Black and White Image Colourization Using Modified UNet and Patch GAN

Vicky Chen, Jack Yang, Ibraheem Hameed, Mikhail Semyanovskiy
University of Toronto
https://github.com/Ibraheem014/colorGANvas

January 10, 2025

1 Abstract

Our project develops a model to colourize black-and-white images using data from online repositories like ImageNet and self-captured grayscale images. We employ a modified Generative Adversarial Network (GAN) architecture, where grayscale images are input into a UNET-based generator that incorporates a Recurrent Neural Network (RNN) at the smallest kernel size to enhance temporal coherence. This RNN bottleneck helps maintain temporal consistency across images, ensuring realistic color transitions by leveraging contextual feature learning. In the training phase, the generator encodes and decodes the data, while the discriminator provides feedback, enhancing the generator's accuracy. During testing, we retain only the generator, omitting the evaluation step. This hybrid approach balances spatial and sequential dependencies, aiming to improve colorization quality by preserving contextual details.

2 Introduction

Our project uses deep learning to convert black-and-white images into vibrant, realistic colour versions. Image colourization is complex requiring a model to understand object context, textures, and environmental cues to predict plausible colours. Traditional approaches, like the scribble-based method [1], demand user input, limiting scalability and efficiency. Therefore, automating this process with deep learning not only saves time but also allows scalability and adaptability to various types of grayscale images.

This project is significant for a few reasons. First, colourization enhances the interpretability and visual appeal of historical black-and-white images. Important for industries like media and film to preserve cultural heritage. For instance, Peter Jackson's documentary *They Shall Not Grow Old* [2] famously used colourization to vividly portray World War I footage, offering a level of immersion that black-and-white imagery could not achieve. Additionally, colourized images can improve machine vision applications, as colour can provide additional data for downstream tasks like object detection and scene segmentation.

Deep learning excels in feature extraction and pattern recognition, essential for accurate colour prediction. Using CNNs and/or GANs, we can train models to learn colourization patterns on extensive data, such as ImageNet. This data-driven approach helps generalized well across different image types–from natual landscapes to human portraits–improving colour accuracy and addressing ambiguities in grayscale inputs. Our approach included training on large datasets, classifying imgaes by classes, fine-tuning models, and experimenting with architectures to achieve high-quality, realistic colourization.

This project aims to harness deep learning's capabilities for a scalable, effective solution to image colorization.

3 Background & Related work

Automated image colourization, converting grayscale images into RGB images, is an actively researched area in computer vision. Traditionally, manual colourization involved expert artist but advances in deep learning have led to scalable, efficient, and realistic colourization with minimal human intervention. We discuss key studies and developments that inform our approach.
One notable study by Zhang et. al [3] utilized a CNN autoencoder for automatic colorization, treating it as a multinomial classification problem to predict chrominance values (AB channels in the LAB colour space). Their model, trained on ImageNet, demonstrated that large-scale training can produce plausible, contextually accurate colourizations, introducing a novel approach to colourization that avoids the averaging effects seen in regression-based methods.
More recent developments involve GANs, as exemplified by Isola et al. [4]. who used GANs for tasks like converting black-and-white images to colour through their work on Image-to-Image Translation with Conditional Adversarial Networks. This method uses a discriminator to improve the realism and variety of generated colours, showing that GANs, with their adversarial training, help produced more photorealistic and contextually accurate results compared to traditional autoencoders.
For our project, enhanced the UNET generator architecture proposed by Isola et al.[4] by incorporating an RNN as the bottleneck and exploring hybpid models that blend classification with adversal training. This adaptation aims to further the practical application of deep learning-based colourization, achieving high-quality results across various image types.

