CHROMA COMPONENT GENERATION OF GRAY IMAGES USING MULTI-SCALE CONVOLUTIONAL NEURAL NETWORK

Tien-Ying Kuo*, Yu-Jen Wei† and Bin-Yen You†
National Taipei University of Technology, Taipei, Taiwan, R.O.C
E-mail: * tykuo@ntut.edu.tw
† {yjwei, byyou}@image.ee.ntut.edu.tw

***Abstract*— In this paper, we solved the problem of the colorization algorithms using convolutional neural networks. There are existing algorithms predicting chroma components to generate colorization images, but the color of generated images is usually dim and the saturation of them is poor. Thus, we proposed to solve the colorization problem by a pyramid-alike multi-scale convolutional neural networks and convert the color space of image from RGB to HSV to predict the chroma components. Experiments indicate our algorithm can produce colorized images with more accurate color and higher saturation than the existing work.**

I. INTRODUCTION

There are lots of colorless gray photos from the old day. If those old photos can be colorized, they would be more vivid and joyful to the viewers. Even nowadays, there are still many cases of gray-scale images generated somewhere in our daily life, such as infrared images, satellite images, and electron microscope images. Coloring those images will help viewers analyze and interpret the information presented in the image more intuitively and correctly. To colorize a gray-scale image with rational color, it is necessary to first correctly recognize the contents of images with the help of prior knowledge. The manual colorizing approach by humans can be very accurate but requires huge amounts of labors and time, whereas the computer can save the efforts by predicting the corresponding color from analyzing the context of objects inside images, but it is a challenging problem.

The computer-based gray-scale image colorization can be categorized into three types of approaches: scribble-based[1], example-based [2], and learning-based [3-6] approaches. The scribble-based approach is a semi-automatic approach. It needs the assistance of humans to draw the color scribbles to describe each object feature of target images, and computers will colorize target images according to the scribbles. The example-based approach can automatically colorize images using the color of a similar feature in the pre-selected reference image by humans. Scribble-based and example-based methods can save a lot of time compared to all-manual coloring, but each processed image still needs to be assisted by providing relevant color information.

The learning-based approach can automatically analyze the features of gray-scale images and predict correct colors by neural networks. The convolutional neural network (CNN) updates the parameters through the learning from a large number of color images, so that it can predict reasonable colors that match the input. Most of the learning-based work in the literature [3-6] often generate the chroma components in YUV or Lab color space, which are then combined with the original gray-scale images to generate colorized results. Zhang et al. [3] proposed a learning-based approach regarding color prediction as a regression problem. They classified the ab-pairs of Lab color space into 313 categories and generated ab-pairs according to the features in images. As human eyes are less sensitive to chroma, Guadarrama et al. [4] proposed to generate delicate low-resolution chroma compositions pixel by pixel through PixelCNN architecture, and then improve image details using bilinear interpolation and the CNN. The PixelCNN used in [4] will increase the color pixel accuracy but it is an extremely time-consuming process. Iizuka et al. [5] used two parallel CNNs to respectively extract global and local features, and they added the scene classification label to optimize the training of network. Cao et al. [6] made the prediction based on conditional GAN. The generator adopts convolutional architecture without dimensionality reduction to replace U-net. All of the above learning-based colorization methods do not require human assistance, but the common problem is the resulted images are dim and low-saturated in color. The objective of our paper is to propose a fully automatic colorization algorithm based on learning-based methods while solving the above problems.

