Journal of Artificial Intelligence and Data Mining (JAIDM), x(x): xxx-xxx, xxxx

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Automatic Grayscale Image Colorization using a Deep Hybrid Model

Kourosh Kiani*, Razieh Hemmatpour, and Razieh Rastgoo

Electrical and Computer Engineering Faculty, Semnan University, Semnan, Iran.

Article Info

Article History:
Received 07 August 2020
Revised 16 March 2021
Accepted 13 May 2021

DOI:10.22044/jadm.2021.9957.2131

Keywords:
Deep Learning, Convolutional Neural Network (CNN), Image Colorization, Encoder-decoder, Inception-v2, Computer Vision.

**Corresponding author: Kourosh.kiani@semnan.ac.ir (K. Kiani).*

Abstract

Image colorization is an interesting yet challenging task due to the descriptive nature of obtaining a natural-looking color image from any grayscale image. To have a fully automatic image colorization procedure, we propose a convolutional neural network (CNN)-based model to benefit from the impressive capabilities of CNN in the image processing tasks. Harnessing from the convolutional-based pre-trained models, we fuse three pre-trained models (VGG16, ResNet50, and Inception-v2) in order to improve the model performance. The average of three model outputs is used to obtain more rich features in the model. We use an encoder-decoder network to obtain a color image from a grayscale input image. To this end, the features obtained from the pre-trained models are fused with the encoder output to input into the decoder network. We perform a step-by-step analysis of different pre-trained models and fusion methodologies to include a more accurate combination of these models in the proposed model. Results on the LFW and ImageNet datasets confirm the effectiveness of our model compared to the state-of-the-art alternatives in the field.

1. Introduction

Image colorization consists of automatically adding color to a grayscale or black-and-white image without a direct human assistance. In other words, it contains the color assigning to each pixel in the grayscale image. This is easy in the human mind. For instance, the person simply understands that the color of the sky is blue. In this way, she/he restores the rest of her/his mind, and applies the color to the image. Thus the result is acceptable and desirable but when we have a grayscale image, we will take a brief look at it since our understanding of a color image is much larger than an image with a grayscale one [1]. It is also noteworthy that the human eye can distinguish millions of colors, while it can only detect about 20 to 30 levels of the grayscale spectrum [2].

The colorization of the grayscale image is a challenging research area because there are no plausible colors in some scenes and images without any prior knowledge. Furthermore, some objects do not have a homogenous color in the adjacent pixels. Also some objects can have different colors at different times without any change in their appearance (see Figure 1) [3].

Automatic image colorization includes adding colors to grayscale images without a direct human assistance. There are many interesting applications for image colorization such as an old movie or image colorization [4], distinguishing the tissues from each other in a medical image [5], and better analysis of satellite imagery and image compression [6]. In many problems, we are required to predict the pixel values in different parts of the input images and exploit information from these parts. In this way, automatic image colorization can help to facilitate the processing procedure of this mechanism. With the advent of Artificial Intelligence, especially deep learning in the recent years, many research areas benefit from the impressive advantages and capabilities of these techniques such as computer networks [7-11], computer vision [12-19], speech recognition [20], and natural language processing [21]. While some works have been proposed for image colorization using deep learning techniques [22-27], there are

still some challenges in this area that are required to be resolved. This is a very difficult task since it is an ill-posed problem that usually needs the user assistance in order to achieve a high quality. In this work, we are in line with the previous deep-based models, and benefit from the deep learning techniques for automatic image colorization. In order to propose a fully automatic approach capable of generating a real colorization of an input grayscale image, we propose a Convolutional Neural Network (CNN)-based model including three parallel CNN models to benefit from the fused features of these models. Our contributions can be listed as follow:

- A deep hybrid model including three parallel CNNs and an encoder-decoder network is proposed for automatic image colorization. Our model is simple yet efficient that benefits from the complementary features of three pretrained models.
- We analyzed different fusion methodologies and pre-trained models in order to obtain the most accurate one to use in the model.
- Benefiting from the deep learning capabilities, the proposed model outperforms the state-of-the-art alternatives in automatic image colorization.

The rest of this paper is organized as what follows. The related works are introduced in Section 2. The proposed model is described in detail in Section 3. The experimental results are provided in Section 4. Finally, we conclude the work in Section 5.