4 Data

For this project, we used the ImageNet dataset [5], one of the most widely used datasets for computer vision tasks and specifically common for image colourization research. We chose not to focus on colourizing any specific category of images, which makes ImageNet ideal, since it contains over a million images across more than 1000 diverse classes. This variety enabled our model to learn colour patterns and associations across a wide range of objects, textures and scenes. ImageNet poses several challenges for colourization due to its diversity in content and colour distribution. Images vary significantly in terms of lighting, context, and composition, which means the model must learn to colourize both simple and complex objects consistently and accurately. For example, the model needs to differentiate between and appropriately colour objects like blue skies, green trees, and human faces, even when grayscale inputs provide limited visual cues. To mitigate potential biases in our dataset arising from variations in volume and diversity, we maintained consistent classes across all datasets and ensured that the dataset splits (train, validation, and test) are proportionally balanced in terms of object types. This approach guarantees that each dataset includes a diverse range of object types, reducing the risk of overfitting and promoting robust generalization.
To prepare the ImageNet data for training we followed two key preprocessing steps to ensure consistency. Firstly the dataset will be resized to 256x256 pixels to strike a balance between detail and computational efficiency. This resolution is commonly used in colourization models and allows the model to capture essential textures and edges without excessive computational cost. Each colour image is converted to grayscale to create paired grayscale/colour training samples. For grayscale conversion we converted the image from RGB colourspace to LAB and extracted the L channel (luminance). This separation is advantageous because the model's task is to predict the missing colour values (the A and B channels) based on the grayscale input (the L channel).

5 Model Architecture

To start off, we defined our model's encoder as the architecture from Figure 1 . The input L channel of a picture (1x256x256) is down-sampled using convolutions with the number of channels growing proportionate to the conv dimensions to achieve higher dimensional space with lesser computational cost. At the final stages of encoder, the number of channels is locked at 512, while number of features in spatial dimensions is scaled down to 2 by 2. This is prepared to unroll feature quadrants of the picture as a sequence, serving as input to the RNN bottleneck.

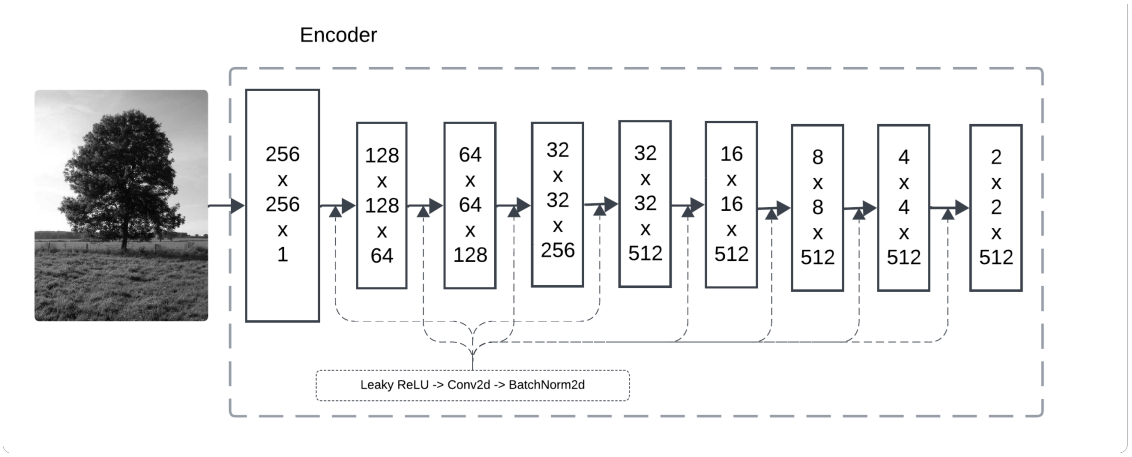

Figure 1: Encoder

The Bottleneck, a bidirectional RNN that processed the quadrant-wise decomposition of the encoder's output as its input sequence, as illustrated in Figure 2. We used a RNN generator pattern similar to a text application in an attempt to approach the picture as a sequence. In our architecture, we passed the feature through the first four GRU nodes to generate (2x512) hidden states, resulting from forwarding the generated feature quadrants from two different directions. The output of the four nodes was concatenated into (4x1024) tensor which was then passed into a sequence of four more GRU nodes alongside hidden states. This time the nodes reused the output and hidden states of each previous GRU node. The rationale for such an architecture was that, without reusing the outputs of the first four GRU nodes, there was significant colour loss. The results from each stage were then concatenated into (16x1024) tensor which was subseqently max pooled, convoluted and activated to reduce it back to (2x2x512).

