Image Colorization using U-Net and GAN

V. Guna Chowdary
Department of Computer Science and Engineering
PES University
Bangalore, India
v.gunachowdary@gmail.com

Vishnu Prakash
Department of Computer Science and Engineering
PES University
Bangalore, India
vishnuprakash156@gmail.com

Praneeth Cheepurupalli
Department of Computer Science and Engineering
PES University
Bangalore, India
chpraneeth2003@gmail.com

Swastika sinha
Department of Computer Science and Engineering
PES University
Bangalore, India
Sinhaswastika07pes@gmail.com

Shylaja S S
Department of Computer Science and Engineering
PES University
Bangalore, India
shylaja.sharath@pes.edu

***Abstract*—This paper introduces a novel method for colorizing greyscale images utilizing a conditional Generative Adversarial Network (GAN) architecture with U-Net encoding, trained exclusively on the COCO dataset. Employing the LAB color space for training, the model learns to predict the 'AB' channels based on the 'L' channel of greyscale images. By conditioning the generator on both the greyscale input and the corresponding ground truth color image in the LAB color space, the model captures intricate color details while preserving image structure. Experimental evaluation demonstrates superior performance in generating realistic and vibrant colorizations compared to existing methods.**

***Index Terms*—L*a*b colorspace, U-Net, GAN, COCO Dataset**

I. Introduction

The task of imagining their colors may initially seem overwhelming, given that a significant portion of the information (two of the three dimensions) has been omitted. Upon closer examination, however, it becomes apparent that many regions within each image offer clues based on the scene's semantics and surface texture: grass typically appears green, the sky usually appears blue, and a ladybug is commonly red. Admittedly, these semantic hints may not apply universally; for instance, the colors of croquet balls resting on the grass may not accurately reflect reality, though they could be a reasonable guess. Nonetheless, the objective of this paper is not necessarily to recreate the precise true colors but rather to generate a convincing colorization that could potentially deceive a human observer.

Colorization of images involves the task of predicting RGB colors for grayscale images or frames in videos to enhance their visual appeal and perceptual quality. Advances in deep learning methods for image colorization have been significant over the past decade, highlighting the necessity for a comprehensive review and evaluation of these methods.

Every image has a rank-3 (height, width, color) array with the last axis containing the color data. RGB color space has 3 numbers for each pixel indicating how much Red, Green, and Blue the pixel is.In L*a*b color space, we have again three numbers for each pixel but these numbers have different meanings. The first number (channel), L, encodes the Lightness of each pixel and when we visualize this channel (the second image in the row below) it appears as a black and white image. The *a and *b channels encode how much green-red and yellow-blue each pixel is, respectively.

When using L*a*b, we can give the L channel to the model (which is the grayscale image) and want it to predict the other two channels (*a, *b) and after its prediction, we concatenate all the channels and we get our colorful image. But if we use RGB, we have to first convert our image to grayscale, feed the grayscale image to the model and hope it will predict 3 numbers for you which is a way more difficult and unstable task due to the many more possible combinations of 3 numbers compared to two numbers.

We have employed the pix2pix framework, which utilizes Conditional Generative Adversarial Networks (CGANs) and L1 loss, to address the image colorization task. In our proposed approach, a generator model takes grayscale images as input and produces colorized versions, while a discriminator evaluates the realism of the generated colorizations. Both the generator and discriminator are conditioned on the grayscale input images, allowing them to incorporate contextual information during training. The training process involves minimizing both L1 loss, which measures the pixel-wise difference between generated and ground truth color images, and adversarial loss, which encourages the generator to produce colorizations

that are indistinguishable from real images according to the discriminator. This combination of techniques enables the generation of plausible colorizations that closely resemble real images, improving the aesthetic and perceptual quality of grayscale photographs.

