SIP (2021), vol. 10, e15, page 1 of 12
doi:10.1017/ATSIP.2021.13

ORIGINAL PAPER

Two-stage pyramidal convolutional neural networks for image colorization

YU-JEN WEI,[1] TSU-TSAI WEI,[1] TIEN-YING KUO[1] AND PO-CHYI SU[2]

The development of colorization algorithms through deep learning has become the current research trend. These algorithms colorize grayscale images automatically and quickly, but the colors produced are usually subdued and have low saturation. This research addresses this issue of existing algorithms by presenting a two-stage convolutional neural network (CNN) structure with the first and second stages being a chroma map generation network and a refinement network, respectively. To begin, we convert the color space of an image from RGB to HSV to predict its low-resolution chroma components and therefore reduce the computational complexity. Following that, the first-stage output is zoomed in and its detail is enhanced with a pyramidal CNN, resulting in a colorized image. Experiments show that, while using fewer parameters, our methodology produces results with more realistic color and higher saturation than existing methods.

Keywords: Image colorization, Convolutional neural network, Image pyramid

Received 30 June 2021; Revised 13 September 2021

I. INTRODUCTION

Due to photographic equipment limitations, there exist many grayscale images in both the past and present, such as legacy photos, infrared images [1], thermal images [2], radar images [3], and electron microscope images [4]. These images can be made more vivid and appealing by colorizing them. A user can utilize image tools to manually colorize the objects in the grayscale image, such as blue skies, black asphalt, and green plants, based on empirically conjecture for the suitable color. However, the manual colorizing procedure is laborious and time-consuming, making the job considerably more difficult if the user is unfamiliar with these objects.

Colorization is recent active research yet a difficult subject in the realm of image processing, with the goal of quickly predicting and colorizing grayscale images by analyzing image content with a computer. Existing colorizing algorithms can be classified into three categories depending on the information provided by humans: scribble-based [5–10], example-based [11–15], and learning-based [1–4, 16–32] methods.

The scribble-based methods require a user to enter scribbles into a computer to instruct the colorization algorithm. The user draws correct color scribbles for the textures of various objects in grayscale images, and then the computer automatically propagates the colors of the scribbles to the pixels with the same texture. Levin's method [5] developed a quadratic cost function under the assumption that neighboring pixels having similar luminance should also have similar colors. They used this function with color scribbles to produce fully colorized images, but there exists a color bleeding problem in these results. Based on [5], Huang *et al.* [6] solved this problem by incorporating adaptive edge detection to enhance the boundaries of objects in images. Yatziv and Sapiro [7] used the distance between the pixel and its surrounding scribbles to determine the color of the pixel. [8–10] used convolutional neural network (CNN) to colorize grayscale images using color provided by color scribbles.

The example-based methods must be provided with prescreened reference color images that have similar features as the grayscale images to assist in colorizing the grayscale images. [11] converted the color space of pre-selected reference color images from RGB to Lab, and then analyzed and compared the luminance channel in its Lab color space with the grayscale image to be colorized, and provided the corresponding chroma information to the grayscale image based on similarities in textures. Gupta *et al.* [12] used super-pixel segmentation to reduce the computational complexity of [11]. [14] proposed a CNN-subnet technique with two CNN sub-nets: a similarity sub-net and a colorization sub-net. It uses the similarity sub-net to build bidirectional similarity maps between a grayscale image and a source image, then utilizes the colorization sub-net to

[1] Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan, R.O.C

[2] Department of Computer Science and Information Engineering, National Central University, Taoyuan City, Taiwan, R.O.C

Corresponding author:
Tien-Ying Kuo
Email: tykuo@ntut.edu.tw

generate a colorized image. [15] designed a spatially corresponding feature transfer module that uses self-attention to learn the relationship between the input sketch image and the reference image. Although scribble-based and example-based methods can save a large amount of time when compared to all-manual colorizing, each processed image must still be assisted by providing relevant color information. In addition, the colorizing performance is greatly influenced by the color scribbles and the selected reference image used.

