Image Colorization using Conditional Wasserstein GANs

Sohan Arun

***Abstract*—Automatic image colorization, the task of adding plausible color to grayscale images, is a significant yet challenging problem in computer vision due to the inherent ambiguity and multi-modality of potential colorings. This paper presents an image colorization framework based on a Conditional Wasserstein Generative Adversarial Network with Gradient Penalty (cWGAN-GP). The proposed generator utilizes a ResU-Net architecture, integrating residual blocks within a U-Net structure to effectively learn hierarchical features and preserve fine-grained details necessary for realistic color synthesis. The model operates in the Lab color space, using the L (lightness) channel as conditional input to predict the corresponding a and b (chrominance) channels. A convolutional critic network, trained to differentiate between real and generated color distributions using the Wasserstein distance and stabilized by a gradient penalty and R1 regularization, guides the generator. The generator is optimized through a combination of an L1 reconstruction loss and the adversarial loss. Qualitative visual assessments and quantitative metrics, including Inception Score (IS) and Fréchet Inception Distance (FID), are employed to evaluate the performance, demonstrating the model's capability to produce high-quality, perceptually convincing colorized images with improved training stability.**

***Index Terms*—Image Colorization, Conditional GAN, Wasserstein GAN, WGAN-GP, ResU-Net, L*a*b* Color Space, Deep Learning**

I. Introduction

Image colorization, the process of adding color to grayscale images, holds significant value in various applications, from restoring historical photographs and films to enhancing the visual appeal and comprehensibility of scientific or artistic content. Despite its importance, automatic colorization presents a substantial challenge due to the inherent ambiguity of the problem: a single grayscale image can have multiple plausible and perceptually valid colorizations. This ill-posed nature makes it difficult for traditional algorithms to produce consistently realistic results [7]. This paper addresses this challenge by proposing a deep learning model designed to automatically generate plausible and realistic full-color images. Specifically, the primary objective is to develop a system that takes the lightness (L) channel of an image in the L*a*b* color space as input and predicts the corresponding chrominance (a* and b*) channels to reconstruct a vibrant and natural-looking color image [7].

A. Related Work

Early approaches to automatic image colorization often relied on user-guided scribbles or matching statistics between reference color images and target grayscale images [7]. The advent of Generative Adversarial Networks (GANs) revolutionized image generation, enabling the synthesis of highly realistic images [1]. Conditional GANs (cGANs), notably exemplified by the Pix2Pix framework, extended this capability to image-to-image translation tasks, directly learning mappings between input and output image domains [2]. However, training standard GANs is often plagued by challenges such as mode collapse and instability [8]. Wasserstein GANs (WGANs) were introduced to mitigate these issues by employing the Wasserstein distance and a critic network, leading to more stable training and improved sample quality [3], [4]. Concurrently, U-Net architectures, with their characteristic encoder-decoder structure and skip connections, have proven highly effective for tasks requiring precise localization, such as image segmentation and translation [5]. Furthermore, Residual Networks (ResNets) have enabled the successful training of significantly deeper neural networks by facilitating gradient flow through identity shortcut connections, enhancing feature representation capabilities [6]. This study builds upon these advancements, integrating a cWGAN framework with a ResU-Net generator to tackle the image colorization problem.

B. Contribution

The primary contributions of this work are as follows:

- The application of a Conditional Wasserstein GAN with Gradient Penalty (cWGAN-GP), further augmented with R1 regularization, is presented for the image colorization task, aiming to achieve robust and stable training dynamics.
- A ResU-Net generator architecture is designed and implemented, which integrates Residual Blocks within a U-Net structure. This configuration is intended to effectively learn and reconstruct chrominance information while preserving essential image details from the grayscale input.
- The L*a*b* color space is systematically employed to simplify the learning problem by decoupling the image's lightness component from its chrominance components, thereby allowing the model to focus on predicting color.
- A comprehensive evaluation of the proposed colorization model is conducted, encompassing both qualitative visual assessments and quantitative metrics, specifically the Inception Score (IS) and Fréchet Inception Distance (FID), to gauge performance.

The remainder of this paper is organized as follows: Section II details the methodology, including the network architectures, loss functions and implementation details. Section III presents and discusses the results. Finally, Section IV and V concludes the paper and suggests future work.