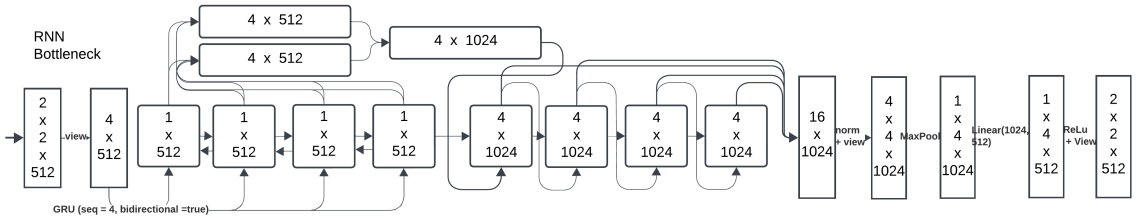

Figure 2: RNN Bottleneck

The decoder is the last component of our generator. It utilizes multiple layers that upscale the RNN output, integrating it with outputs from the encoder layers to maintain continuity of information. As illustrated in Figure 3, In the decoder, we used multiple groups of layers that upsampled the output of the RNN in combination with the output from the encoder layers. This approach allowed us to carry over information from earlier layers. The result of the deconvolutions was then activated by a TanH function and combined with the input L channel along with a 2x256x256 AB model

result to produce the LAB result. We then converted this to RGB for assessment.

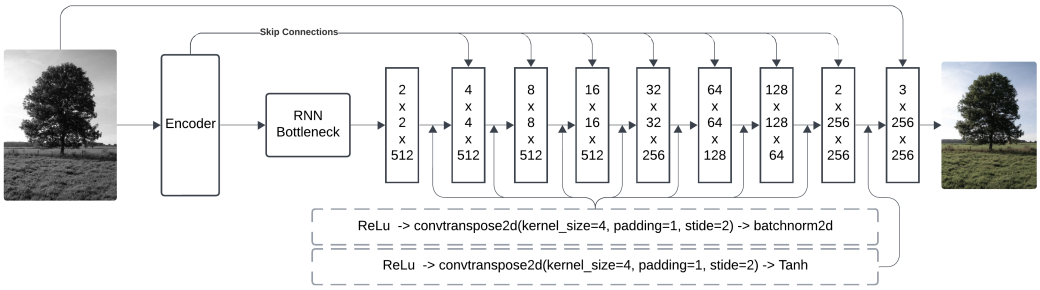

Figure 3: The final architecture of the generator: Encoder, RNN bottleneck and Decoder

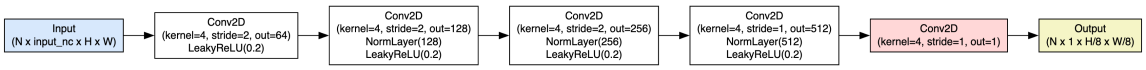

Figure 4: Discriminator

The PatchGAN discriminator is a convolutional neural network designed to classify whether image patches are real (from the dataset) or fake (generated by the generator). Figure 4 illustrated the architecture of the PatchGAN discriminator as described. The network started with an initial convolutional layer that reduced the spatial resolution by half while increasing the number of channels (ndf). It was followed by several intermediate layers (defined by n_layers) that repeatedly downsampled the image using convolution, normalization, and LeakyReLU activation. Each successive layer doubled the number of filters up to a maximum of ndf * 8. The final layers consisted of a convolutional block with stride 1, increasing the depth of features, and an output layer that generated a single-channel feature map representing the discriminator's judgment on patch-level realism. The overall architecture progressively extracted hierarchical features and reduces spatial dimensions, allowing it to focus on local details at the patch level, making it suitable for tasks like image-to-image translation.