In contrast to the above methods, the learning-based methods just require grayscale images to be fed into a CNN for automatic color image generation. The CNN requires a large number of images to train neural network parameters by analyzing the features of grayscale images to produce appropriate chroma components. The existing learning-based methods convert the color space of images from RGB to YUV or to Lab color space during the training of the CNN to learn the correlation between luminance and chrominance in the images. [17, 18] both used the pre-trained VGG [33] to extract features from grayscale images. Varga and Szirányi [17] utilized the features to predict chroma component for each pixel, while Larsson *et al.* [18] used them to predict the probability distributions of chroma component. [19–23] incorporated semantic labels into the training of colorization models, such as object registration [19–22] and scene classification [23]. [20–23] used two parallel CNNs to solve colorization task and semantic task, respectively, and the colorization task exploits the features extracted from the semantic task to improve colorization performance. [24] designed a set of networks, each of which aims to colorize a specific class of object. They classified the objects in images first and then colorized them with the matching network. Zhang *et al.* [25] proposed a learning-based approach regarding color prediction as a classification problem. [25, 26] classified the ab-pairs of the Lab color space into 313 categories and generated ab-pairs according to the features in images. As human eyes are less sensitive to chrominance, Guadarrama *et al.* [27] employed the PixelCNN [34] architecture to generate delicate low-resolution chroma components pixel by pixel, and then improved image details with a refinement network. The PixelCNN used in [27] will increase the accuracy but it is an extremely time-consuming process.

[28–32] made the prediction based on generative adversarial networks (GAN) [35] and GAN variants. Cao *et al.* [30] replaced the U-net of the original generator with a convolutional architecture without dimensionality reduction. All of the aforementioned learning-based colorization methods do not require human assistance and produce faster colorizing results than the other two categories of approaches, but a common problem among them is the resulting images are subdued and have a low saturation in color. In addition, existing neural network-based algorithms often designed complex network architectures to achieve more delicate results, resulting in models with huge parameter counts that are challenging to apply to real-world applications.

The pyramid concept in image processing is to exploit multiple scales of features to make a more accurate prediction, which has recently been used in convolution neural networks [36, 37]. [36] combined features from different scales by nearest neighbor up-sampling the lower scale features and then fusing them with the higher scale features. These multi-scale features are extracted from different layers of the backbone convolution neural networks. The results are predicted individually for each scale to ensure that the features at each scale are meaningful. [37] implemented different sizes of down-sampling and 1 × 1 convolutions on the features obtained in the last layer of the backbone neural networks, and then up-sampled these results to original size before making final predictions based on the concatenation of these modified features and original features.

The objective of our paper is to overcome the aforementioned problem using a fully automatic colorization algorithm based on the learning method. To allow a model to accurately predict chroma information without human hints while also minimizing the difficulty of model training at the same time, we adopted a coarse-to-fine generation approach. Our method is composed of two stages: preliminary chroma map generation and chroma map refinement. We predicted preliminary chroma components by the low-resolution chroma map generation network (LR-CMGN), aiming to obtain coarse color information from grayscale images first, allowing the model to converge more quickly. Then, to enhance the quality of generated chroma map, we generated a high-resolution color image by the refinement colorization network (RCN), which is designed with a pyramid model to reduce the number of model parameters. It is worth noting that we adopted the HSV color space in our method since we observed that machine learning behaves more like humans in this color space and can learn better color and saturation than other color spaces commonly used in this research area, such as Lab and YUV. The contributions in this paper are listed as follows:

- We presented a pyramidal structure of CNNs that predict the images of chroma components H and S in the HSV color space.
- The pyramidal structure can reduce the computing load of the model by analyzing smaller sizes of features and generate more reasonable chroma maps by analyzing information at multiple scales.
- The new loss function is designed for the properties of the H component of the HSV color space, ensuring that the colorized image obtains superior color and saturation.

II. PROPOSED METHOD

In this section, we introduce our colorization method. We first explain why we adopt the HSV color space instead of the YUV and Lab color spaces. The LR-CMGN and RCN