II. Methodology

A. Lab* Color Space for Colorization

The L*a*b* color space is adopted for the image colorization process in this project. This color model characterizes colors through three channels: L* for lightness, a* for the green-red chromatic dimension, and b* for the blue-yellow chromatic dimension. A significant reason for choosing the L*a*b* space is its inherent separation of the achromatic lightness (L*) information from the chromatic color (a*, b*) information. This decoupling simplifies the task for the generator model; given the L* channel (which represents the input grayscale image), the network learns to predict only the a* and b* color channels. The full-color image is then reconstructed by merging the original L* channel with these predicted a* and b* channels. Such a methodology enables the generator to focus its predictive power on chrominance, conditioned on the existing luminance.

B. Dataset & Preprocessing

The dataset used in this study was sourced from the MIR-FLICKR25k image collection, which comprises 25,000 high-quality color images annotated with user-generated tags and categorized into 24 semantic groups. While the original dataset consists solely of RGB images, a preprocessed version made available by the Kaggle community was utilized in this work. In this version, each image was converted into the *L**, *a**, and *b** channels of the Lab* color space and stored as *.npy* NumPy arrays.

Due to time constraints, a subset of 5,000 image pairs, each resized to *224* × *224* pixels, was selected for both training and evaluation. During data loading, the NumPy arrays were converted to PyTorch tensors, which inherently normalized the pixel values to the *[0, 1]* range. For visualization and RGB reconstruction, the *L** channel tensor was rescaled by multiplying by 100 to restore the expected *[0, 100]* range, while the *a** and *b** channels were transformed to approximate the *[-128, 128]* range using:

$$ab_{\text{tensor}} = (ab_{\text{tensor}} - 0.5) \times 256 \tag{1}$$

It is important to note that the same set of 5,000 samples was used for both training and testing the model. To introduce variability during training, the dataloader employed random shuffling, while the test dataloader maintained a fixed sample order to ensure consistent evaluation.

C. Generator Architecture: ResU-Net

The generator is implemented as a ResU-Net that converts a single–channel L* input ($1 \times 224 \times 224$) to a two–channel a*b* output (2x224x224) of identical spatial resolution (Fig. 1). It follows the U-Net encoder–decoder template, but incorporates residual learning and bilinear up-sampling. Key design details are:

1) **Encoder blocks:** Each down-sampling unit consists of a 2×2 max-pool with stride as 2 followed by a ResBlock. The ResBlock stacks two 3×3 conv – BN – ReLU layers (*stride* 1, padding 1) with an identity shortcut; feature-map depth doubles at successive levels ($64 \to 128 \to 256 \to 512$), enabling the network to learn increasingly abstract context.
2) **Bridge:** A single ResBlock at the bottleneck provides the deepest feature representation.
3) **Decoder blocks:** Up-sampling is performed with bilinear interpolation, rather than a learnable transposed convolution, thereby avoiding checkerboard artefacts. The up-sampled feature map is concatenated with the correspondingly sized encoder output (skip connection) and passed through another ResBlock. Channel depth is halved symmetrically ($512 \to 256 \to 128 \to 64$). No explicit cropping is required because symmetric padding keeps feature-map sizes aligned.
4) **Residual blocks:** Each ResBlock follows the pattern Conv – BN – ReLU – Conv – BN with a post-addition ReLU. This arrangement preserves gradients and allows deeper feature reuse.
5) **Regularisation:** Dropout value of 0.2 is applied after every encoder block and the bridge to mitigate over-fitting while leaving the decoder path dropout-free.

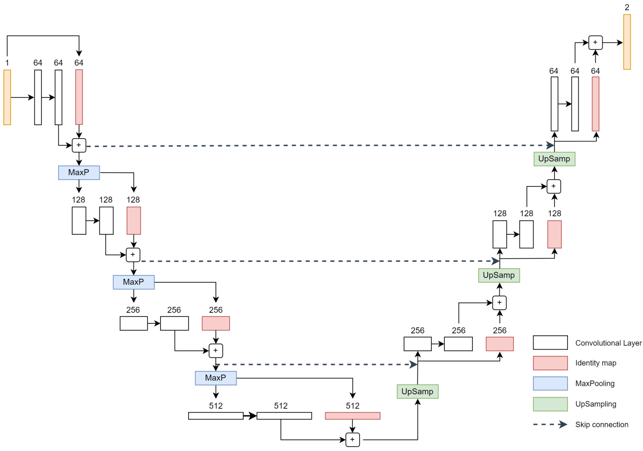

Fig. 1. ResU-Net generator. Left: encoder (down-sampling), centre: bottleneck bridge, right: decoder (up-sampling). Skip connections propagate low-level detail.

