IMAGE COLOURIZATION VIA CONVOLUTIONAL NEURAL NETWORKS AND DEEP LEARNING

Youssef Fikry
Student# 1006682626
youssef.fikry@mail.utoronto.ca

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

Peter Leong
Student# 1010892955
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: 7

1 INTRODUCTION

While colour photography processes first emerged in the 1890s, colour photography did not become widely accessible until the 1970s (Science & Museum, 2020). Consequently, most historical photographs remain in black and white, lacking the visual richness that modern viewers are accustomed to. Moreover, individuals who undergo cataract removal as part of vision restoration procedures often struggle to interpret grayscale images, rendering many historical photographs inaccessible to them Vogelsang et al. (2024). This project aims to leverage deep learning to automatically colourize black and white images, with the goal of restoring visual information and improving accessibility for all audiences. Traditional, non-deep learning colourization methods tend to produce desaturated results and require extensive human input, limiting their scalability (Cheng et al., 2016). In contrast, deep neural networks such as convolutional neural networks (CNNs) can effectively learn spatial and semantic features, enabling realistic colourization without user intervention (Zhang et al., 2016). This makes deep learning a promising and scalable solution for image colourization.

2 BACKGROUND & RELATED WORK

The challenge of image colourization has been addressed through a range of methods, particularly within deep learning. Even among deep learning-based solutions, researchers have proposed a variety of architectures, which can be broadly categorized into five groups: simple colourization neural networks, user-guided colourization networks, diverse colourization networks, multi-path networks, and exemplar-based approaches (Zěger et al., 2021).

Simple colourization neural networks use feedforward convolutional neural networks (CNNs) to directly map grayscale inputs to colour outputs. One of the most influential examples is the work by Zhang et al. (2016), who proposed a fully convolutional network that predicts the *a* and *b* channels in the CIELAB colour space. Their architecture comprises several convolutional layers, each followed by ReLU activations and batch normalization, and is trained as a classification task over quantized ab values, producing more vivid outputs than regression-based methods.

User-guided colourization networks incorporate human input to guide the colourization process. Zhang et al. (2017) extended their earlier work by accepting user-provided colour "scribbles" as input alongside the grayscale image. The network learns to propagate these hints across the image while minimizing differences from the target colour, allowing interactive and controllable colourization.

Diverse colourization networks aim to generate multiple plausible colourizations for a single grayscale input. For instance, Vitoria et al. (2020) used a generative adversarial network (GAN) to produce diverse outputs by learning a conditional distribution over colourizations. This approach addresses the inherent ambiguity in mapping grayscale to colour.

Multi-path colourization networks extract features at multiple spatial resolutions to improve accuracy and context-awareness. Iizuka et al. (2016) proposed a model with both global and local feature pathways, enabling the network to learn both scene-level semantics and fine-grained textures. This structure helps ensure coherent colourization across different image regions.

Exemplar-based colourization networks transfer colour information from reference images to the target. In Su et al. (2020), instance segmentation is used to match regions between the target and exemplars, and two separate colourization networks process this information before merging their outputs. This instance-level guidance simplifies the task compared to end-to-end full-image colourization and enhances accuracy in semantically similar scenes.

3 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 Repurposing Online Datasets

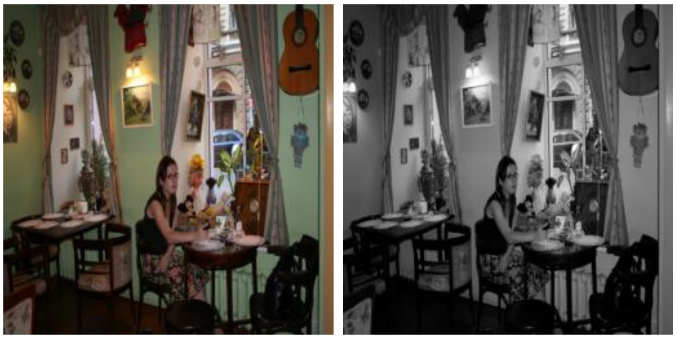
The human image dataset, originally intended for human detection, contains a diverse range of 17,300 images of people in different environments (Elmenshawii, 2023). Furthermore, the animal image dataset, initially developed for image classification contains 5,400 images of 90 different animals (Banerjee, 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.