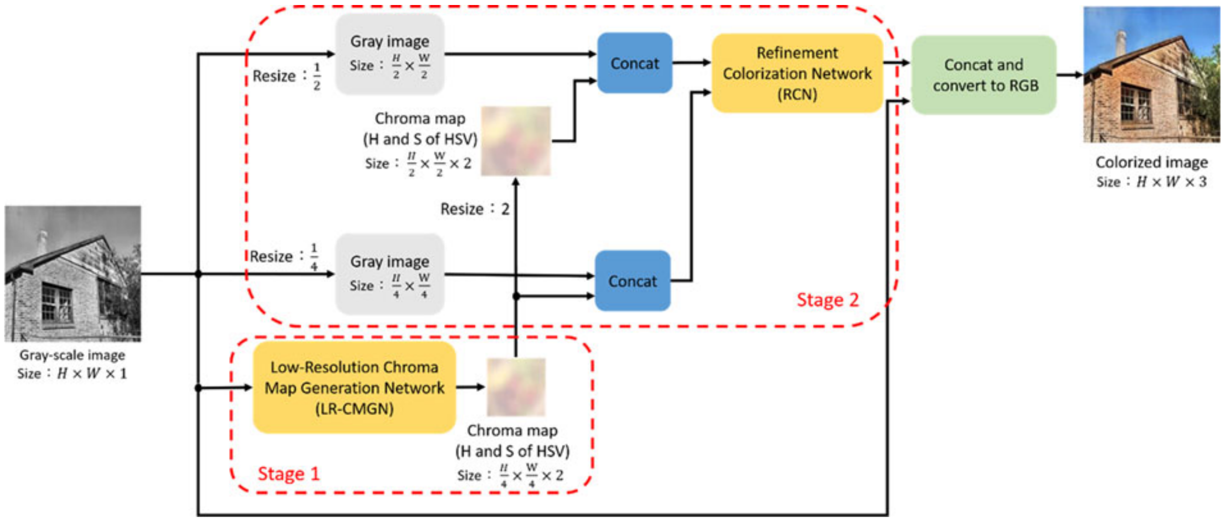


Fig. 1. Flowchart of proposed architecture.

architectural designs, as well as their training methodologies, are then detailed. Our colorization procedure, as illustrated in Fig. 1, involves two sub-networks: the LR-CMGN and the RCN, and is carried out in the HSV color space. The V component of HSV is used as a grayscale image for the LR-CMGN input and then generates low-resolution chroma maps. Afterwards, multiple scales of grayscale images and chroma maps are sent into the RCN, which outputs detailed chroma maps and, finally, the colorized image.

A) Analysis of color spaces

[25] explained the issue where the results of colorization using the Lab color space, despite being extensively utilized for modeling human perception such as [38], are tend to be low saturation. We believe that the definition of ab channels in Lab makes it hard to represent the saturation in a straightforward way because two parameters, a and b, are both involved with the saturation nonlinearly. As a consequence, during the training process, with existing loss functions, the difference in saturation between the prediction of model and ground truth is not easily minimized. The YUV color space is also a widely adopted color space in colorization tasks, but its chroma channels U and V have the same issue as the a and b parameters of the Lab color space.

We were inspired by this viewpoint and speculate that if neural networks can learn color saturation in a more informative way, this issue may resolve itself during the training process. We focused on the color spaces that enable saturation as an independent channel and investigated whether they can make the saturation of output images closer to the ground truth or not.

The HSV channel definition, where S stands for saturation, perfectly fits our needs. Furthermore, the HSV color space can match the human vision description more properly than Lab and YUV, allowing our model to learn in a more human-like manner and making the generated results more consistent with our perceptions. Taking into account all of the abovementioned factors, we selected HSV as our color space for training models to generate relevant chroma components.

B) Modeling

The way humans viewing images is to focus solely on either the whole or small parts of the image rather than on each pixel accuracy, making slight variations in pixel level, particularly in chroma components, difficult to detect. We conducted a basic experiment to confirm the practicality of this concept. We converted a set of randomly picked RGB images to the HSV color space and compared them to the images with low-resolution chroma maps. Figure 2(b) was created by downsizing the chroma maps of the original images in Fig. 2(a) to 1/4 size and then scaling up to the original size using bilinear interpolation. The results displayed in Fig. 2 show that, even if there might be some little artifacts in the image details when created from the low-resolution chroma maps, they can still reflect the majority of color information and the little artifacts will be removed later in our refining colorization network.

The above concept is incorporated in the design of the LR-CMGN to predict low-scale chroma maps with a size 1/16 of grayscale images. This design can lower the complexity of the whole model and make it easier to predict low-scale chroma maps correctly. This model consists of 12 layers of 3×3 convolutions, where all layers employ ReLU as the activation function except for the 12th layer, which uses tanh as the activation function, as shown in Fig. 3.