D. Critic Architecture

The critic network is a convolutional neural network (CNN) designed to differentiate between real and generated colorized images, conditioned on the input L* channel. Its architecture comprises a sequence of convolutional blocks. Each block typically consists of a 4x4 convolutional layer with a stride of 2 for downsampling, followed by Instance Normalization (except in the first block) and a LeakyReLU activation function with a negative slope of 0.2.

The input to the critic is a 3x224x224 tensor, formed by concatenating the L* channel (the conditioning grayscale image) with either the ground truth a*b* channels (for real samples) or the a*b* channels predicted by the generator (for fake samples). After several such convolutional blocks that progressively reduce spatial dimensions and increase feature depth (3 → 64 → 128 → 256 → 512 filters), an adaptive average pooling layer reduces each feature map to a single value. This is followed by a flattening operation and a final linear layer that outputs a single scalar value. This scalar is used inside the Wasserstein loss (Section II-E) to indicate how *real* or *fake* a colourisation is. Fig. 2 depicts the pipeline.

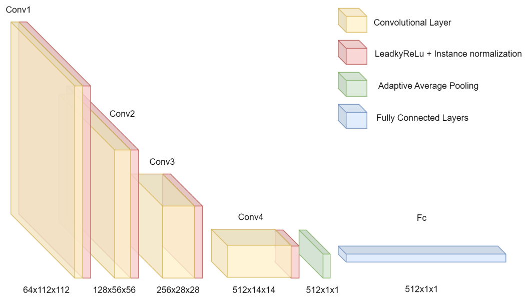

Fig. 2. Critic Network Architecture. The diagram shows the sequence of convolutional blocks, each performing downsampling. The input is the L* channel concatenated with either real or generated a*b* channels. The network outputs a single scalar value.

E. Training Framework: Conditional WGAN–GP

Both networks are trained in a Conditional Wasserstein GAN with Gradient Penalty (cWGAN–GP) setting, where the L^* channel acts as the common condition.

Critic loss: The critic seeks to maximise the *Wasserstein-1* distance between real and generated joint distributions:

$$\mathcal{L}_D^{\text{WGAN}} = \mathbb{E}_{(\mathbf{L},\mathbf{ab})\sim P_{\text{real}}}\big[D(\mathbf{L},\mathbf{ab})\big] - \mathbb{E}_{(\mathbf{L},\hat{\mathbf{ab}})\sim P_G}\big[D(\mathbf{L},\hat{\mathbf{ab}})\big]. \tag{2}$$

Gradient penalty: To satisfy the 1-Lipschitz condition required by the Wasserstein metric, we add the gradient penalty[4]:

$$\text{GP} = \mathbb{E}_{\hat{\mathbf{ab}}}\Big[\big(\|\nabla_{\hat{\mathbf{ab}}} D(\mathbf{L},\hat{\mathbf{ab}})\|_2 - 1\big)^2\Big], \tag{3}$$

where $\hat{\mathbf{ab}}$ is an interpolation between real and fake chrominance channels.

R1 regularisation: An additional R1 term penalises the squared gradient norm on the same interpolates, further smoothing the critic:

$$\text{R1} = \mathbb{E}_{\hat{\mathbf{ab}}}\big[\|\nabla_{\hat{\mathbf{ab}}} D(\mathbf{L},\hat{\mathbf{ab}})\|_2^2\big]. \tag{4}$$

Total critic objective:

$$\mathcal{L}_D = -\mathcal{L}_D^{\text{WGAN}} + \lambda_{\text{gp}}\,\text{GP} + \lambda_{\text{r1}}\,\text{R1}, \quad \lambda_{\text{gp}} = \lambda_{\text{r1}} = 10. \tag{5}$$

Generator loss: The generator is trained using an L1 reconstruction loss, which encourages pixel-wise accuracy in predicting the a^*b^* chrominance channels. Although no explicit adversarial loss is applied, the generator implicitly benefits from adversarial feedback through the alternating update scheme with the critic.

$$\mathcal{L}_G = \lambda_{\text{recon}}\,\mathbb{E}\big[\|\hat{\mathbf{ab}} - \mathbf{ab}\|_1\big], \quad \lambda_{\text{recon}} = 100, \tag{6}$$

F. Metrics

Model performance was assessed using a two-tier strategy, combining visual checks with established quantitative scores. All quantitative metrics were derived from features extracted using a pre-trained Inception-v3 model, ensuring a consistent evaluation pipeline.