2. A Brief Introduction to CNN

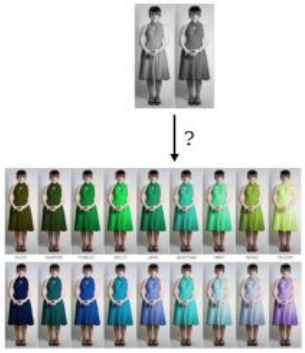

Figure 1. Possible different colors for an image.

A CNN is a Neural Network (NN) including one or more convolutional layers that has been introduced in 1995 by Yann LeCun and Yoshua Bengio [28]. A convolution is inherently sliding a filter over the input. It is a specialized kind of linear operation. Generally, CNN is a particular kind of NN for processing the data that has a grid-like topology. It is similar to ordinary NN, and includes some neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product, and follows with a non-linearity. The main intuition behind that is to look at the smaller portions of the image instead of looking at an entire image in one glance. Different layers such as convolution, Fully Connected (FC), activation, dropout, and pooling are used in a CNN. The CNN-based models successfully outperform the state-of-the-art alternatives in many practical applications such as image processing [29, 30] and classification [12-18].

3. Related Work

Current works in image colorization can be grouped into three categories: scribble-based models, example-based models, and learning-based models. While the first category employs the color interpolation techniques based on the color scribbles provided by a user, the example-based models transfer the color information from a ground truth image to a target grayscale image. The learning-based models use a learning procedure for the image variables corresponding to the image color. Among these categories, the learning-based methods have attracted high research interests in recent years with the advent of deep learning techniques. Deep learning obtained an outstanding superiority over other machine learning algorithms in various artificial intelligence domains [12-19].

The first category of image colorization models needs a user to manually add colored marks to a grayscale image and then smoothly propagate them across the whole image, based on an optimization method. A major weakness of this category is that it needs user intervention to provide the colored marks on the grayscale image. This is time-consuming and needs expertise that makes it hard to include these annotations in a large amount of data.

In the second category, the color information is transferred from a reference image to the grayscale image without any user intervention. Since the methods of this category need an accurate feature matching between the reference image and the corresponding grayscale image, satisfactory results cannot be achieved if feature matching is not performed precisely [31]. Image brightness and contrast are two crucial factors in this category.

In the third category, the color values are estimated using the learning process applied to the training images [22-26]. One of the learning-based methods is the neural network, where color images are firstly trained according to their color values. After that, the grayscale image is fed into the trained network to predict the color values. The methods of this category are effective yet computationally expensive [31]. For example, Baldassarre et al.

proposed a deep-based model by combining a shallow CNN with the InceptionResNet-v2 pre-trained model. This model is able to process images of any size and aspect ratio. Based on the results obtained, the proposed model achieves the "public acceptance" of the generated images using a user study [32]. A feed-forward and two-step CNN-based model [22], a deep-based model using VGG16 with the loss of cross-entropy [23], a CNN-based model for image colorization using the CIELUV color space [25], an end-to-end CNN-based model using a style transfer method to use the global features of the image [25], a deep-based model to predict the per-pixel color histograms of the input image [1], and a four-steps deep model [26] are some of the proposed methods in this category.

With the advent of deep learning in the recent years, CNN, as a deep-based model, has presented impressive capabilities in image processing. In this way, we propose a CNN-based model for the automatic image colorization. The details of the proposed model is presented in the next section.

4. Proposed Model

Given the fact that the pixel color is highly dependent on the features of its adjacent pixels, using the CNN is a suitable option for image colorization. In the case of having only a black-white or grayscale image, finding the exact color is complicated. There is not enough information for a network to estimate the pixel colors. For example, for a gray image of a car, there are several valid options for car color. To estimate a suitable color, we need more information to learn the model to match a grayscale input image to the corresponding color of the output image. One of the most successful learning-based models in recent years is the CNN model. CNN confirmed impressive capabilities in image processing. In this way, we propose a CNN-based model for automatic image colorization. Details of our model are explained in the following (See Figure 2).

4.1. Color Space

Choosing an appropriate color space has an important effect on the image colorization. For example, the RGB color space does not work well in all ways. It is better to transfer the image into another color space. Among the color spaces, CIE Lab and YUV are used in many tasks. The color space CIE Lab is a color-independent model, independent of any particular device such as a printer, scanner, and display screen. It defines a comprehensive model that encompasses a wide color space. The RGB color space is not suitable for our goal due to the color distribution in all the three layers or channels. That is why it is better to use another color space where the intensity of light and color is separated such as YUV and CIE Lab. Among the existing color spaces, CIE Lab is most similar to the human visual system due to the color space structure. In this color space, L contains the light intensity values. Changing the color space has two important advantages. First, reusing the L layer in order to produce the output of the final image since the brightness does not change in the image. Secondly, there is no need to predict three layers or color channels.