3.2 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).

3.3 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.



human_0004.jpg

bw_human_0004.jpg

Figure 1: Sample Data Pair from Training Set

3.4 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.

3.5 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.

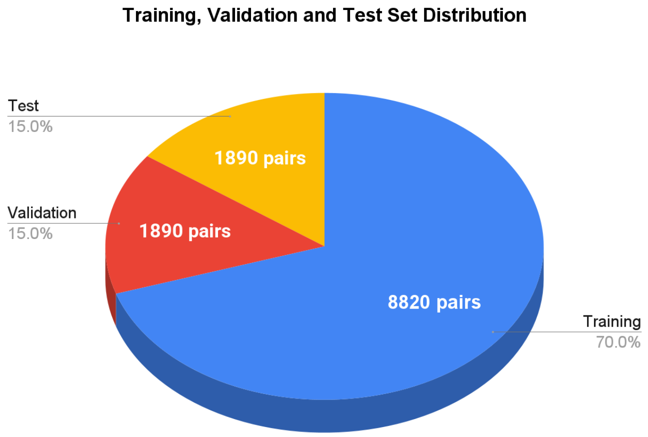


Figure 2: Training, Validation and Test Set Distribution of 12600 Image Pairs

4 Illustration

The top row shows the encoder, composed of blocks with two convolutional layers followed by max pooling, which progressively downsample the grayscale L channel from 256x256 to 32x32 while increasing feature depth from 64 to 512. The bottom row depicts the decoder, which uses upsampling blocks to restore full resolution and predict the two chrominance channels (a and b). Dashed arrows

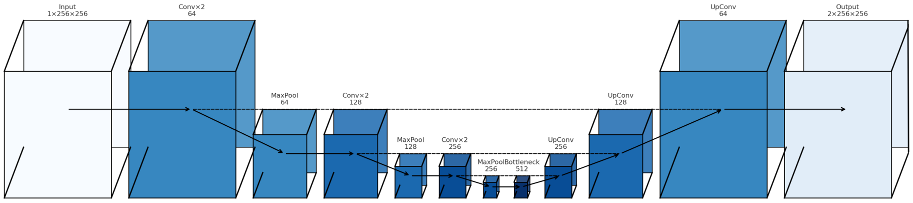


Figure 3: Encoder-decoder colourization network with skip connections.

indicate the three skip connections that transfer fine-grained features from corresponding encoder layers to the decoder stages.

5 Architecture

Our primary model for colourization is a convolutional encoder-decoder network, a common architecture for image-to-image translation tasks (Leatvanich, 2025). Operating in the CIE Lab colour space, the model takes the L channel (lightness) as input and predicts the a and b chrominance channels. To ensure input-target consistency, we derive the L channel directly from the same RGB image used to generate the ground truth a,b channels. The predicted chrominance values are then combined with the L channel and converted back to RGB for visualization. This use of Lab space decouples intensity from colour, aligning with human perception and typically yielding more realistic results than RGB-based models (Leatvanich, 2025).

The architecture resembles a simplified U-Net. The encoder consists of convolutional layers with ReLU activations and batch normalization that progressively downsample the input while extracting high-level features. For example, the model starts with a 3×3 convolution producing 64 feature maps, followed by layers with increasing filter counts (e.g., 128), each halving spatial resolution. A bottleneck with 256 filters captures abstract representations. The decoder upsamples these back to the original resolution using transposed convolutions, predicting the a,b channels. Skip connections between encoder and decoder layers help preserve edge and texture details (Leatvanich, 2025).

We initialized the encoder with ResNet-18 weights pretrained on ImageNet, leveraging semantic features learned from large-scale datasets to better infer plausible colours for common objects (Olah & Yang, 2022). The decoder and output layers were trained from scratch.