Although the low-resolution color map can represent the color of the whole image, there are still some differences from its ground truth, such as the stripes on the windows of the house and the car in Fig. 2. We proposed the RCN as a solution to this problem. The RCN architectural design is based on the image pyramid, and the performance of the model is enhanced by inputting and analyzing images of multiple scales, as shown in Fig. 4.

We already know from the experiment in Fig. 2 that the chroma components in HSV color space are fine to be

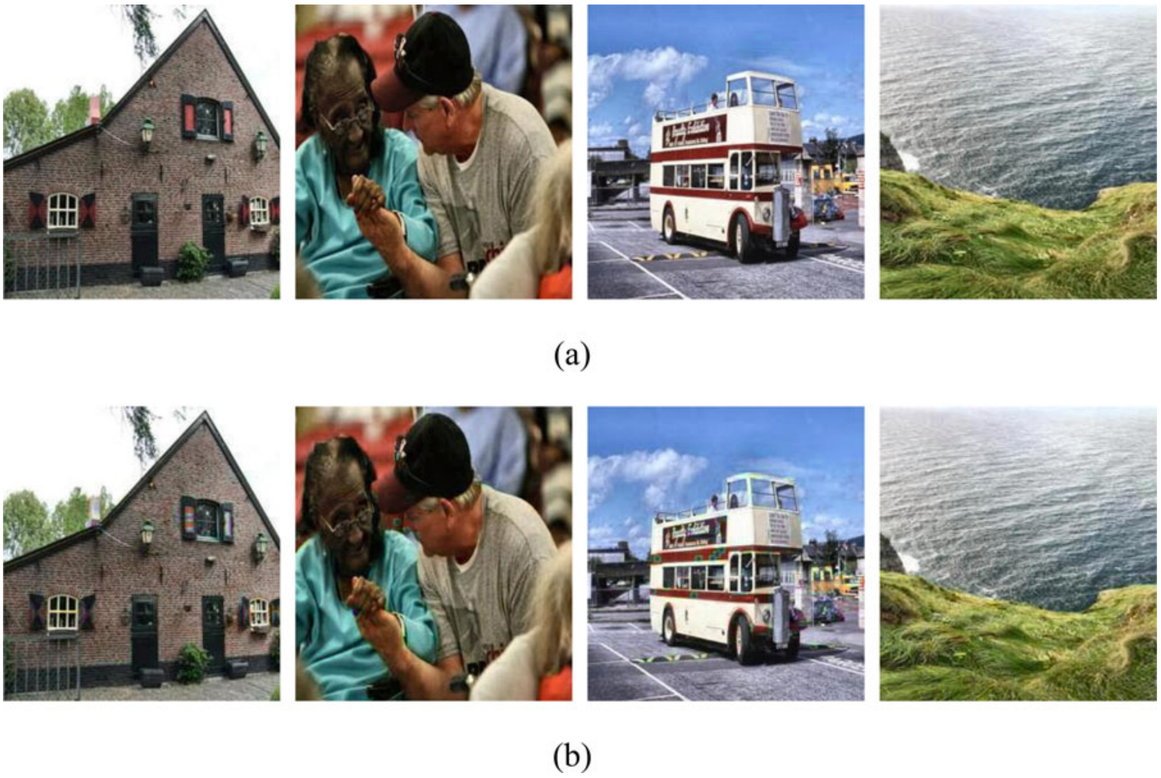


Fig. 2. Consequences of using low-resolution chroma maps. (a) Original images. (b) Images created using low-resolution chroma maps.

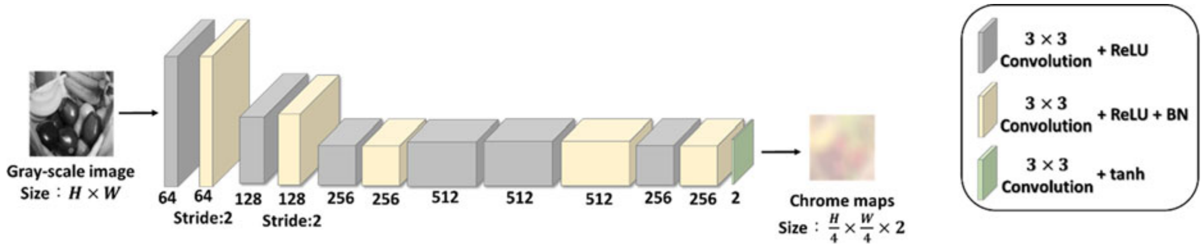


Fig. 3. Low-resolution chroma map generation network (LR-CMGN).