6 Results

Our experimental results demonstrate both the strengths and limitations of our RNN-enhanced U-Net + PatchGAN architecture compared to the baseline U-Net + PatchGAN model by Isola et. al [4]. We present a series of test cases that highlight various scenarios where our model shows improvement, as well as areas where both approaches face challenges. In scenarios involving distant landscapes and architectural elements, both models demonstrate strong performance. Figure 5 shows the colourization results for a castle and a volcanic landscape. These images represent ideal cases for both architectures, as they contain clear structural elements and well-defined colour boundaries. The success in these cases can be attributed to the consistent colour patterns typically found in such scenes such as blue skies, beige stone, and natural vegetation colours which are well-learned by both models during training.

Input | Isola et. al | Our model | Ground truth

Figure 5: Both models performed best in simple outdoors scenes with landscapes and buildings

Input | Isola et. al | Our model | Ground truth

Figure 6: Colourizing Animal photographs

When processing images of animals, our RNN-enhanced architecture shows notable improvements over the baseline model. Figure 6 presents colourization results for three subjects: a frog and two dogs. The baseline model exhibits a tendency to diffuse animal colours beyond their natural

boundaries, which causes visible colour bleeding in the surrounding areas. In contrast, our RNN-enhanced model demonstrates superior handling of fur textures, maintaining more precise colour boundaries and natural variation in fur patterns. This improvement can be attributed to the RNN bottleneck's ability to maintain temporal coherence across similar textural regions, which results in more natural-looking fur colourization.

Input Isola et. al Our model Ground truth

Figure 7: A Colourized Scene with a Human Subject

Human subjects present unique challenges for colourization models. Figure 7 shows results for an image of a person standing in front of a subway car. Our model demonstrates improved performance in facial colourization, producing more natural skin tones compared to the baseline. However, both models struggle with consistent skin tone application across different body parts, particularly in the hands. This limitation suggests that while our RNN enhancement improves local colour coherence, challenges remain in maintaining consistency across spatially separated but semantically related regions.

Input Isola et. al Our model Ground truth

Figure 8: Colourizing darker, less colourful images - this is the scenario where our model shows the greatest improvement over the baseline

Our model demonstrates superior performance when handling scenes with darker, more subdued colour palettes. As shown in figure 8, we tested two challenging scenarios: a scene from Game of Thrones featuring the Iron Throne, and a portrait-style album cover. In the throne scene, while the baseline model introduces artificial blue artifacts, our RNN-enhanced model more accurately captures the authentic metallic tones and muted atmosphere. Similarly, in the album cover comparison, our

model faithfully reproduces the intended dark green colour scheme, whereas the baseline model incorrectly colours the scene with orange and red tones.

These results suggest that while both architectures perform well on straightforward cases like landscapes, our RNN-enhanced model offers significant improvements in handling complex textures, maintaining colour consistency, and accurately reproducing subtle colour palettes. The most notable gains are observed in cases requiring fine texture preservation (as in animal fur) and accurate reproduction of muted or nuanced colour schemes.

7 Discussion

Our project has demonstrated significant advancements in image colourization through integrating an RNN within a U-Net + PatchGAN framework. This hybrid model has shown to enhance colour accuracy and contextual integrity across a variety of test cases, particularly in handling complex textures and maintaining colour consistency across temporal coherent scenes.
Despite these enhancements, our results also revealed limitations in the model's ability to consistently apply colours across different semantic regions, particularly in human subjects. This issue highlights a potential gap in the model's spatial understanding, which could be addressed by further refining the network architecture or training process to improve semantic segmentation capabilities.
These findings point to significant opportunities for improvement, emphasizing the need for ongoing adjustments and testing to enhance the robustness and accuracy of colourization techniques.

8 Limitations

While our model demonstrates promising results in image colourization, several notable limitations persist in both the baseline U-Net + PatchGAN architecture and our modified version with the RNN bottleneck.

