"Image Colorization with WGAN"

Giacomo Gonella†, Pietro Picardi†, Filippo Zanatta†

***Abstract*—This paper presents a novel approach for the colorization of grayscale images using conditional Generative Adversarial Networks (cGANs). The field of computer vision and image processing has seen a significant increase in the development of deep learning techniques for solving various tasks, including image colorization. Our approach tries to advance the state of the art by using a Wasserstein Generative Adversarial Network (WGAN) and the Structural Similarity Index (SSIM) as evaluation metric to train the network. The generator network is implemented using a UNet architecture which leverages the power of skip connections to improve the quality of the generated images. The discriminator network is based on a PatchGAN, which allows for the effective discrimination of the generated images while keeping the network relatively lightweight. Our results aim at showing that the use of a WGAN and SSIM loss results in more visually appealing colorization compared to traditional GANs and other loss functions.**

***Index Terms*—cGAN, UNet, PatchGAN, WGAN, Colorization**

I. Introduction

Generative Adversarial Networks (GANs) are a class of deep learning algorithms that have seen widespread use in various domains such as image synthesis, style transfer, and more recently, image colorization. In particular, the Conditional Generative Adversarial Network (cGAN) architecture has been proven to be particularly effective in the colorization task, where the goal is to automatically color grayscale images.

Colorization is a challenging task as it requires a deep understanding of the semantic and the contextual relationships between colors and objects in an image. Traditional colorization methods are based on hand-designed rules and heuristics, which can lead to inconsistent results and a lack of control over the final output. cGANs, on the other hand, provide a flexible and scalable solution to this problem by allowing the model to learn the relationships between colors and objects directly from data. With the ability to control the colorization process through input conditioning, cGANs have shown remarkable results in achieving photorealistic colorization. In this paper, we will focus on the application of cGANs in the colorization task.

This paper presents a novel approach to colorization using cGANs that leverages Wasserstein Generative Adversarial Network (WGAN) training and the Structural Similarity Index (SSIM) metric. The use of WGAN training offers several advantages over traditional GAN training methods, such as improved stability and a clearer notion of distance between the generated and real data distributions. The SSIM metric, on the other hand, provides a more robust and sophisticated evaluation of the quality of the generated color images, taking into account not only pixel-level differences but also structural information.

The combination of WGAN training and SSIM evaluation is expected to result in better colorization results as it enables the model to generate more photorealistic color images and provides a more accurate evaluation of the quality of these images. In this paper, we aim to demonstrate that this approach can significantly improve the performance of cGANs in the colorization task and to explore the potential limitations and trade-offs of this approach. Through this research, we hope to provide valuable insights into the design and optimization of cGANs for colorization and other image-to-image translation tasks.

†Department of Information Engineering, University of Padova,
email: {name.surname}@studenti.unipd.it

II. Related Work

In the colorization task, cGANs have shown promising results in generating high-quality colored images from grayscale inputs. The pix2pix project [1], in particular, has demonstrated the effectiveness of using cGANs for image-to-image translation tasks. The project utilizes a pix2pix architecture, where a generator network generates a colored output image from a grayscale input image, and a discriminator network is trained to distinguish between real and fake outputs. In several experiments, pix2pix has shown the ability to generate photorealistic colored images, preserving the fine details and semantic structures of the original grayscale images.

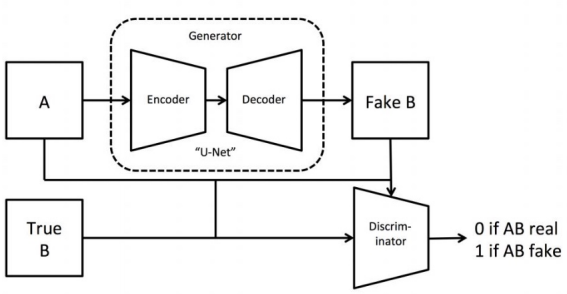
The pix2pix approach has also been applied to other image-to-image translation tasks, including style transfer and semantic segmentation, demonstrating the versatility of the cGAN architecture. The success of cGANs in the colorization task highlights their potential as a tool for a wide range of image-to-image translation tasks, and provides a foundation for further research in this area.