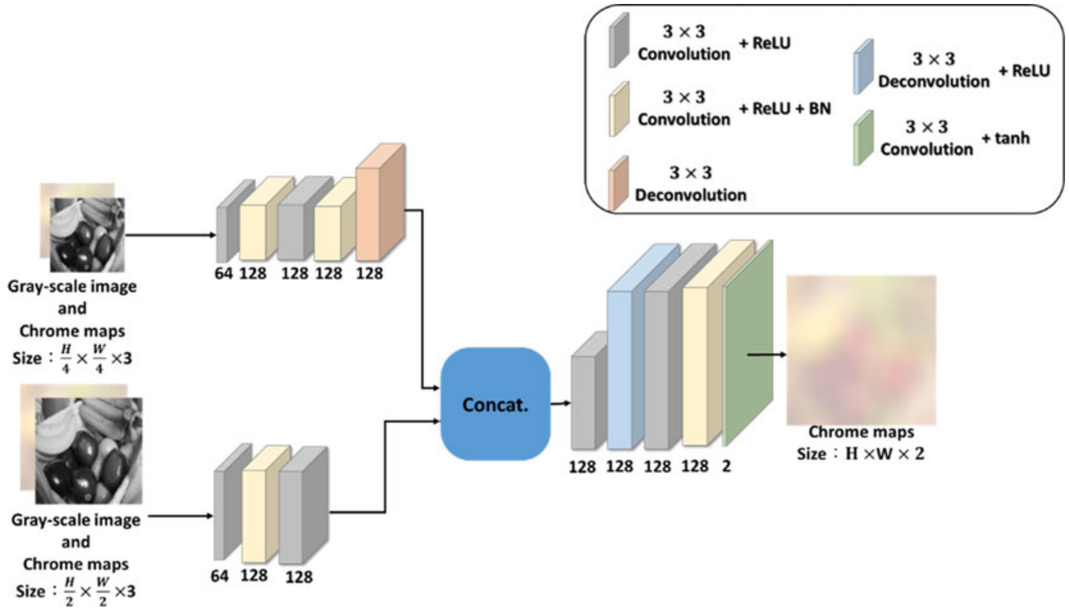
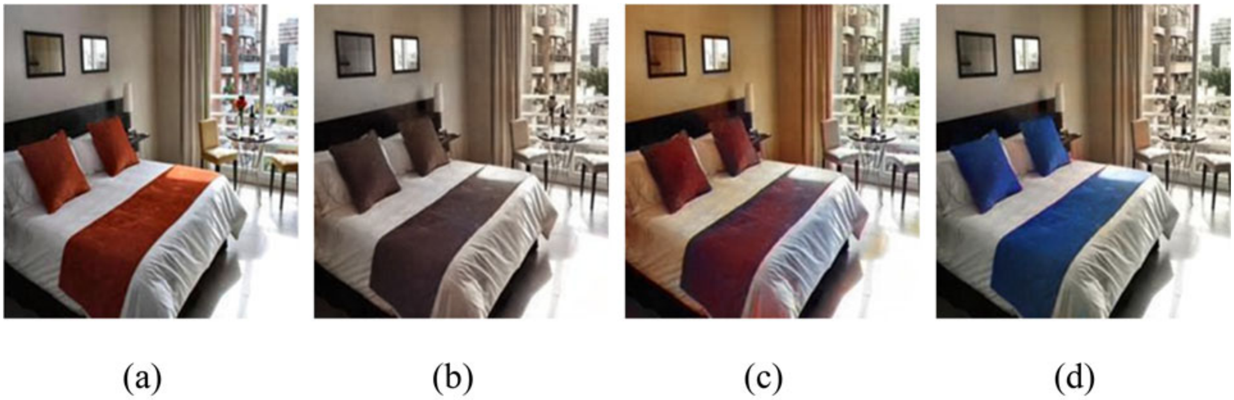

3 × 3 Convolution + ReLU
3 × 3 Convolution + ReLU + BN
3 × 3 Deconvolution
3 × 3 Deconvolution + ReLU
3 × 3 Convolution + tanh
Gray-scale image and Chrome maps
Size：$\frac{H}{4} \times \frac{W}{4} \times 3$
64 128 128 128 128
Gray-scale image and Chrome maps
Size：$\frac{H}{2} \times \frac{W}{2} \times 3$
64 128 128
Concat.
128 128 128 128 2
Chrome maps
Size：H ×W × 2

Fig. 4. Refinement colorization network (RCN).