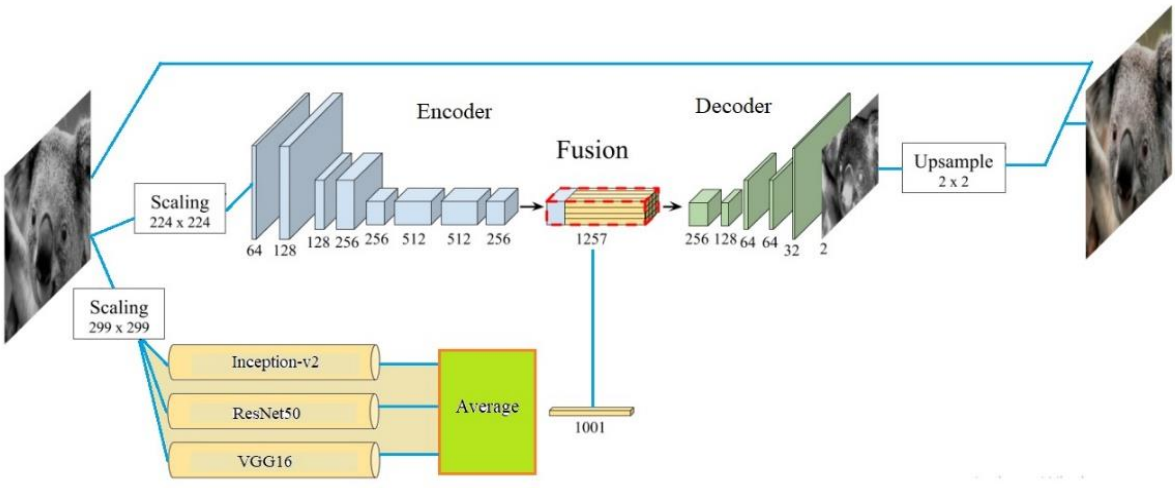

Figure 2. Proposed model including two pre-trained CNNs and an encoder-decoder network.

4.2. Training Phase

The input layer of the proposed model is the grayscale layer that receives the grayscale image in order to obtain the corresponding color image in

the output layer. In order to create the final color image, the L channel of the image is used again. The output of the convolution filters is also used to convert an input layer into two color layers. For this end, different filters are used and directed toward the output layer. In order to use the values in the error calculation, the target color values expected by the network should be normal. This is achieved using the Tanh activation function that holds values between -1 and 1. It performs the split operations into 128 values in the real values.

In the next step, we transfer the color space from RGB to the CIE Lab space. While the light intensity layer is considered as the input layer, the other two layers are the target output for the network. Then the light intensity layer is converted into three layers to enter simultaneously into the model including the encoder model. The encoder part of the proposed model includes three parallel pre-trained models. Then the output of the encoder model, achieved from the three CNNs, is averaged and fused to the decoder part of the proposed model. Then after predicting the output values and the operation of increasing the sample of the network, the operation of the error calculation is performed and repeated. After the network training, the predicted output must be converted to the image. Since the output values are between -1 and 1, we are required to return to the CIE Lab space. To this end, the output values are multiplied by 128 in order to obtain the correct values. Then the network input is the intensity layer of the image. The values a, b, and the final image are saved after converting space from Lab to RGB.

The proposed image colorization network consists of four main steps, as follow:

- An encoder block that extracts the low-level features using the convolutional filters.
- The CNN pre-trained models used to connect the high-level and low-level features using the fusion layer.
- A fusion block that concatenates the encoder outputs and the output feature maps of the three parallel CNNs.
- A decoder block to obtain the output image using the convolution filters and up-sampling method.

The outputs of the fusion block are fed into the decoder block in order to obtain the color image. The Adam function is used as an optimizer during the optimization process. It should be noted that if the input images are very similar, the results obtained are very satisfactory. However, the model does not have accurate results for the non-similar images.