Rather than formulating colourization as classification over discrete colour bins (Olah & Yang, 2022), we use a regression-based approach that directly predicts continuous a,b values. The model uses a linear output head without tanh activation to avoid chroma saturation and preserve the full colour range.

To encourage vibrant, diverse colourization and reduce bias toward brownish tones, we apply a weighted colour loss informed by a dataset-wide chrominance prior. We also use a robust (Huber) pixel loss, which better handles uncertainty and avoids bias toward low-chroma predictions. All augmentations were applied consistently to both input and target images, with vertical flips excluded to preserve semantic orientation. Finally, to reduce training time, prior computations were cached and updated once per epoch, reducing training from six hours to one.

6 Baseline Model

Our main baseline was a shallow encoder-decoder convolutional neural network, designed to be lightweight and fast to train. The model accepted the L channel of a grayscale image as input and predicted the a and b chrominance channels. The encoder consisted of two convolutional layers: the first mapped the input to 64 channels using a 3x3 3x3 kernel with padding 1, followed by a ReLU activation; the second expanded to 128 channels using the same kernel size and activation. A max pooling layer with a 2x22x2 kernel and stride 2 then reduced the spatial resolution by half. This was followed by a bottleneck layer with 256 filters and a ReLU activation, which further processed the compressed feature representation (Leatvanich, 2025).

The decoder upsampled the feature map using a transposed convolution that reduced the number of channels from 256 back to 128 while restoring the original resolution, followed by ReLU. A final convolutional layer projected to 2 output channels corresponding to the a and b values, and a Tanh activation was applied to constrain the outputs to the normalized Lab colour range of [-1,1][-1,1] (Rosebrock, 2019).

We made no architectural or hyperparameter changes to this baseline throughout the project. It lacked skip connections, pretrained components, or deep feature extraction, and as expected, often produced desaturated or muted results. Nonetheless, it served as a clear and reproducible benchmark against which the improvements from our full model could be evaluated.

7 Quantitative Results

We used a combination of smooth loss and colourfullness metric to calculate the loss values for our model. The colourfullness metric encouraged the model to use more saturated colours instead of resorting to brown hues, which is an issue we previously encountered.

Validation and training loss was documented after every epoch and the results were graphed, as shown in Figure 4. We observed some overfitting, as the final training loss was 0.004423, and the final validation loss was 0.007595.

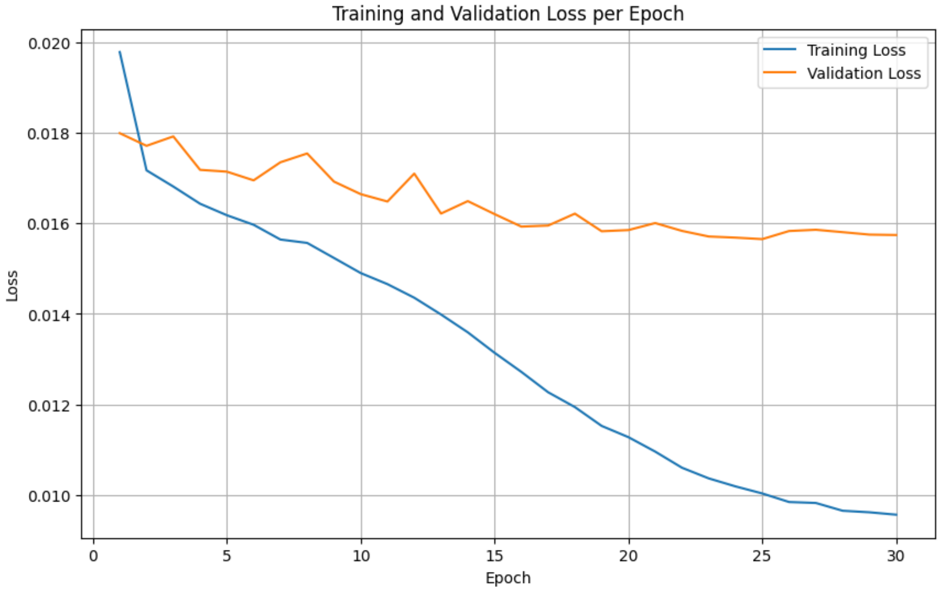


Figure 4: Learning curve showing training and validation loss of model over 30 epochs.

