GRAY-SCALE IMAGE COLORIZATION USING DEEP LEARNING

Parsa Kazerooni, Utku Mehmetoğlu
Computer Engineering Department
Yıldız Teknik Üniversitesi, 34220 Istanbul, Türkiye
{parsa.kazerooni, utku.mehmetoğlu}@std.yildiz.edu.tr

***Özetçe* —Projemiz, gri seviye resimleri renklendirmek için derin öğrenme tabanlı bir model oluşturmayı amaçlıyor. Buradaki ana amacımız bir insanı otomatik olarak, tatmin etmek olduğundan, göze hoş görülen renklendirmeler üreten bir yaklaşım izlendi. Sistem, test zamanında bir CNN'de ileri besleme geçişi olarak uygulanır ve bir milyondan fazla renkli görüntü üzerinde eğitildi. Performans, DeltaE gibi matematiksel yaklaşımlar kullanarak çıktıları test eden yaklaşımlar ve insan geri bildirimini inceleyen bir "insan Turing testi" anketi kullanarak değerlendirilir.**

Anahtar Kelimeler—Renklendirme, Derin Öğrenme, CNN, ResNet, Delta E

***Abstract*—Our project aims to build a deep-learning based model in order to colorize gray-scale images. Since our main goal is to satisfy a human viewer in an automatic way, we chose study methods that predicts colors that are appealing to the eye. The system is implemented as a Self-supervised feed-forward pass at test time in a pre-trained Convolutional Neural Network. Its performance is measured by both numeric analysis such as DeltaE and by a "Human Turing Test" survey that checks for human satisfaction rate.**

Keywords—Image Colorization, Deep Learning, CNN, ResNet, Delta E

I. Introduction

Gray scale images are images whose pixels are constructed only by the amount of brightness and nothing else. However human eye perceives colors in different aspects, brightness, color, intensity and etc. Thus such information on a gray-scale photo is lost. Our aim in this project is to recover these lost colors as much as possible using the power of deep-learning. A human can paint a gray-scale image based on their past experiences which is a lifetime of familiarity for colourful objects, landscapes and life forms. But given such case a person may not paint a gray-scale photo to the exact same colors which the photo originally consist of, but still the painted image in most cases will look pleasant to human eye and satisfies a viewer. In case of teaching a machine exposing it to a large image dataset, which we prepared before, and feeding it gray-scale channels as input and other channels as output to teach our model is our goal in this project. We aim to produce a pleasant to look, vivid and sharp images and to test the pleasantness to the human eye we are going to use a "Human Turing test" where we survey a group of people with mostly artificially colored and native images where we ask our test group whether the image is actually natural or colored by machine.

II. Preview

For years, artists have been actively colorizing historical footage, old pictures. To help the process, the usage of artificial intelligence was introduced. Automatic Image colorization is still an ongoing research area with multiple solutions and methods. Some of them are discussed below:

A. User-guided colorization

User-guided networks was introduced in 2004 in an article proposed by Levin et al. [1]. The method needed user's provided scribbles with the target colors, then it colorized by filling each region with the related scribble using least-square method. Then, Huang et al. [2] propose an adaptive edge-detection based colorization method to reduce the amount of artifacts created by poor region detection. Also texture features were implemented by Qu et al. [3] and Luan et al. [4] to reduce the amount of scribbles needed. [5]

B. Plain Networks

This was an interesting method at the time and it was used by artists and animators. But the results were heavily depended by users abilities. The utilization of CNNs for image processing in the last years, grew significantly. Deep Colorization Method proposed by Z. Cheng [5] was introducted by using image descriptors as an input for a deep neural network. Iizuka's method was introduced a fusion layer for combining global priors and local image features as an input [6]. But most of the most of the methods were generating desaturated results. Zhang el al. [7] found a solution to this issue by using classification methods. To increase the diversity of colors in the final output, class-rebalancing is used during training. [8]

1) Deep Colorization: In this study various images from image-sets are clustered into different image colorization (ie. landmark images, city images, view images, person images) using an adaptive image colorization technique. Later chrominance values for the data-set is used to train the model. The model basically provides a chrominance value for a given image based on clusters. The architecture used in this model consists of a neural network which has the equal size of neurons in both input and output neurons, further the neuron size in the hidden layer is set to half of that count. ReLU is utilized as activation function in this method. In feature design work, images are separated as 3 levels of features as low high and mid. Adaptive Image Clustering

technique is also applied. The reference images are clustered with k-means algorithm and peak signal to noise ratio is computed between the original image and colorized result. If the computed result fails to meet a certain threshold the image is removed from the image-set. Also semantic clustering is applied to clustering method which helps to differentiate between images that are globally similar and semantically different from each other. [5]