II. Related Work

In recent years, many solutions have been proposed to colourize images using deep learning. *Automatic Image Colorization with Simultaneous Classification* [1] presented a technique to automatically colourize grayscale images that combines both global priors and local image features. Based on Convolutional Neural Networks, the deep network features a fusion layer that allows to elegantly merge local information dependent on small image patches with global priors computed using the entire image. The entire framework, including the global and local priors as well as the colorization model, is trained in an end-to-end fashion. The paper leverages an existing large-scale scene classification database to train the model, exploiting the class labels of the dataset to more efficiently and discriminatively learn the global priors. The model consists of four main components: a low-level features network, a mid-level features network, a global features network, and a colorization network. The output is the chrominance of the image which is fused with the luminance to form the output image.

Colourful Image Colorization [2] suggests ways to solve the problem of hallucinating a plausible colour version of the photograph. It proposes a fully automatic approach that produces vibrant and realistic colorizations. The problem is posed as a classification task and uses class-rebalancing at training time to increase the diversity of colours in the result. The system is implemented as a feed-forward pass in a CNN at test time and is trained on over a million colour images. They have evaluated the algorithm using a "colorization Turing test," asking human participants to choose between a generated and ground truth colour image. It successfully fools humans on 32% of the trials, significantly higher than previous methods.

Image colorization using similar images [3] presents a new example-based method to colourize a grey image. As input, the user needs to supply a reference colour image which is semantically similar to the target image. Features are extracted from these images at the resolution of superpixels, and are used for the colorization process. A superpixel representation speeds up the colorization process. It also empowers the colorizations to exhibit a much higher extent of spatial consistency in the colorization as compared to that using independent pixels. They have adopted a fast cascade feature matching scheme to find correspondences between superpixels of the reference and target images. Each correspondence is assigned a confidence based on the feature matching costs computed at different steps in the cascade, and high confidence correspondences are used to assign an initial set of chromatic values to the target superpixels. To further enforce the spatial coherence of these initial colour assignments, an image space voting framework draws evidence from neighbouring superpixels to identify and to correct invalid colour assignments.

III. Methodology

One of the most important aspects into colorising an image, is loading the image itself. When an image is loaded onto a machine learning model, a rank-3 array will be generated, where the last axis contains the color data for the image. There are multiple color spaces like RGB, L*a*b among many others. In this project we will be using the L*a*b color space. In this color space, there are three numbers for each pixel: first number, L, encodes the lightness of each pixel and the *a and *b numbers indicate how green-red and yellow-blue a pixel is respectively. The reason for using L*a*b color space is that, we can only pass the L channel of an image and make the model predict the *a and *b channels. Doing this is not intuitive and efficient for other color spaces. In our approach two losses are used: L1 loss, which makes it a regression task, and adverserial GAN loss which can be used to solve the problem in a unsupervised way.
In a Generative Adversarial Network there is a Generator model and a Discriminator model that work together to learn to solve a problem. In this setting, a grayscale image is passed into the generator, which then generates a color image with the *a and *b channels. The discriminator takes these two channels and compares it with the input grayscale image and decides whether the image is real or fake. Taking a look at the math, let's consider $\boldsymbol{x}$ as the grayscale image, $\boldsymbol{z}$ as the input noise for the generator, and $\boldsymbol{y}$ as the 2-channel output we want from the generator. The loss function for the conditional GAN is shown in Fig. 1, where G represents the Generator and D represents the Discriminator.

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))]$$

Fig. 1. Loss function for a GAN

To further improve the model and help introduce some supervision in the task we introduce the L1 loss. L1 loss is preferred over L2 loss because it reduces the effect of producing images that are predominantly gray.This measures the loss between the predicted colors and the actual colors as shown in Fig. 2.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1]$$

Fig. 2. L1 Loss function

Using L1 loss alone, the model will be conservative and not produce the best results. Hence, to get the best of both

worlds, the adversarial GAN loss the L1 loss are combined. The combined loss function is shown in Fig. 3.

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G)$$

Fig. 3. Loss function for a GAN

In the Fig.3 the lambda is a coefficient introduced into the equation to balance the contribution of the two losses into a single consolidated loss.

