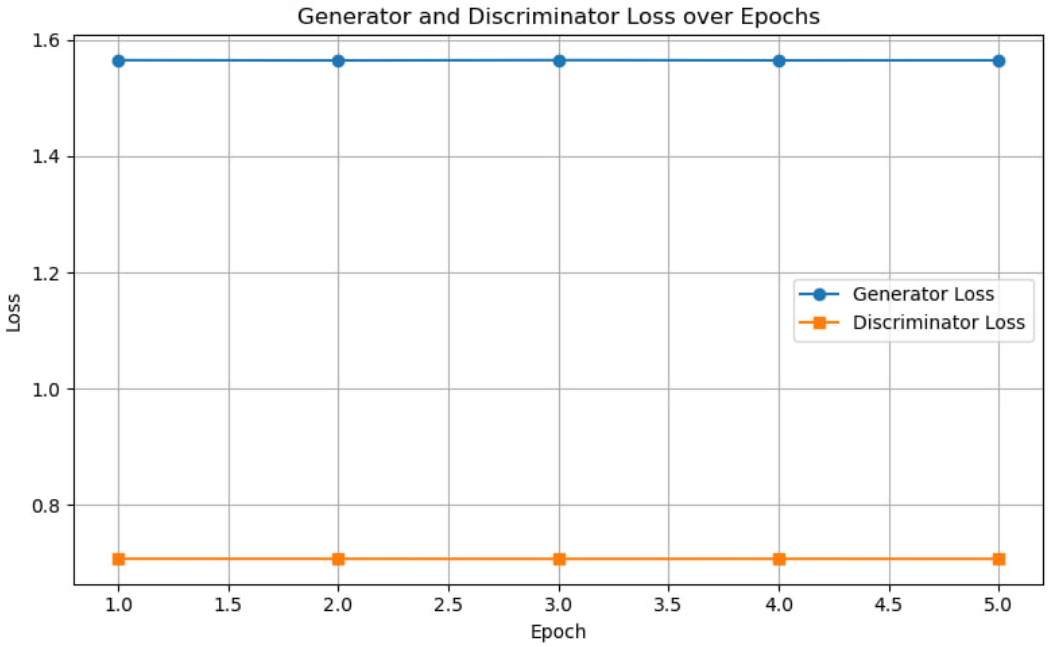

Figure 3: Generator and discriminator loss values over 5 epochs of training. The architecture will be investigated further during the final phase of the project to determine why the loss values fluctuate by such small amounts.

Qualitative: The baseline model produced smooth but somewhat desaturated colorizations. It tended to predict average colors in ambiguous regions, resulting in low color diversity. See Figure 4 for representative examples.

Figure 4: Sample output from the baseline model on grayscale input images.

3.2.4 Challenges

The primary challenge encountered during development was the tendency of the baseline model to learn overly conservative color predictions. Without a discriminator or explicit diversity loss, the network favored colorizations close to the dataset mean. Additionally, selecting the appropriate normalization strategy for LAB space and constraining output to valid chrominance ranges required careful tuning.

3.2.5 Feasibility Assessment

Despite its simplicity, the baseline model demonstrated that the colorization task is learnable with a low-capacity network, although the outputs lacked the vividness and fidelity achieved by our primary GAN-based model. The baseline thus served its role in confirming the viability of end-to-end colorization from grayscale input and establishing a benchmark for model improvement.

3.3 Primary Model

Our colouriser combines a lightweight ResNet-18 encoder with a U-Net decoder. This design keeps global scene context (from the pretrained encoder) while preserving fine edges through skip connections, all within a 15 M-parameter budget suitable for Google Colab.

3.3.1 Architecture

The encoder copies the first four blocks of ResNet-18, pretrained on ImageNet, but the opening convolution is modified to ingest a single-channel L image. Feature maps are therefore produced at 256×256, 128×128, 64×64 and 32×32. A symmetric decoder upsamples in three steps (transpose-conv $\rightarrow$ 3 × 3 conv × 2, BatchNorm, ReLU). Skip links inject the corresponding encoder features to keep edges crisp. A final 3 × 3 layer with tanh activation outputs the a and b chroma channels. Total capacity is 15 M parameters, comfortably below Colab's memory ceiling. The whole network sits under 25 MB on disk, so checkpoints save quickly.

3.3.2 Training Progress

We train with AdamW (lr = 3×10^{-4}, batch = 32) on the cleaned 12.6 k-image set, and the learning rate decays each epoch via cosine annealing. Five epochs on a pro A100 GPU took just under 20 minutes. Validation L1 dropped from 0.07 to 0.0064, and SSIM climbed from 0.58 to 0.74. Qualitatively the model recovers plausible sky blues and fur textures that the two-layer baseline could not capture. The model also learns to colorize human skin tones, but it struggles with complex scenes like forests and crowds, where it tends to produce muddy colors.

3.3.3 CHALLENGES

On some photos the network over-uses warm hues, giving a slight brown cast. Early inspection shows that brown-dominant scenes are over-represented in the training set and the loss does not penalise colour bias. We plan to rebalance LAB histograms per batch and add a small perceptual term to discourage monotone outputs.

3.3.4 FEASIBILITY ASSESSMENT

After five epochs the model already lifts SSIM from 0.58 (baseline) to 0.74, showing it can learn credible colour cues. Remaining errors are mostly a warm "brown" bias. Because these mistakes are consistent, modest adjustments should yield clear gains. We plan to (i) rebalance training batches so cool and warm tones appear equally, and (ii) add a light perceptual loss to discourage over-use of one hue. These changes require no extra parameters or memory. With a 20-epoch run and a cosine learning-rate schedule, we expect further improvement toward an SSIM of 0.80 while staying within course limits.

REFERENCES

Sourav Banerjee. Animal image dataset (90 different animals), 2022. URL https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals.

Fares Elmenshawii. Human dataset, 2023. URL https://www.kaggle.com/datasets/fareselmenshawii/human-dataset/data.

Elisa Mariarosaria Farella, Salim Malek, and Fabio Remondino. Colorizing the past: Deep learning for the automatic colorization of historical aerial images. *Journal of Imaging*, 8(10):269, Oct 2022. doi: 10.3390/jimaging8100269.

Jeff Hwang and You Zhou. Image colorization with deep convolutional neural networks, 2016. URL https://cs231n.stanford.edu/reports/2016/pdfs/219_Report.pdf.

Inc. Kaggle. Level up with the largest ai and ml community. URL https://www.kaggle.com/.

Justin Olah and Jenny Yang. Let there be color: Deep learning image colorization, 2022. URL https://cs231n.stanford.edu/reports/2022/pdfs/109.pdf.

Arnaud Rougetet. Landscape pictures, 2020. URL https://www.kaggle.com/datasets/arnaud58/landscape-pictures.

National Science and Media Museum. A short history of colour photography, Jul 2020. URL https://www.scienceandmediamuseum.org.uk/objects-and-stories/history-colour-photography.

Marin Vogelsang, Lukas Vogelsang, Priti Gupta, Tapan K. Gandhi, Pragya Shah, Piyush Swami, Sharon Gilad-Gutnick, Shlomit Ben-Ami, Sidney Diamond, Suma Ganesh, and Pawan Sinha. Impact of early visual experience on later usage of color cues. *Science*, 384(6698):907–912, 2024. doi: 10.1126/science.adk9587. URL https://www.science.org/doi/abs/10.1126/science.adk9587.