2) Colorful Colorization: The network stacks two or three convolutional layers, ReLU layer and Batch normalization layer together to form eight blocks. To have vibrant colors, the method rebalances the loss based on the pixel rarity in training time. By doing this, less colorful backgrounds such as sky, won't affect the colorization process. It's a multi-modal scheme that each pixel has a probability distribution for each color. [7] [9]

C. Various methods

It's worth mentioning some different approaches and specific use cases as well.

1) Domain-specific Networks: Such as Infrared Colorization by Limmer et al. [10] which colorizes RGB images to Near Infrared (NIR) images using a multi-branch CNN. Additionally Song et al. proposed Radar image colorization [11] which acts as a feature-extractor on raw single-polarized radar images. Since domain knowledge is applied, their performance is usually high, but the use case is specific and it can not be generalized.

2) Text-based methods: Text-based Networks apply methods which can be a part of user-guided techniques, But instead of colored scribbles, they are guided through text hints. Manjunatha et al. [12] deployed a model by using Colorful Colorization's classifier FCNN [7] to be trained with image captions. Text2Color was also introduced by Bahng et al. [13]. It's implemented by using two conditional GANs, to generate a pallete by words, and to use the pallete for the colorization. They can be advantageous in the situations were image descriptions are already available, but they require an accurate text input. [9]

III. Approach

The colors of an image can be represented different. The most common color spaces are RGB (Red, Green, Blue), CMYK (Cyan, Magenta, Yellow, Key). CIELAB is a color space which represents the colors by their L* as luminance, and a* channel as red- green pair, and b* as yellow-blue pair. CIELAB is a good representation to human vision, since our aim is to generate colors to seem natural to humans, CIE Lab color space is preferred. Also, by using CIE Lab, the luminance channel can be directly achieved by the gray-scale image. This means that our model only needs to predict ab channels. But using RGB color space, forces us to predict three different values.

A. Problem space

Given a image in CIE Lab colorspace with the size of $H \times W$, the input is its Lumminance channel demonstrated as $X_L \in \mathbb{R}^{W \times H \times 1}$. The target is to predict a and b channel values. The target is to obtain a accurate mapping between Luminance values and ab values for prediction. denoted as below

$$\mathscr{F} : \mathbf{X_L} \rightarrow \left(\hat{\mathbf{X}}_\mathbf{a}, \hat{\mathbf{X}}_\mathbf{b}\right) \tag{1}$$

The colorized image, $\hat{X} \in \mathbb{R}^{W \times H \times 3}$, is constructed by combining X_L with $\hat{X}_a$ and $\hat{X}_b$.

$$\hat{X} = (X_L, \hat{X}_a, \hat{X}_b) \tag{2}$$

B. Objective Function

MSE (Mean Squared Error) is used as the objective function. it's defined by taking the arithmetic mean for the squared distance of each pixel's individual channel values between the prediction and the ground truth.

$$L = \frac{1}{2HW} \sum_{k \in \{a,b\}} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(X_{k_{i,j}} - \hat{X}_{k_{i,j}}\right)^2 \tag{3}$$

We also considered to use CMSSIM as the objective function but its details is explained in the Section 4.1.1.

C. Network Architecture

ResNet-18: is a CNN trained on ImageNet dataset, presented in the article "Deep Residual Learning for Image Recognition" [14]
The Network Architecture is shown in Figure 1.

layer name	output size	18-layer
conv1	112×112	
conv2_x	56×56	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$
	1×1	
FLOPs		1.8×10^9

Figure 1 ResNet18 Architecture [15]

IV. Material and Dataset Prepration

We use MirFlicker[16] as the dataset to train and test our network. it contains pictures varying from green landscapes to human-made objects.
Preprocessing: Converted to Lab colorspace, extracted L

and ab channels into individual numpy arrays.
After the training is done, our network tries to predict the best combination of ab values. By adding the lightness channel (which is given as input), we would be able to construct the full picture. (Figure 2)

Figure 2 a picture with ab channels only, added lightness values to recreate the original picture

V. Implementation

A. Training The Model

The dataset is splitted into 20% dedicated as train set and 80% as test set. Each training epoch consists of 1250 loops and the Batch Size of 16. Total of 15 Epochs are iterated through the learning process.