	(a)	(b)	(c)	(d)
PSNR		24.15	21.66	19.653
SSIM		0.9332	0.9076	0.8879

Fig. 5. Problems in using objective image quality assessment metrics to evaluate colorization results. (a) Ground truth. (b) Iizuka *et al.* [23]. (c) Zhang *et al.* [25]. (d) Proposed method.

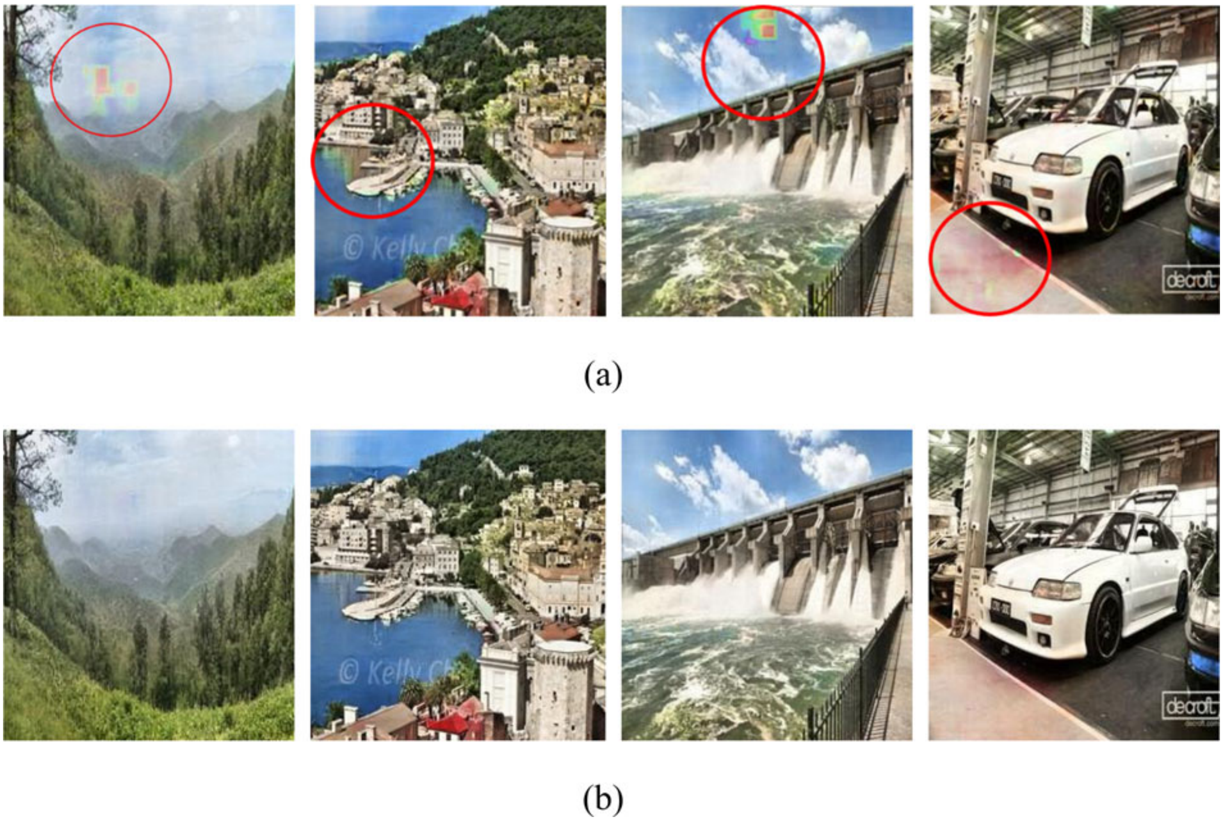


Fig. 6. Images produced by colorization methods with different stages. (a) Only first-stage. (b) Proposed two-stage method.

down-scaled due to the deficiency for humans to notice little changes in chroma channels at pixel level. Based on this assumption, our design differs from the existing pyramid structures in that they used more than two scales of information, whereas our method generates original-scale chroma maps from only low- and middle-scale information, allowing us to reduce the computational load of the model while still obtaining sufficient feature quality. Our implementation is more similar to [37] than other works that used the pyramidal concept, in that we modified the last layer output of a network to obtain multi-scale features rather than collecting different scale features from the feed-forward process.