For the error calculation, the actual values for a, b, and their predicted values should be normalized between -1 and 1. While a ReLU activation function is used in the middle layers, in the last layer, the Tanh activation function is used. In order to normalize the real values of a and b, they are divided into 128 to be in the range of -1 and 1. By doing so, we will be able to calculate the error value. After calculating the final error, the network updates the filters so that the total error is reduced. Here, we define the error function of the model as follows:

$$C(X,\theta)=\frac{1}{2HW}\sum_{k\in\{a,b\}}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(X_{k_{i,j}}-\tilde{X}_{k_{i,j}}\right)^2 \quad (1)$$

where θ refers to all the parameters of the model, X is the original image, the X i,j value is referred to the pixel value in the original image, the $\tilde{X}$ i,j value is referred to the predicated pixel value in the colored image, and H,W are the height and weight of the original image. This equation is simply generalizable to batch. This will be obtained by averaging the cost of all images in the category. For each batch, we will have the following function:

$$\frac{1}{|B|}\sum_{X\in B}C(X,\theta) \quad (2)$$

where B is a batch size. During training, this error is used in order to update the network using the Adam optimizer.

4.3. Test Phase

At this step, a quarter of all images were used in order to test the model. We performed the following steps to test the model:

- Pre-processing: In this step, we normalized the input images, and then transferred them to the CIE Lab space. In this space, the light, color, and intensity layers are separated. These layers are simultaneously fed to the pre-trained models.
- Prediction: In this step, the proposed model that is trained after 1000 epochs is used to estimate and predict the color values of the test images.
- Post-processing the model output: After the previous step, the generated output only contains the values between -1 and 1. These outputs are required to be scaled to the color range of the CIE Lab color space. This is achieved by multiplying the output values by 128.
- Output Image: In this step, the output values of 256 × 256 × 2 dimension are required to be added to the network input with the 256 × 256 × 1 dimension in order to obtain an image with the 256 × 256 × 3 dimension.

- Save the image: Finally, in order to save and view the output, the image created in the previous step is taken from the CIE Lab color space into the RGB space and then saved.

5. Results

In this section, we present the results achieved.

5.1. Datasets

One of the most important issues for any proposed model is to select an appropriate dataset for training. In most automatic image colorization methods, the ImageNet dataset [33] is utilized. One of the advantages of this dataset is the large number of images in different groups and classes (14.000.000 images). Furthermore, a free access to these images is the other benefit of this dataset. Also in order to test the effectiveness of the proposed system on the facial images, a collection of images of the people in the LFW dataset [34] is used. However, this dataset is used to identify the individuals. This dataset includes 13,000 human images with the upper body in various backgrounds. The advantage of this dataset is to provide a large number of face images with different backgrounds. A free access to it is another advantage of this dataset. Given that the images of this dataset have the 250 × 250 dimension, we are required to resize them to input to the models.

5.2. Implementation Details

We used the Adam optimizer with an initial learning rate of 0.001μ. The model is trained on separately two datasets including 17,000 images from both datasets, three-quarters of the images are used for training, and the rest are used for testing. The size of the batch is also 20. In order to train the modal, GPU GTX 1080 is used. The number of training steps is 1000.

5.3. Evaluation Metric

There are several ways to compare two images such as the Structural Similarity Index Measure (SSIM), Mean Square Error (MSE), Mean Average Error (MAE), and Peak Signal-to-Noise Ratio (PSNR). Meanwhile, in the colorization problem, the MSE and PSNR criteria are considered as the most commonly used methods. Evaluation of the image colorization problem is not easy due to its descriptive nature. For example, it cannot be precisely determined by taking into account the MSE criterion because two images may have a small MSE value but the coloration is not done properly. For this reason, some articles [24, 26] have used a poll of users in order to evaluate the final coloring effect. We used the MSE and PSNR metrics in our evaluation, as Table 1 shows. In this table, we report the results of four models, the details of which are as follows:

Model 1 [32]: This model is our baseline model. It includes an Inception-ResNet-v2 and an encoder-decoder model.

Model 2: This model includes Inception-ResNet-v2, VGG16, and an encoder-decoder model.

Model 3: This model includes Inception-ResNet-v2, VGG16, ResNet50, and an encoder-decoder model.

Model 4 [23]: This model includes the VGG16 model for image colorization.

We performed a step-by-step analysis on our model. As the results obtained show, our final model, Model 3, presents a better performance compared to the other models.

5.4. Comparison with Other Models

The results of the proposed model with the state-of-the-art models are shown in Figure 3, Figure 4, and Table 1. As one can see in these figures and table, the proposed method can colorize better than in some samples of the ImageNet dataset. In the LFW dataset, the proposed model is completely better. According to the values obtained from the criteria given in Table 1 and the observation of the results in Figures 3 and 4, the proposed system improves the colorization of images, especially the facial images. This comes from the fact that our model is simple yet efficient that benefits from the complementary features of three pretrained models as well as a well-suited fusion methodology. It should be noted that the empty cells in Table 1 are due to the unavailability of the results corresponding to that configuration.