This paper builds upon the successful pix2pix project and aims to achieve improved results in the colorization task through the use of Wasserstein GAN (WGAN) training and the Structural Similarity Index (SSIM) metric. WGAN training has been shown to address some of the challenges faced by traditional GANs, such as instability in the training process and mode collapse, resulting in more stable and consistent results. The SSIM metric provides a more robust evaluation of image quality compared to traditional metrics such as mean squared error, making it better suited for evaluating the performance of cGANs in the colorization task.

By incorporating WGAN training and the SSIM metric, this paper aims to demonstrate that cGANs can be further optimized to produce even higher-quality results in the coloriza-

tion task. Through this research, we hope to provide insights into the strengths and limitations of cGANs and further advance the state of the art in image-to-image translation.

III. Processing Pipeline



The proposed cGAN architecture for the colorization task consists of a generator network, UNet, and a discriminator network, PatchGAN. The generator network takes in a grayscale image as input and produces a corresponding colored image as output. The discriminator network, on the other hand, takes both the grayscale image and the generated colored image as input and classifies whether the generated image is real or fake. The two networks are trained adversarially where the generator network tries to produce images that are indistinguishable from real colored images, while the discriminator network tries to correctly distinguish real and fake images.

Autoencoder and PatchGAN are two types of neural networks that have been utilized in the design of the cGAN architecture. An autoencoder is a neural network that is trained to reconstruct its input, and the UNet used in this work is a variant of autoencoder that has a symmetrical architecture. The use of UNet allows for the preservation of low-level information, such as edges and textures, in the colorization process.

PatchGAN is a type of discriminator network that operates on image patches instead of full images, making it more computationally efficient. Additionally, it has the advantage of being able to capture local variations in the image and produce more fine-grained results compared to traditional discriminators that work on full images. By using PatchGAN as the discriminator network, the cGAN is able to produce more accurate and high-quality colorized images compared to other possible solutions.

IV. Colorspaces

In our experiments, we tested both the RGB and the Lab color spaces for colorization, using two separate training processes for each. All images (taken from the Coco dataset) were resized to a shape of 256x256 before being fed into the cGAN. The use of two different color spaces was motivated by the desire to compare the results obtained from each and to understand the impact of the choice of color space on the performance of cGANs for colorization.

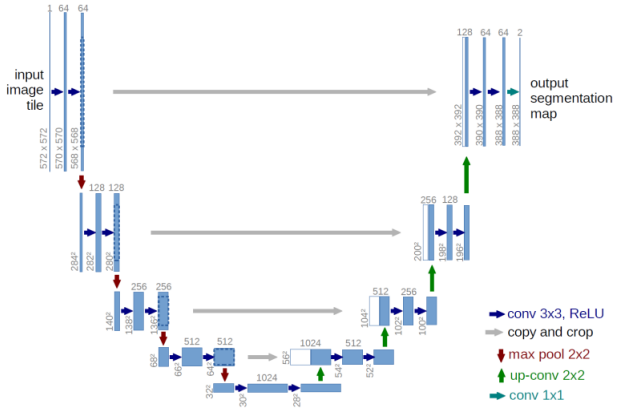
The RGB color space represents colors as a combination of red, green, and blue values, while the Lab color space represents colors as a combination of lightness and two color channels, a and b. The use of the Lab color space has several advantages over the RGB color space for the colorization task. First, the Lab color space is perceptually uniform, meaning that equal steps in the a and b channels correspond to equal perceived changes in color. This makes it easier for the model to learn the relationships between colors and objects in the image. Second, the Lab color space separates the luminance information (represented by the L channel) from the chrominance information (represented by the a and b channels). This allows the model to more easily control the brightness and saturation of the generated colors, leading to more photorealistic results.

Another advantage of the Lab color space is that it is less sensitive to the changes in the lighting conditions compared to the RGB color space. In RGB, small changes in the lighting conditions can result in large changes in the pixel values, making it difficult for the model to learn the underlying color relationships. In Lab, the L channel represents the luminance information, which is relatively insensitive to changes in the lighting conditions, allowing the model to more easily learn the color relationships.

In summary, the use of the Lab color space for colorization has several advantages over the RGB color space, including the perceptually uniform representation of colors, the separation of luminance and chrominance information, and the insensitivity to changes in lighting conditions. These advantages are expected to result in better performance for cGANs in the colorization task.