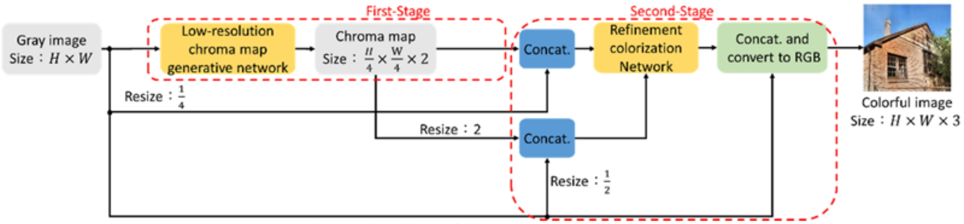

First-Stage
Second-Stage
Gray image
Size：$H \times W$
Low-resolution chroma map generative network
Chroma map
Size：$\frac{H}{4} \times \frac{W}{4} \times 2$
Concat.
Refinement colorization Network
Concat. and convert to RGB
Colorful image
Size：$H \times W \times 3$
Resize：$\frac{1}{4}$
Resize：2
Concat.
Resize：$\frac{1}{2}$

Fig. 1 Flowchart of the proposed architecture

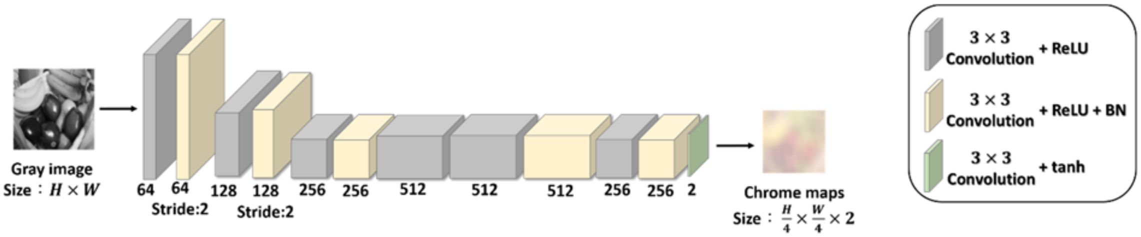


Fig. 2 First stage model

Our method has the following contributions:

- The image pyramid alike structure of convolutional neural network models is used to predict the chroma components H and S of HSV color space. This multi-scale structure can produce the chroma maps more accurately by finding from lower and middle resolutions.
- The pyramid alike structure can save network size as the feature size of networks is smaller in lower resolutions and the overall layer can be shallower due to the help of low-scale information.
- The HSV color space is adopted and the loss function for the H component of HSV is tailored to obtain a more correct color. Thus, the colorized image can have better color and saturation.

II. Proposed Method

We propose a pyramid model by incrementally predicting color from lower to higher resolutions of chroma, through which the color components H and S of HSV are generated. The flowchart is shown in Fig. 1.

We adopt HSV color space in our method due to its characteristics that can preserve linearity and match the human vision description properly, which we found in this space the learning-based method behaving more like humans can learn better color and saturation than other color spaces such as Lab and YUV that are commonly used in other work of this research area.

We use the V component of HSV as the gray-scale image input into our model and predict the approximate chroma components in the lower resolution forms via convolutional layers. Then, we feed the chroma output image as well as the original gray-scale image into our rest model to refine the correct resolution of chroma and finally produce the colorized image.

A. Modelling

Our model starts to predict the low-scale chroma maps with 1/16 size of original images because humans usually start to determine a scene with blobs of color in a global sense, and the low-scale chroma prediction will be easier to predict correctly. Our first-stage model is simply composed of 12-layer 3×3 convolutions and generates the low-scale chroma output. The design of predicting low-scale chroma can reduce the complexity of the whole model. The first 11 layers all take the ReLU as activation functions whereas the 12th layer uses tanh as activation function, as shown in Fig. 2.

Once the low-scale chromas are determined, we introduce the concept of image pyramids to our second-stage model designing, which enables the efficiency improvement of models by extracting and analyzing the features of multiple scales, as shown in Fig. 3. First, we concatenate the first-stage output with a low-scale input image, i.e., 1/16 downsized original grayscale images. We also up-scale the concatenated first-stage output to a middle-scale image, i.e., 1/4 downsized image. Unlike the standard pyramid generating the target layer only from contiguous layers, we generate the full size of chroma images using all of the lower scales of images directly. The concatenated images of low- and middle-scale images go through their own CNN networks to extract features in each of two scales through two convolutional sub-networks. Both outputs of the low and middle scale features are concatenated again by de-convolving the smaller size of branches from size 1/16 output to 1/4 size. After that, the concatenated feature will pass through a convolutional network to be deconvolved into full-scale chromas map of H and S, which can be then used to colorize the gray-scale input image.