B. Testing

As shown in figures 3, the model predictions are quite accurate or at least the fact that they're artificially colored is unnoticeable. But there are some human-made objects that are in a completely different color, some of those aren't usually linked with a specific color and no human would tell if it's in a "wrong" color, but some of them do have a globally-known color such as Turkish Flag. The Classification Method improves the colorization of these kind of objects.

VI. Performance Analysis

A. Color Distance Metrics

To measure how close the predicted image is, there are many approaches. e.g. Evaluating based on human feedback such as "Colorization Turing Test" used by Richard Zhang's Colorful Colorization method [7]. It determines the realness of the generated image by the percentage of the 'yes' answers of the question "Is it real?" in a survey. It can be a good metric since it's directly related to the target of this research. But it can be a slow process. Therefore We investigate and define autonomous methods.

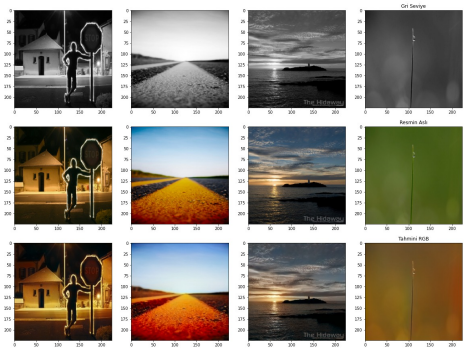

Figure 3 Gray Scale, Actual and Predicted image

1) CMSSIM and other Metrics: Multiscale Structural Similarity (MSSIM) is evaluated by Luminance, Contrast, Color, Gradient and other properties of two images to determine their similarity [17]. By calculating the Euclidean distance between two colors and applying a threshold filter, we can estimate how accurate the color prediction performed [18]. But it can be inconsistent if not performed correctly, more details are examined in [19].

2) Mean Delta E: Delta E (Empfindung) is used to measure the difference between two colors perceived by human. Table 1 describes the Delta E values. [20] It has the potential to be the perfect candidate metric to determine the similarity of a generated image with the ground truth.

Table 1 Delta E ranges and Perception

Delta E	Perception
≤ 1.0	Not perceptible by human eyes.
1 - 2	Perceptible through close observation.
2 - 10	Perceptible at a glance.
11 - 49	Colors are more similar than opposite
100	Colors are exact opposite

It is usually applied to "perceptually uniform" colorspaces such as CIELab. there are three versions of deltaE, dE76, dE94, and dE00. **dE76** is very similar to Euclidiean Distance Metric as shown below.

$$\Delta E^*_{ab} = \sqrt{(L^*_2 - L^*_1)^2 + (a^*_2 - a^*_1)^2 + (b^*_2 - b^*_1)^2} \qquad (4)$$

18 years later, **dE94** came out as an improvement. It introduced weighted parameters such as luminance, hue,

chroma. The formula is shown below

$$
\begin{aligned}
\Delta E^*_{94} &= \sqrt{\left(\frac{\Delta L^*}{k_L S_L}\right)^2 + \left(\frac{\Delta C^*_{ab}}{k_C S_C}\right)^2 + \left(\frac{\Delta H^*_{ab}}{k_H S_H}\right)^2} \\
\Delta L^* &= L^*_1 - L^*_2 \\
C^*_1 &= \sqrt{a^{*2}_1 + b^{*2}_1} \\
C^*_2 &= \sqrt{a^{*2}_2 + b^{*2}_2} \\
\Delta C^*_{ab} &= C^*_1 - C^*_2 \\
\Delta H^*_{ab} &= \sqrt{\Delta E^{*2}_{ab} - \Delta L^{*2} - \Delta C^{*2}_{ab}} = \sqrt{\Delta a^{*2} + \Delta b^{*2} - \Delta C^{*2}_{ab}} \\
\Delta a^* &= a^*_1 - a^*_2 \\
\Delta b^* &= b^*_1 - b^*_2 \\
S_L &= 1 \\
S_C &= 1 + K_1 C^*_1 \\
S_H &= 1 + K_2 C^*_1
\end{aligned}
\tag{5}
$$

Finally in 2000, **dE2000** or **00** was introduced, it solved some issues with eD94 but it's more complex. [21] [22]

We used dE94 because it's a good balance between accuracy and complexity. Since It's a per-pixel metric, there should be a generalization method for the whole image. By taking mean of pixel delta E values, we can obtain a single metric to discuss the color similarity of the prediction and ground truth.