V. Learning Framework

- **UNet:**



The UNet architecture is a fully convolutional neural network used in image segmentation tasks. It consists of an encoder and a decoder, both made up of a series of convolutional, batch normalization, and activation layers. The encoder downsamples the input image, reducing the spatial dimensions and increasing the number of channels to capture high-level features. On the other hand, the decoder upsamples the feature maps and combines them with the information from the encoder through skip connections. These connections allow the

decoder to access both high-level features from the encoder and the fine-detail information from the input, resulting in a more precise segmentation.
For the colorization task, the UNet has been used as the generator to produce a colorized version of the grayscale input. The encoder is responsible for learning the high-level features of the input, such as object boundaries and shapes, while the decoder uses the skip connections to add color information to these features. This approach allows the generator to balance the trade-off between preserving the fine-detail information from the input and producing a plausible colorization. The use of skip connections also helps to prevent the vanishing gradient problem, enabling the network to learn deep representations of the input. Overall, the UNet architecture provides an effective and flexible solution for the colorization task, allowing for a high level of control over the output while producing visually appealing results.

- **PatchGAN:**
The PatchGAN used in this colorization task is a type of convolutional neural network designed for image-to-image translation tasks. The main idea behind the PatchGAN is to restrict the receptive field of the discriminator network to smaller image patches, rather than the full image. This allows the discriminator to effectively evaluate the local structure of the generated image and determine its realism on a patch-by-patch basis.
The PatchGAN consists of several convolutional layers followed by batch normalization and leaky ReLU activation functions. The input to the PatchGAN is the concatenation of both the generator output and the original grayscale image. This is done to condition the discriminator's output on the input, allowing it to better assess the consistency of the generated image with the original image.
The output of the PatchGAN is a binary prediction of whether each patch of the generated image is real or fake. The advantage of this approach is that it reduces the number of parameters in the network, allowing it to be trained more efficiently, and providing a more stable convergence of the model. Furthermore, it allows the generator to produce high-resolution colorizations while the discriminator focuses on local details, leading to a more realistic colorization result. The use of the PatchGAN in combination with the UNet generator results in a powerful cGAN architecture for the colorization task.

- **WGAN:**
Wasserstein GAN (WGAN) proposes a new cost function using Wasserstein distance, continuous and almost differentiable everywhere, which allows to train the model to optimality, avoiding the vanishing gradient problem present when using JS divergence. The Wasserstein distance (Earth Moverâ€™s distance) is computed between real data distribution P_r and generated data distribution P_g. Discriminator of GAN is changed to a critic in WGAN: it estimates a scalar value function, i.e. the Wasserstein distance between P_r and P_g, allowing it to act less strictly than a normal discriminator. WGAN-GP proposes an alternative to weight clipping to ensure smooth training. Instead of clipping the weights to a small range, it adds a loss term that keeps the L2 norm of the discriminator gradients close to 1. An interpolation image is used alongside the generated image before adding the loss function with gradient penalty as it helps to satisfy the Lipschitz constraint. Batch normalization is avoided for the critic (discriminator). Batch normalization creates correlations between samples in the same batch. It impacts the effectiveness of the gradient penalty which is confirmed by experiments in various papers. WGAN-GP achieves faster convergence and higher stability while training, and a better assignment of weights.

- **SSIM:**
The Structural Similarity Index (SSIM) is an index that measures the similarity between two images by taking into account the changes in structure, luminance, and contrast. It has been used in various image processing tasks such as image quality assessment, video compression, and image and video restoration. In the case of the colorization task, SSIM has been used as a metric for the evaluation of the generated images. The advantages of using SSIM for this task are that it provides a more robust measurement of image similarity compared to simple pixel-wise error metrics, such as mean squared error (MSE). It is better suited to this task as it takes into account both the differences in luminance and contrast, which are important factors in colorization, and the changes in structural information, which are also important in maintaining the overall structure of the image. Additionally, SSIM provides a more perceptually meaningful measurement of image similarity, which is essential in image processing tasks where the objective is to produce visually appealing results.

VI. Results

Let us begin by highlighting that evaluating the results of a task involving images is rather difficult since neural networks don't perceive these data as humans do and interpreting numerical metrics is quite challenging. Therefore we present to the reader both the actual generated images and some numerical measurements.

Model	L1+BCE (generator loss)	WGAN generator loss	MSE	SSIM
cGAN + RGB space	7.406	/	0.009	0.869
cGAN + Lab space	5.247	/	0.004	0.788
WGAN + RGB space	/	2.461	0.018	0.431

Looking at the generated images, the best results are obtained through the model proposed by Pix2Pix: these appear to be the more realistic ones, even though quite different from the ground truth.