Table 1. Results of comparison of the proposed model with the state-of-the-art models on the LFW and ImageNet datasets.

Model	PSNR		RMSE	
	LFW	ImageNet	LFW	ImageNet
Model 1 [32]	0.334	0.331	-	-
Model 4 [23]	-	-	-	0.299
Model 2 (ours)	0.342	0.334	0.311	0.296
Model 3 (ours)	**0.356**	**0.342**	**0.291**	**0.294**

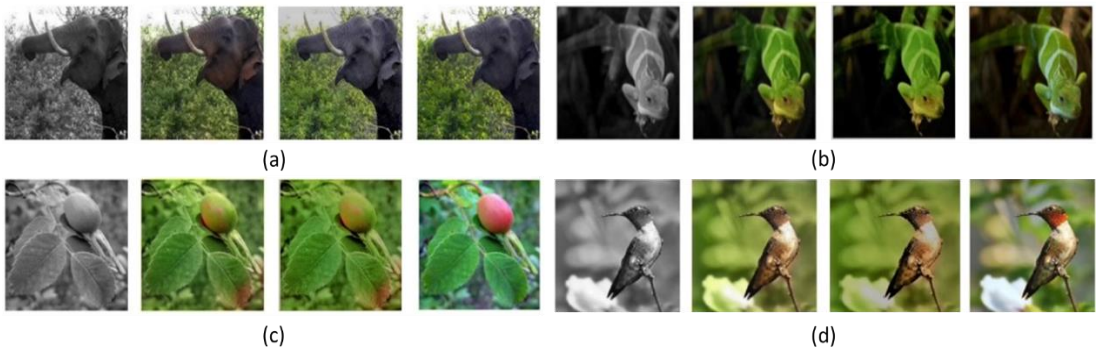

Figure 3. Comparison of the results of the proposed model (Model 3) with [23] on the ImageNet dataset. For each example, four images correspond to the input image, results from [23], Model 3, and ground truth, respectively.

Figure 4. Comparison of the results with state-of-the-art models: top-left: ImageNet, top-right, and top-bottom: LFW.

References

[1] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," *ECCV*, 2016.

[2] M.E. Valentinuzzi, "Understanding the human machine: a primer for bioengineering," *World Scientific*, Vol. 4, 2004.

[3] V.K. Bagaria and K. Tatwawadi, "CS231N Project: Coloring black and white world using Deep Neural Nets", *Stanford University*, 2016.

[4] X. Gu, M. He, and M. Gu, "Thermal image colorization using Markov decision processes," *Memetic Computing*, Vol. 9, pp. 15-22, 2017.

[5] I. Virag, L. -Tivadar, and M. Crişan-Vida, "Client-side Medical Image Colorization in a Collaborative Environment," *Studies in health technology and informatics*, pp. 904-908, 2017.

[6] T. Horiuchi, "Color image coding by colorization approach," *Journal on Image and Video Processing*, Vol. 1, pp. 158273, 2018. https://doi.org/10.1155/2008/158273.

[7] R. Rastgoo and V. Sattari-Naeini, "A neurofuzzy QoS-aware routing protocol for smart grids" *22nd Iranian Conference on Electrical Engineering (ICEE)*, pp. 1080-1084, 2014. DOI: 10.1109/IranianCEE.2014.6999696.

[8] F. Bordbar, R. Rastgoo, M.A. Askarzadeh, and M.S. Tavallali, "Prediction of Residential Natural Gas Consumption Using Artificial Neural Network," *The 9th International Chemical Engineering Congress & Exhibition (IChEC 2015)*, pp. 1-4, 2015.

[9] R. Rastgoo and V. Sattari-Naeini, "Tuning parameters of the QoS-aware routing protocol for smart grids using genetic algorithm," *Applied Artificial Intelligence*, Vol. 30, No. 1, pp. 52-67, 2016.

[10] R. Rastgoo and V. Sattari-Naeini, "Multi-Constraint Optimal Path Finding for QoS-Enabled Smart Grids: A Combination Approach of Neural Network and Fuzzy System," *Journal of Computing and Security*, Vol. 4, No. 2, pp. 47-61, 2017.

[11] R. Rastgoo and V. Sattari-Naeini, "Gsomcr: Multi-constraint genetic-optimized qos-aware routing protocol for smart grids," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, Vol. 42, No. 2, pp. 185-194, 2018.