(a) Input (b) Isola et. al (c) Our model (d) Ground Truth

Figure 9: Both models struggle to interpet the lighting on the scoop of ice-cream in this situation. Isola et. al incorrectly colours the shadow to the right of the scoop, while our model incorrectly interprets the darker lighting on the right half of the scoop as yellow

Both architectures face challenges when dealing with images containing high levels of noise or complex lighting conditions. In particular, the models sometimes struggle to differentiate between shadows and actual colour variations, leading to incorrect colouring in the final output. This limitation is especially pronounced in scenes with multiple light sources or strong directional lighting.
Another notable limitation concerns the handling of semantically ambiguous objects. While the models perform reasonably well with common objects that have consistent colour associations (like blue skies or green vegetation), they often produce inconsistent or unrealistic results for objects that could plausibly appear in multiple colours, such as clothing, furniture, or abstract art pieces. This

ambiguity becomes particularly challenging in indoor environments where colour choices are more arbitrary and less constrained by natural laws.

(a) Input (b) Isola et. al (c) Our model (d) Ground Truth

Figure 10: Comparison of colourization results on a sample indoor image. Note how both models struggle with complex patterns and textures

The models also demonstrate limitations in preserving fine details during the colourization process. Small textural elements or intricate patterns sometimes become blurred or lose definition, particularly in areas where colour boundaries should be sharp and well-defined. This effect is more pronounced in our RNN-enhanced model, where the sequential processing of spatial information can occasionally lead to slight smoothing of fine details.
These limitations suggest several directions for future work, including the exploration of alternative architectural designs that might better handle complex indoor scenes, the development of more efficient RNN implementations to reduce computational overhead, and the investigation of novel loss functions that could better preserve fine details while maintaining colour consistency across similar patterns.

9 Ethical Consideration

In developing our image colorization model, we considered several ethical implications related to data collection and model application. Firstly, we ensured compliance with intellectual property laws when using ImageNet and acknowledge that ImageNet may contain biases in terms of ethnicity, culture, and environment [6]. These biases could potentially affect our model's performance across different demographic groups and cultural contexts, as the training data may not equally represent all populations. Additionally, we acknowledge that the colorization of historical photographs could potentially lead to misrepresentation of historical events if colors are inaccurately applied. This is particularly concerning for historically significant images where color choices could influence public perception and understanding of past events. The automated nature of our colorization system means that while it can process images efficiently, it may not always capture the historical accuracy that a human expert with period-specific knowledge could provide.

10 Conclusion

This project developed an automated image colorization system using a modified GAN architecture with an integrated RNN bottleneck. Our results demonstrate that this hybrid approach effectively preserves both spatial content and temporal coherence, leading to more realistic color transitions compared to traditional methods. While exploring various enhancement techniques, we encountered challenges with implementing a KL-based saturation loss function, which was ultimately excluded from the final model due to initialization issues [7] [8]. This experience highlighted potential areas for future improvement, particularly in weight initialization strategies and the handling of color saturation. Moving forward, promising directions include investigating the relationships between

color saturation, brightness, and contrast, as well as developing feedback mechanisms to better capture these interactions. Despite current limitations, our RNN-enhanced model represents a significant step forward in automated image colorization, with potential applications in historical image restoration and computer vision systems.

References

[1] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, p. 689–694, Aug. 2004. [Online]. Available: https://doi.org/10.1145/1015706.1015780

[2] C. Leithart, "How peter jackson brought wwi to life in "they shall not grow old"," *Frame.io Blog*, 2019. [Online]. Available: https://blog.frame.io/2019/06/06/peter-jackson-they-shall-not-grow-old/

[3] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," *CoRR*, vol. abs/1603.08511, 2016. [Online]. Available: http://arxiv.org/abs/1603.08511

[4] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[6] R. Steed and A. Caliskan, "Image representations learned with unsupervised pre-training contain human-like biases," *CoRR*, vol. abs/2010.15052, 2020. [Online]. Available: https://arxiv.org/abs/2010.15052

[7] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," *CoRR*, vol. abs/1603.06668, 2017. [Online]. Available: http://arxiv.org/abs/1603.06668%7D

[8] C. Ballester, H. Carrillo, M. Clément, and P. Vitoria, *Analysis of Different Losses for Deep Learning Image Colorization*. Cham: Springer International Publishing, 2023, pp. 821–846. [Online]. Available: https://doi.org/10.1007/978-3-030-98661-2_127