1) **Qualitative inspection:** Side-by-side visual comparison of the input L^* channel, the generator's colorized output, and the corresponding ground-truth color image. This step is crucial for identifying perceptual artifacts such as color bleeding, unnatural hues, or lack of detail that numerical scores might not fully capture.
2) **Quantitative metrics:** Two standard scores for generative models were employed:
   - **Inception Score (IS).** This metric evaluates both the quality (recognizability of objects) and diversity of the generated images. A pre-trained Inception-v3 model is used to obtain class probability distributions $p(y|x)$ for each generated image x. The IS is then calculated based on the KL divergence between this conditional distribution and the marginal class distribution $p(y)$ (approximated from the generated samples). Higher IS values generally indicate better image quality and diversity.
   - **Fréchet Inception Distance (FID).** FID measures the similarity between the distribution of real images and the distribution of generated images in the Inception-v3 feature space. Features are extracted for both sets of images. The mean vector (μ) and element-wise standard deviation vector (σ) of these features are calculated for both the real and generated image sets. The FID, as implemented in this project, is then computed using these statistics as:

     $$\text{FID} = 2 \cdot \|\mu_{\text{real}} - \mu_{\text{gen}}\|_2^2 + \|\sigma_{\text{real}} - \sigma_{\text{gen}}\|_2^2.$$

     Lower FID values suggest that the distribution of generated images is closer to that of real images, indicating higher fidelity and diversity.

G. Hyperparameter Configuration

The training of the cWGAN-GP model involved a specific set of hyperparameters, crucial for achieving the reported performance.

- **Optimizers:** Both the generator and the critic networks were optimized using the Adam optimizer.
  - Learning Rate: 2×10^{-4} for both optimizers.
  - Betas: $(\beta_1 = 0.5, \beta_2 = 0.9)$ for both Adam optimizers.
- **Loss Function Weights:**
  - L1 Reconstruction Loss (λ_{recon}): 100. This term encourages the generator to produce outputs close to the ground truth chrominance channels.
  - Gradient Penalty (λ_{gp}): 10. This weight scales the gradient penalty term in the critic's loss, enforcing the 1-Lipschitz constraint.

  - R1 Regularization (λ_{r1}): 10. This weight scales the R1 regularization term, further stabilizing the critic by penalizing large gradients with respect to interpolated samples.
- **Training Parameters:**
  - Batch Size: 1. Due to the nature of image-to-image translation and available computational resources, training was conducted with individual samples per batch.
  - Epochs: 150. The model was trained for a total of 150 passes over the entire training dataset.

III. Results and Discussion

A. Results

The performance was evaluated by calculating the Inception Score and Fréchet Inception Distance. Table I presents these scores. The Inception Score for the generated images is compared against the score for real images from the test set, providing a reference. The FID score measures the distance between the feature distributions of real and generated images.

TABLE I
Quantitative Evaluation Metrics for Image Colorization.

Metric	Real Images	Generated Images
IS	4.3412 ± 1.7963	4.3125 ± 1.8372
FID	N/A	13.1076

Note: IS values are reported as mean ± standard deviation.

Qualitative results are presented in Fig. 3. These results display side-by-side comparisons of the input grayscale L channel, the corresponding ground truth color image, and the colorized image produced by the generator .

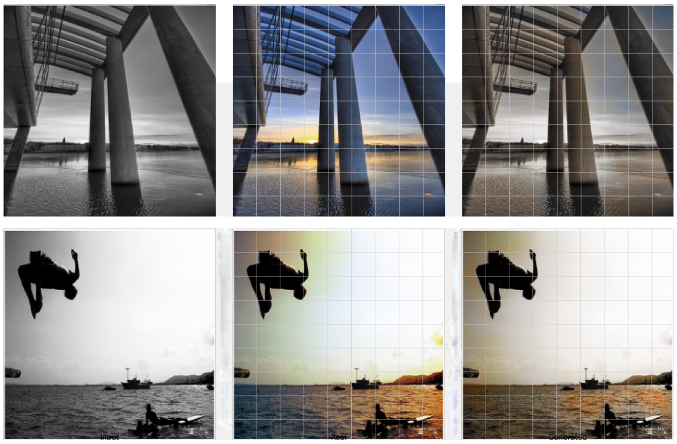

Fig. 3. Examples of image colorization. From left to right in each row: Input Grayscale, Ground Truth Color, Generated Color.