B. Training

Places365 dataset [7] is used as our training set, which contains 365 categories of photos including natural scenery, humans, buildings, and so on. The dataset is divided into a training set, validation set, and test set. There are 1.8 million images in the training set, 18 thousand in the validation set, and 320 thousand images in the test set. The resolution of each image is 256×256. The training and validation sets are used to train the first stage and the second stage models. To input an image in the form of HSV color space, we convert the color space of training images from RGB to HSV, and normalize the values of three channels in the range between -1 to 1. V is used and set as the input, and HS is set as ground truth.

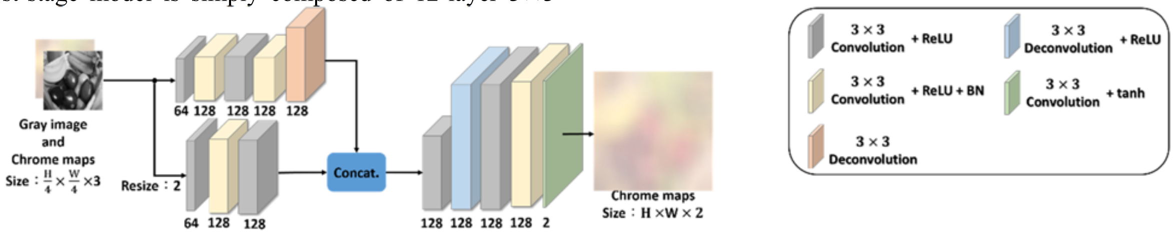
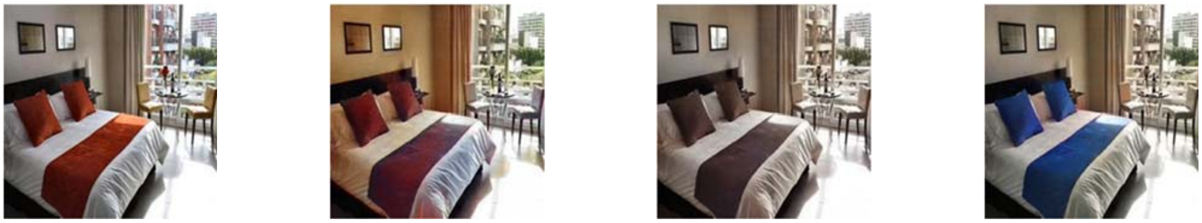


Fig. 3 Second stage model



Ground truth	Zhang et al. [3]	Iizuka et al. [5]	Proposed
PSNR	21.66	24.15	19.653
SSIM	0.9076	0.9332	0.8879

Fig. 4 Colorization results

Because we find the low-scale 1/16 sized of chroma first and use it as a base to find the higher resolution, the training of the second-stage model will be affected by the result of first-stage. We first train the first-stage model until it converges, and then train the second-stage model. To train the first-stage model, the size of the first-stage result is in 1/16 size of the original image, so the ground truth needs to be shrunk into 1/16 size by bilinear interpolation. After training is almost converged, we also fine-tune the first stage of models during the second stage of model training. Adam optimizer is adopted in our model training, which easily finds the best solution of models. The weight initialization of the model is very important and will affect the training accuracy. The He initialization [8] is adopted in our training because it has better effect in the model with the ReLU activation function.