Unfortunately from the experiments carried out in our project, it is not clear whether WGAN is a better option than the one proposed by Pix2Pix, despite its previously presented theoretical benefits: the results are slightly better than the one obtained exploiting the RGB color space, but clearly inferior to the ones involving the Lab one.

A possible explanation of our inconclusiveness can be found in the fact that WGANs require a long and computationally demanding training (at least 200 epochs over a set of at least 10.000 images): the main idea behind them is to train the discriminator/critic to the point where its loss stabilizes on a restricted range of low values, forcing the generator to overcome a "strong opponent".

With these drawbacks in mind, adopting WGAN for image-to-image translation may not be a promising avenue (see also [2]). We want to stress, however, that this should not be considered conclusive evidence, as the demanding training times and limited access to computing resources limited our ability to perform a deeper study of this approach.

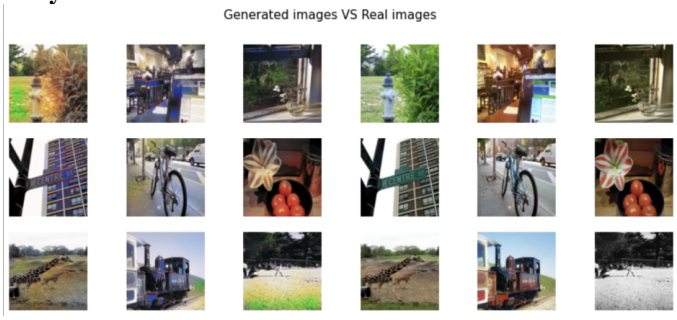
That being said, a very clear outcome is given by our results: the color space plays a pivotal role in the colorization task. As previously stated the Lab color space offers numerous advantages over the RGB one: the most important being the reduced number of channels that the generator has to learn.

VII. Concluding Remarks

In this paper, we have presented a novel approach for colorizing grayscale images using a conditional Generative Adversarial Network (cGAN) architecture, consisting of a UNet generator and a PatchGAN discriminator. Our approach uses Wasserstein Generative Adversarial Network training with Earth Mover's Distance to minimize the gap between generated and target images and also employs Structural Similarity Index Measure (SSIM) as a metric to evaluate the visual quality of the colorized images.

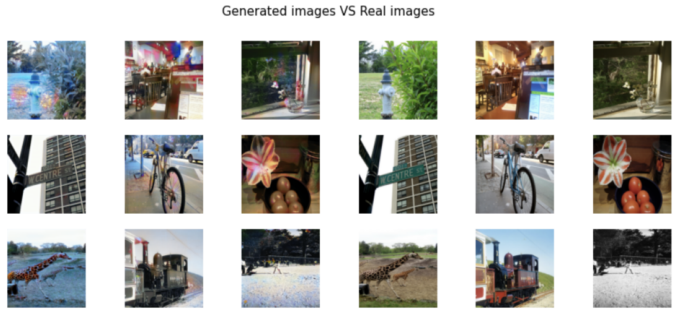
Even if the proposed approach does not represent a significant advancement in the field of grayscale image colorization, it provides some interesting suggestions to this challenging problem. Moreover, there is still room for further improvement and in the future, we would like to explore incorporating additional information such as semantic or contextual information to further improve the performance of our approach. Additionally, experimenting with other adversarial loss functions, such as the LS-GAN or the hinge loss, could also yield interesting results and would be a valuable direction for future research.

- **Grayscale to RGB:**



- **Lab to RGB:**

- **Grayscale to RGB WGAN:**



References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*.

[2] N. Makow, "Wasserstein gans for image-to-image translation,"

VIII. Contributions of each member

We did not subdivide the work, because we preferred to meet and work together. Indeed, we can say that each member of the group contributed equally to the entire project, because almost all the code has been written together and all the ideas have been discussed together as well. Hence, it is very difficult to assign each idea to a member of the group, but we can assert that each of us contributed with at least one idea per subproblem. However, a rough division based on where each of us developed more code and had more ideas can be the following:

- **Giacomo Gonella:** UNet generator (implementation; the cGAN training process was implemented with Pietro due to the necessity to alternate between generator and discriminator), LabRGB class;
- **Pietro Picardi:** PatchGAN discriminator (implementation; the cGAN training process was implemented with Giacomo due to the necessity to alternate between generator and discriminator), GrayRGB class;
- **Filippo Zanatta:** WGAN (implementation and training), results visualization and analysis.