As shown in Fig. 4, our pyramidal RCN has two inputs with different scales, which are created by concatenating the output of the LR-CMGN with two different sizes of low-scaled grayscale images. We scale down the LR-CMGN output and the grayscale image using bilinear interpolation

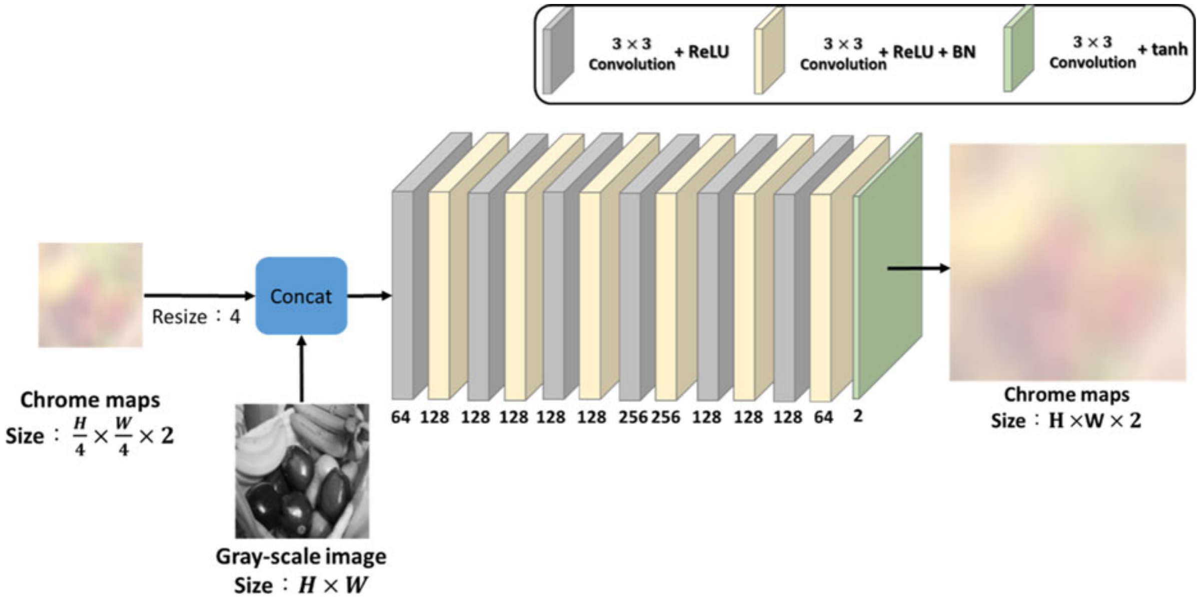


Fig. 7. Original design of refinement network.