In our training model, L1 loss is chosen as the loss function as L1 gives better results than L2 in our test, the formula shown in (1) and (2). The subscripts p and g indicate the predicted result and ground truth, respectively. S_{Loss} represents the absolute difference of saturation of predicted (S_p)and ground-truth (S_g)saturation. Since the H represents the [-180°,179°] of the color ring and is normalized to [-1 to 1], the color ring is dis-continuality at ±180° of H will cause the problem in measuring the color similarity. For example, a value of -0.9 of H in H_p will have almost the same hue to the ground truth (H_g) with value 0.9. However, the value of direct subtraction from $H_p$to H_g cannot reflect the difference. To solve this problem, we have to subtract a value of two from the absolute difference between H_p and H_g when the difference is bigger than one, as shown in (2). Thus loss function can correctly take into account the H similarity.

$$Loss = S_{Loss} + \lambda \times H_{Loss} \tag{1}$$

$$H_{Loss} = \begin{cases} |H_p - H_g| & , if\ |H_p - H_g| \le 1 \\ 2 - |H_p - H_g|, & if\ |H_p - H_g| > 1 \end{cases} \tag{2}$$

III. Experimental results

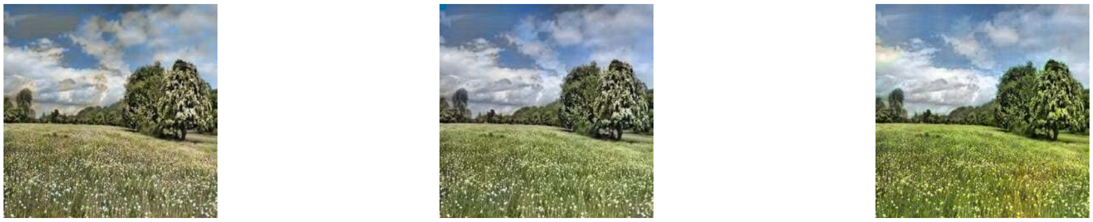
Table I shows the configuration of our test platform. Usually, to evaluate the output quality of image processing algorithms, the PSNR and SSIM are used objective image quality assessment methods, but in our case, these two metrics cannot effectively express the quality of colorized results because some objects have various choices of color appearance. Given an example in Fig. 4, our result is obviously more natural than those of other methods, but our PSNR and SSIM are the lowest. Therefore, we can only subjectively assess the coloring results.

Table I Training environment

CPU	Intel I7-4750K @ 4.00GHz
GPU	NVIDIA GTX 1080
RAM	12GB
Deep learning framework	Tensorflow

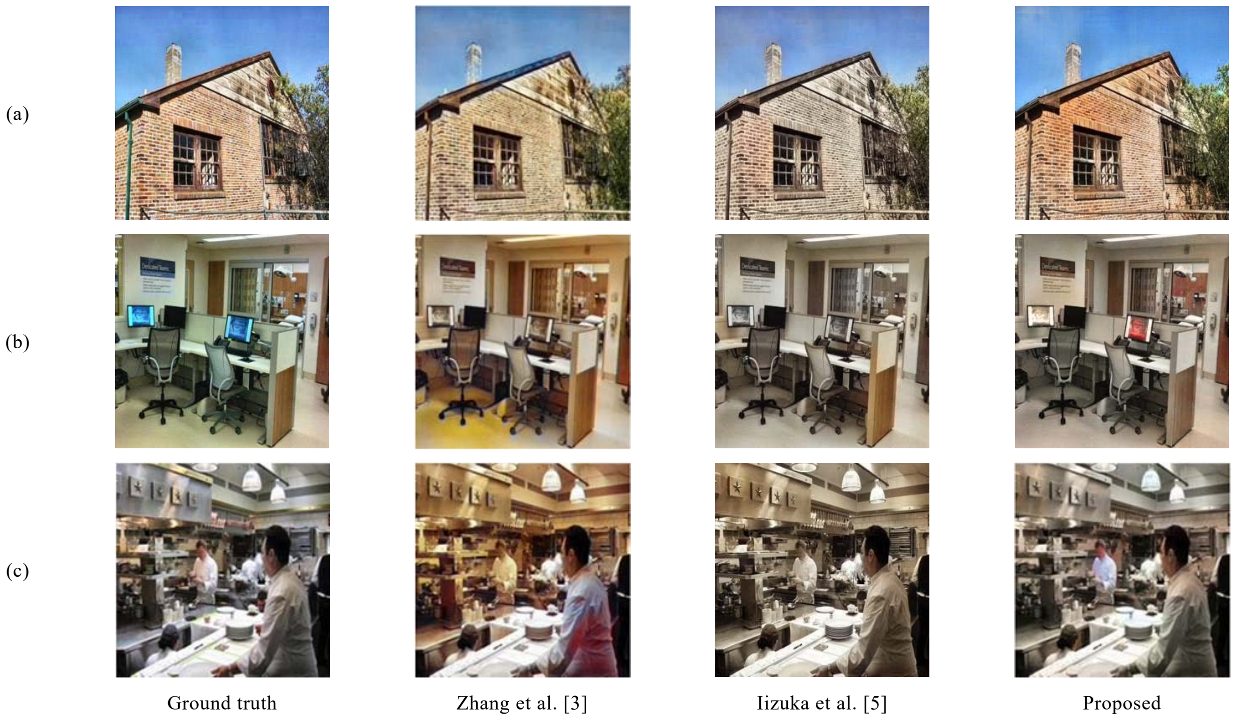
We would first verify the HSV is a proper color space chosen for our model of colorization. We utilize the test set of Places365 to compare the results of different color spaces under the same network models. To compare the different color spaces, we convert the training data into YUV, Lab, and HSV color spaces, and then use the same training method to train under the same network model. Fig. 5 shows an example of generated results. We can observe that the color saturation in HSV is significantly higher than those in other color spaces.