Also, We can visualize the color differences to obtain more details on performance by plotting all the delta E values as a 2d colored map. Figure 4 demonstrates the idea.

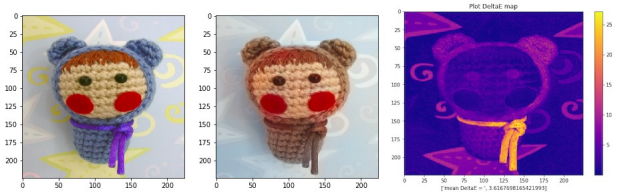

Figure 4 Ground Truth, Prediction, and Delta E Heat-map

Figure 5 gives us a good understanding of the output's color probability distribution. Since it takes the arithmetic mean, some images that are mostly a plain background and a small object, can be misrepresented as a accurate prediction since the impact of the small object is low. This can be fixed by considering weights when calculating overall delta E. But for now, it can be a good approximation for a color similarity measurement.

VII. Human Feedback Study

A simple survey was made with 9 randomly selected images consisting colorized outputs and natural images. the performance was measured by asking *Is the shown image real?* and the yes/no ratio of the answers. The images were not cherry-picked, but most of them contained sky and landscapes which could result to a biased sample, since the colorizing them is usually well performed.

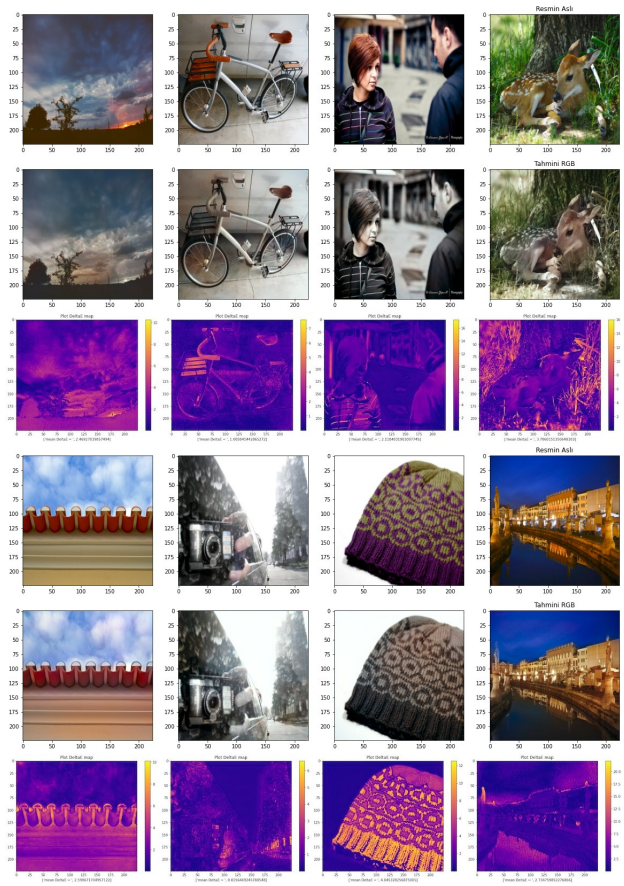

Figure 5 Delta E map for the tests

Figure 6 contains the results of the survey. Each image is evaluated by the percentage of the people who thought the image was real.

VIII. Conclusion

To conclude, we trained a model that is color accurate, good looking and sharp in most cases. Our further analysis showed us that the colorization technique might not be perfectly accurate but it can provide us a good representation of what the gray-scale image might looked like if it was a colorful one. Also in some cases our model struggled with unnatural or extremely vivid colors but nevertheless given the hardware limitations and the colors that disappear due to gray-scale conversion we believe that the model predicted most of the colors right and provided pleasant looking images.

References

[1] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," in *ACM SIGGRAPH 2004 Papers*, 2004, pp. 689–694.