B. Analysis

The quantitative results provide an initial measure of the colorization performance. The IS for the generated images is notably close to the IS of the real images from the test set. This similarity suggests that the generated images possess a comparable degree of feature diversity and object recognizability as the ground truth images, according to the Inception v3 model.

The FID score of 13.1076 quantifies the similarity between the distributions of real and generated images in the feature space of the Inception network. While optimal IS and FID values are dataset-dependent and task-specific, these metrics serve as a standardized benchmark for the model's generative capabilities.

A qualitative assessment of the generated images, as produced during the model inferencing phase, indicates that the model generally produces plausible colorizations. The colors assigned to various objects and scenes often appear natural and contextually appropriate. However, as it is common with generative colorization models, certain artifacts may be present in some outputs. These can include minor instances of color bleeding, where colors extend slightly beyond object boundaries, or patches that may appear somewhat dull or, conversely, overly saturated. The model's ability to handle diverse scenes and objects is generally robust, though the fidelity of colorization and the presence of artifacts can vary depending on the complexity of the input grayscale image, the uniqueness of the objects, and the subtlety of the lighting conditions. For instance, scenes with common objects and clear lighting tend to be colorized more accurately than those with unusual textures or ambiguous illumination.

C. Discussion

The implemented image colorization framework demonstrates several strengths inherent in its design. The adoption of the Conditional Wasserstein GAN with Gradient Penalty framework contributed to a relatively stable training process, mitigating common GAN training issues such as mode collapse, which is often a challenge. This stability is further supported by the inclusion of R1 regularization in the critic's loss function. The ResU-Net architecture employed for the generator proved effective in capturing both global color consistencies across the image and finer local details, facilitated by its skip connections and residual blocks.

The choice of the L*a*b* color space was beneficial, as it allowed the model to focus on predicting the chrominance (a* and b*) channels conditioned on the lightness (L) channel, simplifying the learning task compared to directly predicting RGB values.

Despite its strengths, the model exhibits certain limitations. Like many image colorization approaches, it may occasionally struggle with unusual or abstract color combinations that are underrepresented in the training data. Additionally, achieving perfect color accuracy in fine textures remains challenging. While the current dataset was adequate for demonstrating proof of concept, using a larger and more diverse dataset could further enhance the model's performance, robustness, and generalization to unseen image types.

IV. Conclusion

This research presented the development and evaluation of a generative deep learning framework for automatic image colorization, centered on a Conditional Wasserstein Generative Adversarial Network with Gradient Penalty and a ResU-Net generator. The model, operating within the L*a*b* color space, was trained to predict the a* and b* chrominance channels conditioned on the L lightness channel. The empirical results demonstrate the system's capacity to generate colorizations that are both perceptually realistic and plausible for previously unseen grayscale images.

Quantitative evaluation demonstrated strong performance: the Inception Score for generated images closely matched that of real images, and a Fréchet Inception Distance of 13.1076 was achieved. These metrics, supported by qualitative visual assessments, affirm the model's colorization capabilities.

V. Future Work

While the current model demonstrates strong performance, potential avenues for future research could build upon the presented work. Firstly, expanding the training dataset to include a larger and more diverse collection of images could enhance the model's generalization capabilities to varied visual content. Secondly, further exploration into the loss function components and hyper-parameter optimization could yield performance improvements. This includes investigating alternative reconstruction losses beyond the L1 loss, or fine-tuning the weighting parameters (λ_{recon}, λ_{gp}, λ_{r1}) for the generator and critic. Thirdly, investigating more advanced neural network architectures for both the ResU-Net based generator and the critic could unlock further potential. This might involve exploring attention mechanisms, newer residual block variants, or different critic designs. Finally, applying the developed colorization framework to specific, challenging domains, such as the colorization of historical photographs or medical imagery, presents an interesting direction for practical application and further refinement of the model.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1125–1134.

[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 214–223.

[4] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5767–5777.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[7] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 649–666.

[8] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2234–2242.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6626–6637.

[10] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 8026–8037.

[11] W. Falcon and The PyTorch Lightning team, *PyTorch Lightning*. GitHub, 2019. Accessed: May 24, 2025. [Online]. Available: https://github.com/PyTorchLightning/pytorch-lightning

[12] S. van der Walt et al., "Scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, Jun. 2014, doi: 10.7717/peerj.453.