A. Dataset

For this research we have the COCO dataset. Since the dataset is quite large, having 330,000 images, due to resource constraints, we have taken only 8000 images and have used them for training. For this task, any dataset will work as long as it contains different scenes and locations for better learning of the model.

B. Implementation

In our implementation, we have used the U-Net model as the generator for the GAN. The important aspect to understand here is that, down-sampling and up-sampling methods are added to the middle module at every iteration till it reaches the input and the output modules. The architecture of the U-Net can be seen in Fig. 4. The blue rectangle shows

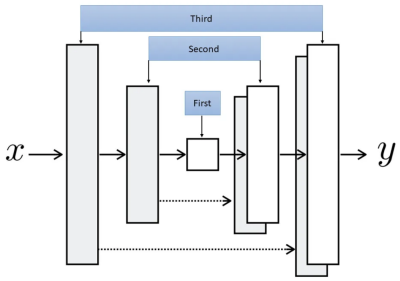

Fig. 4. U-Net Architecture are used in the paper

the order in which the modules are built.

The architecture for the discriminator is straight forward. We implemented a model by stacking Conv-BatchNorm-LeakyReLU layers to tell whether the image is real or fake. We are using a Pitch Discriminator here, where he model outputs one number for every patch of n by n pixels of the input image and for each of them decides whether the image is fake or real separately. Using such a model for the task of colorization is more accurate because the local changes that the model needs to make are very important and deciding on the whole image as in vanilla discriminator cannot take care of the subtleties of this task.

IV. Results

We have trained this model with this dataset on Google Colab, using the T4 GPU. Each training epoch takes around 3-4 minutes. Substantial results can be observed after 20-25 epochs. The results can be seen in Fig. 5.

Fig. 5. Results

V. Conclusion and Future Work

In conclusion, this research has demonstrated a promising approach to greyscale image colorization using a conditional GAN architecture with U-Net encoding, trained exclusively on the COCO dataset and utilizing the LAB color space.

The experimental results underscore the superiority of our method over existing approaches, particularly with a greater number of epochs and increased computing power, suggesting the potential for achieving even higher accuracy. Furthermore, our GAN architecture has shown promise beyond colorization, exhibiting proficiency in image segmentation tasks. This research contributes to the advancement of image colorization techniques and underscores the potential of deep learning models for image segmentation tasks. Future endeavors could explore optimization strategies and extend the approach to diverse datasets to further enhance performance and applicability.

Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks . . .”. Instead, try “R. B. G. thanks. . .”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

[1] Satoshi Iizuka, Edgar Simo-Serra, Hiroshi Ishikawa, “Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification” in SIGGRAPH 2016.

[2] Richard Zhang, Phillip Isola and Alexei A. Efros, “Colorful Image Colorization” at University of California, Berkeley.

[3] Raj Kumar Gupta, Alex Yong-Sang Chia, Deepu Rajan, Ee Sin Ng and Huang Zhiyong, “Image colorization using similar images” in MM ’12, Proceedings of the 20th ACM international conference on Multimedia.

[4] Qing Luan1, Fang Wen, Daniel Cohen-Or, Lin Liang, Ying-Qing Xu and Heung-Yeung Shum, “Natural Image Colorization” at University of Science and Technology of China, 2007.

[5] Saeed Anwar, Muhammad Tahir, Chongyi Li, Ajmal Mian, Fahad, Shahbaz Khan, Abdul Wahab Muzaffar, “Image Colorization: A Survey and Dataset” on 25 Aug 2020.

[6] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong and David Forsyth, “Learning Diverse Image Colorization” at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[7] Shanshan Huang, Xin Jin, Qian Jiang and Li Liu “Deep learning for image colorization: Current and future prospects”, September, 2022.
[8] Guillaume Charpiat, Matthias Hofmann and Bernhard Schölkopf “Automatic Image Colorization Via Multimodal Predictions” at European Conference on Computer Vision, 2008.