[2] Y.-C. Huang, Y.-S. Tung, J.-C. Chen, S.-W. Wang, and J.-L. Wu, "An adaptive edge detection based colorization algorithm and its applications," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 351–354.

[3] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Transactions on Graphics (TOG)*, vol. 25, no. 3, pp. 1214–1220, 2006.

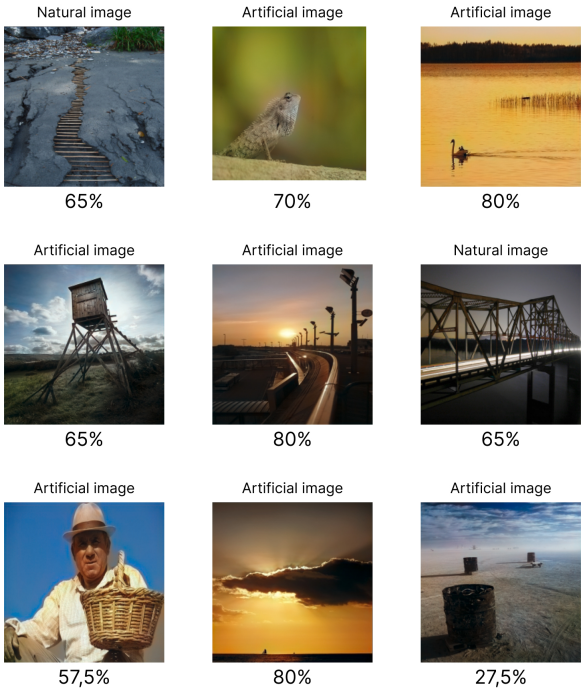

Figure 6 Results of the survey

[4] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Proceedings of the 18th Eurographics conference on Rendering Techniques*, 2007, pp. 309–320.

[5] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 415–423.

[6] S. Iizuka, E. Simo-Serra, and H. 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)*, vol. 35, no. 4, pp. 1–11, 2016.

[7] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666.

[8] L. R.-G. Federico Baldassarre, Diego Gonzalez-Morin, "Deep-koalarization: Image colorization using cnns and inception-resnet-v2," *ArXiv:1712.03400*, Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.03400

[9] S. Anwar, M. Tahir, C. Li, A. Mian, F. S. Khan, and A. W. Muzaffar, "Image colorization: A survey and dataset," *arXiv preprint arXiv:2008.10774*, 2020.

[10] M. Limmer and H. P. Lensch, "Infrared colorization using deep convolutional neural networks," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2016, pp. 61–68.

[11] Q. Song, F. Xu, and Y.-Q. Jin, "Radar image colorization: Converting single-polarization to fully polarimetric using deep neural networks," *IEEE Access*, vol. 6, pp. 1647–1661, 2017.

[12] V. Manjunatha, M. Iyyer, J. Boyd-Graber, and L. Davis, "Learning to color from language," *arXiv preprint arXiv:1804.06026*, 2018.

[13] H. Bahng, S. Yoo, W. Cho, D. K. Park, Z. Wu, X. Ma, and J. Choo, "Coloring with words: Guiding image colorization through text-based palette generation," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 431–447.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[15] Resnet pytorch. [Online]. Available: https://pytorch.org/hub/pytorch_vision_resnet/

[16] Liacs. Mirflickr dataset. [Online]. Available: https://press.liacs.nl/mirflickr/

[17] J. Nilsson and T. Akenine-Möller, "Understanding ssim," 2020. [Online]. Available: https://arxiv.org/abs/2006.13846

[18] M. Hassan Husain and C. Bhagvati, "Structural similarity measure for color images," *International Journal of Computer Applications*, vol. 43, pp. 7–12, 04 2012.

[19] A. Fatima, W. Hussain, and S. Rasool, "Grey is the new rgb: How good is gan-based image colorization for image compression?" *Multimedia Tools and Applications*, vol. 80, pp. 1–17, 01 2021.

[20] Zschuessler. Delta e 101. [Online]. Available: http://zschuessler.github.io/DeltaE/learn/

[21] M. Luo, G. Cui, and B. Rigg, "The development of the cie 2000 colour-difference formula: Ciede2000," *Color Research Application*, vol. 26, pp. 340 – 350, 10 2001.

[22] J. Nobbs, *A Lightness, Chroma and Hue Splitting Approach to CIEDE2000 Colour Differences*, 04 2002, vol. 5, pp. 46–53.