[12] R. Rastgoo and K. Kiani, "Face recognition using fine-tuning of Deep Convolutional Neural Network and transfer learning," *Journal of Modeling in Engineering*, Vol. 17, No, 58, pp. 103-111, 2019.

[13] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Systems with Applications*, Vol. 150, No. 113336, 2020. DOI: https://doi.org/10.1016/j.eswa.2020.113336.

[14] R. Rastgoo, K. Kiani, and S. Escalera, "Multi-Modal Deep Hand Sign Language Recognition in Still Images using Restricted Boltzmann Machine," *Entropy*, Vol. 20, No. 809, 2018.

[15] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools and Applications*, Vol. 79, pp. 22965–22987, 2020. DOI: https://doi.org/10.1007/s11042-020-09048-5.

[16] R. Rastgoo, K. Kiani, and S. Escalera, "Hand pose aware multi-modal isolated sign language recognition," *Multi-media Tools and Applications*, Vol. 80, No. 1, pp. 127–163, 2021.

[17] R. Rastgoo, K. Kiani, and S. Escalera, "Real-time isolated hand sign language recognition using deep networks and SVD," *Journal of Ambient Intelligence and Humanized Computing*, 2021. https://doi.org/10.1007/s12652-021-02920-8.

[18] R. Rastgoo, K. Kiani, and S. Escalera, "Sign language recognition: A deep survey," *Expert Systems with Applications*, Vol. 164, 2021.

[19] M. Kurmanji and F. Ghaderi, "Hand Gesture Recognition from RGB-D Data using 2D and 3D Convolutional Neural Networks: a comparative study," *Journal of AI and Data Mining (JAIDM)*, Vol. 8, No. 2, pp. 177-188, 2020.

[20] M. Asadolahzade-Kermanshahi and M.M. Homayounpour, "Improving Phoneme Sequence Recognition using Phoneme Duration Information in DNN-HSMM," *Journal of AI and Data Mining (JAIDM)*, Vol. 7, No. 1, pp. 137-147, 2019.

[21] A. Torfi, R.A. Shirvani, Y. Keneshloo, N. Tavaf, and E.A. Fox, "Natural Language Processing Advancements by Deep Learning: A Survey," *arXiv: 2003.01200v2*, 2020.

[22] D. Varga and T. Szirányi, "Fully automatic image colorization based on Convolutional Neural Network," *23rd International Conference in Pattern Recognition (ICPR)*, 2016.

[23] J. An, K.G. Kpeyiton, and Q. Shi, "Grayscale images colorization with convolutional neural networks," *Soft Comput*. Vol. 24, pp. 4751–4758, 2020. https://doi.org/10.1007/s00500-020-04711-3.

[24] J. Wang and Y. Zhou, "Image Colorization with Deep Convolutional Neural Networks," *Stanford report*, 2016. http://cs231n.stanford.edu/reports/2016/pdfs/219_Report.pdf.

[25] 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, 2016.

[26] S. Titus and J. Rena, "Fast Colorization of Grayscale Images by Convolutional Neural Network,"

International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET), 2018.

[27] R. Zhang, P. Isola, and A.A. Efros, "Colorful image colorization," *ECCV*, 2016.

[28] Y. LeCun and Y. Bengio, "Convolutional Networks for Images, Speech, and Time-Series," *The handbook of brain theory and neural networks*, Vol. 3361, No. 10, pp. 1, 1995.

[29] N. Majidi, K. Kiani, and R. Rastgoo, "A Deep Model for Super-resolution Enhancement from a Single Image," *Journal of AI and Data Mining (JAIDM)*, Vol. 8, No. 4, pp. 451-460, 2020.

[30] L. Yatziv and G. Sapiro, "Fast Image and Video Colorization using Chrominance Blending," *IEEE Trans. Image Process*, Vol. 15, pp. 1120–1129, 2006.

[31] B. Li, Y.K. Lai, and P.L. Rosin, "Example-based Image Colorization via Automatic Feature Selection and Fusion," *Neurocomputing*, Vol. 266, pp. 687–698, 2017.

[32] F. Baldassarre, D.G. Morín, and L. Rodés-Guirao, "Deep Koalarization: Image Colorization using CNNs and Inception-ResNet-v2," *arXiv:1712.03400*, 2017.

[33] O. Russakovsky, *et al*. " ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, Vol. 115, pp. 211–252, 2015.

[34] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," *University of Massachusetts*, *Technical Report*, pp. 7-49, 2008.