8 Qualitative Results

Our primary model consistently outperformed the baseline, particularly on landscape scenes where the network was able to recover vibrant skies, foliage, and water with relatively realistic hues. In contrast, the baseline produced dull or grayish outputs that lacked semantic colour cues. Figure **??** shows this clear contrast: while the baseline yields low-saturation results, the primary model introduces plausible blue and green tones, suggesting the network successfully learned common associations between luminance and scene content.

Performance was less reliable on images of animals and humans, especially when the ground truth involved bright or saturated colours. In these cases, both the baseline and primary models tended to favour brownish or muted tones, likely due to the loss function penalizing deviations from average colours. This averaging effect is illustrated in Figure 5, where facial features and fur appear washed out or overly neutral.

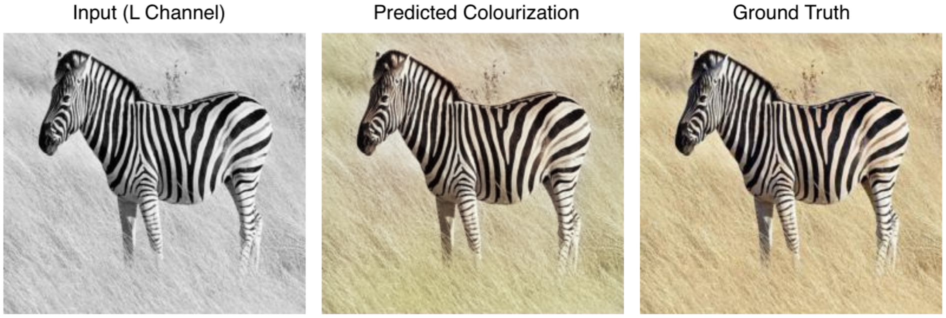


Figure 5: Example output of a zebra image from the primary model. The model successfully captures the zebra's stripes and fur texture, producing a realistic colourization.

We also experimented with several alternate architectures: diverse colourization, quantized outputs using colour bins, multi-path networks, and exemplar-based models. Despite their conceptual appeal, these approaches often underperformed our original model. Figure 6 displays the results from the colour bin model. The exemplar and multi-path models frequently introduced a purplish tint across diverse inputs, while the quantized and diverse variants still exhibited the same dulling behaviour due to similar averaging tendencies. These results highlight that added architectural complexity does not guarantee improved performance, and in some cases, may exacerbate artefacts or introduce new failure modes.

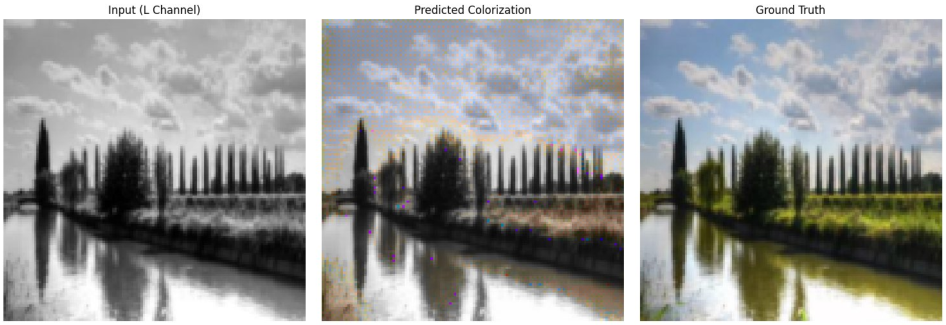


Figure 6: Example output of the colour bin model on a landscape image. The model attempts to predict colours using quantized bins, but the results are often muted and lack vibrancy.