Next, we compare the results of our proposed methods with other popular cited existing works [3, 5], and our results look much more natural in the test images as shown in Fig. 6. The results of other work are a bit dim with low saturation. We can also observe the results of [3] also have color bleeding artifacts in different areas, such as the indoor scene, the chimney of outdoor and humans' clothes. Compared with [3, 5], our results look much closer to the ground truth.



(a) YUV (b) Lab (c) HSV

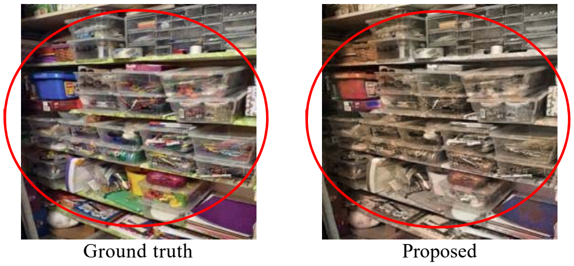
Fig. 5 Comparisons of different color spaces



Ground truth Zhang et al. [3] Iizuka et al. [5] Proposed

Fig. 6 Comparisons of colorization methods

Since we start predicting the chroma channel in the low-scale, it is sometimes difficult to predict the corresponding correct chroma value when the targets are too small in size. This could result in the failure in such areas, such as the objects circled in red in Fig. 7.



Ground truth Proposed

Fig. 7 Failure case

IV. Conclusions

We achieve a colorization algorithm using two-stage multi-scale convolutional neural networks to predict chroma components of HSV. Because the H of HSV is represented by a color ring, we design a tailored loss function for it. Our algorithm first generates the low-scale chroma components and then analyzes multi-scale information through an image pyramid alike architecture to yield colorized images. Compared with other works, our results are closer to the ground truth without the color bleeding artifacts.

Acknowledgment

This work was supported by Ministry of Science and Technology (grant # MOST 109-2221-E-027- 088-)

References

[1] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," ACM Transactions on Graphics (TOG), vol. 37, no. 6, pp. 1-14, 2018.

[2] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," ACM Transactions on Graphics (TOG), vol. 37, no. 4, p. 47, 2018.

[3] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in European Conference on Computer Vision, 2016, pp. 649-666: Springer.

[4] S. Guadarrama, R. Dahl, D. Bieber, M. Norouzi, J. Shlens, and K. Murphy, "Pixcolor: Pixel recursive colorization," arXiv preprint arXiv:1705.07208, 2017.

[5] 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, p. 110, 2016.

[6] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017, pp. 151-166: Springer.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452-1464, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026-1034.