9 Evaluation on New Data

10 Discussion

Our model generally performed well, especially on landscape images, where predicted colours often aligned well with the ground truth. These results suggest the model successfully learned broad structural features such as skies, foliage, and terrain. However, performance declined on images featuring humans and animals, particularly when the ground truth contained highly saturated or vibrant colours. In such cases, both the baseline and our primary model tended toward muted, brownish hues. This behaviour likely arises from the pixel-wise loss, which encourages the model to average colours to minimize error—resulting in desaturated outputs that reduce loss but compromise visual realism.

To explore alternative strategies, we experimented with several advanced colourization approaches, including diverse colourization (generating multiple plausible outputs), quantized prediction using discrete colour bins, multi-path architectures, and exemplar-based methods. Despite their conceptual promise, these alternatives failed to outperform our initial primary model, both quantitatively and qualitatively. In particular, the multi-path and exemplar-based models often produced outputs with a noticeable purple hue, suggesting instability or poor alignment between reference features and target content. Other approaches encountered similar issues of colour averaging, with results again skewing brown or gray to minimize loss.

These findings were unexpected, although the alternative architectures seemed promising in theory, our initial model consistently performed better in both accuracy and visual quality. This shows that adding architectural complexity does not necessarily lead to better outcomes. In fact, some modifications introduced new problems—for example, the exemplar-based and multi-path models often produced a noticeable purple hue. Others, such as the quantized and diverse models, still struggled with the same tendency to average colours, leading to dull or brownish results. Overall, this highlighted the importance of a balanced and well-tuned architecture. Through this process, we learned that effective colourization depends not just on network design, but also on careful attention to loss behaviour, data distribution, and how these elements influence model outputs.

11 Ethical Considerations

The dataset used is publicly available, so there are no copyright or consent concerns. However, there is potential for representation bias, as the dataset may contain imbalances in race, skin tone, animal breeds, or fur colours. For instance, if lighter-skinned individuals are overrepresented, the model may produce unrealistic or culturally insensitive colourizations for darker-skinned individuals. A similar issue applies to animals: limited diversity in breeds or fur patterns may lead to misleading results. While such issues have not visibly appeared in our outputs, they remain a risk, especially if colourized images are used for educational purposes or identification tasks.

There is also a risk of evaluation bias. Without knowledge of the dataset's demographic distribution, we cannot ensure the model performs equally well across all categories. This uncertainty could skew evaluation metrics and lead to misinterpretation of the model's effectiveness.

Finally, because the model generates plausible but unverifiable outputs, there is a risk that users may place undue trust in the results. It is therefore critical to communicate the model's limitations clearly and avoid using it in contexts that require factual accuracy.

REFERENCES

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

Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization, 2016. URL https://arxiv.org/abs/1605.00075.

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

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi 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):110:1–110:11, July 2016. ISSN 0730-0301. doi: 10.1145/2897824.2925974. URL https://doi.org/10.1145/2897824.2925974.

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

Crystal Leatvanich. Image colorization using neural networks, May 2025. URL https://medium.com/@crystal.leatvanich/image-colorization-using-neural-networks-aeafd10e6ef5.

Justin Olah and Jenny Yang. Let there be color: Deep learning image colorization, 2022. URL https://cs231n.stanford.edu/reports/2022/pdfs/109.pdf. CS231n Course Project Report.

Adrian Rosebrock. Black and white image colorization with opencv and deep learning, Feb 2019. URL https://pyimagesearch.com/2019/02/25/black-and-white-image-colorization-with-opencv-and-deep-learning/.

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.

Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization, 2020. URL https://arxiv.org/abs/2005.10825.

Patricia Vitoria, Lara Raad, and Coloma Ballester. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In *Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2434–2443, Los Alamitos, CA, USA, March 2020. IEEE Computer Society. doi: 10.1109/WACV45572.2020.9093389. URL https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093389.

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.

Ivana Zĕger, Sonja Grgic, Josip Vuković, and Gordan Šišul. Grayscale image colorization methods: Overview and evaluation. *IEEE Access*, 9:113326–113346, 2021. doi: 10.1109/ACCESS.2021.3104515.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. *European conference on computer vision*, 2016.

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, 2017. URL https://arxiv.org/abs/1705.02999.