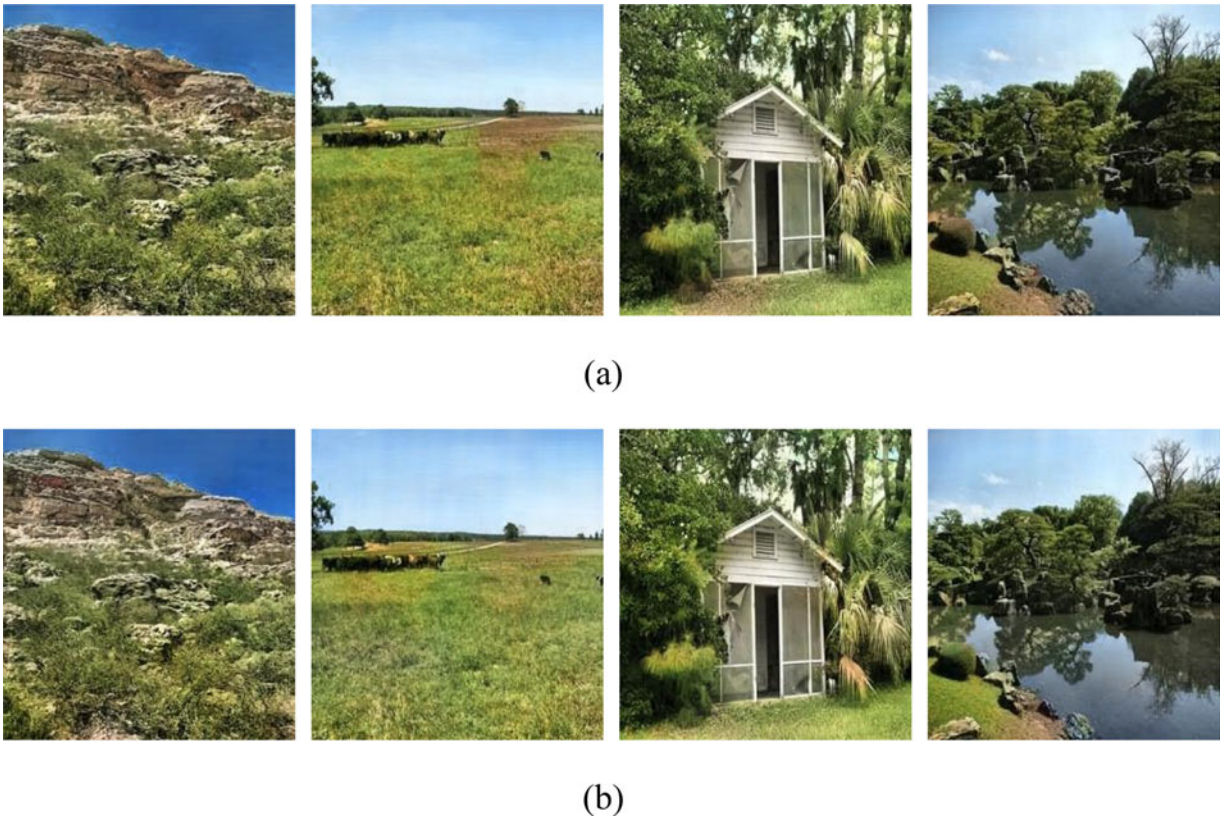


Fig. 8. Results of using different refinement network structures. (a) Original design (13 convolutional layers). (b) Pyramidal structure.

to 1/4 and 1/16 of the original grayscale image size. The low-scale and middle-scale features are extracted from their respective input images using two parallel convolutional sub-networks, and the output features representing two different scales are concatenated and fed into a deconvolutional layer to generate features of the same size as the original grayscale image. The *H* and *S* components of the HSV color space are predicted by analyzing these features, and the predicted results are combined with the grayscale image to form the colorized image.

C) Training details

Our colorization models were trained using the Places365-standard [39] database, which contains 1.8 million training, 36 500 validation, and 320 000 test images covering 365

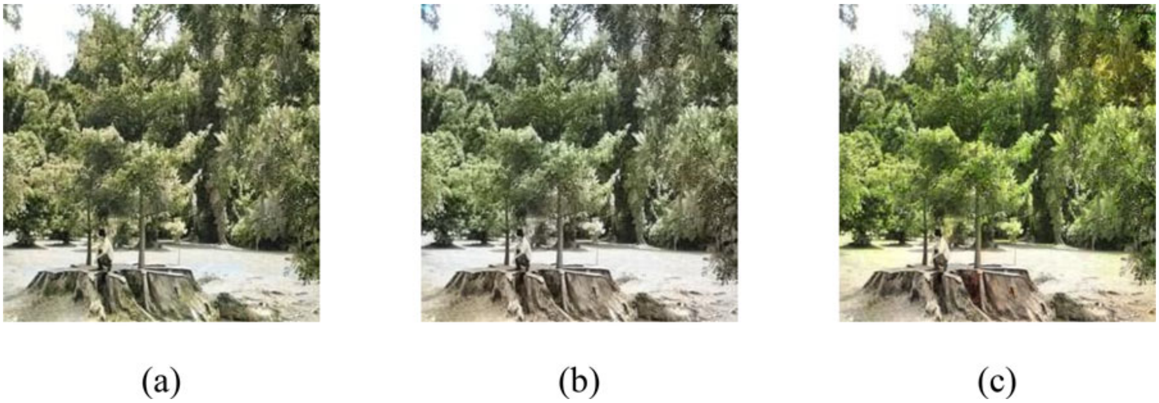


Fig. 9. Comparisons of colorization results using different color spaces. (a) YUV. (b) Lab. (c) HSV.

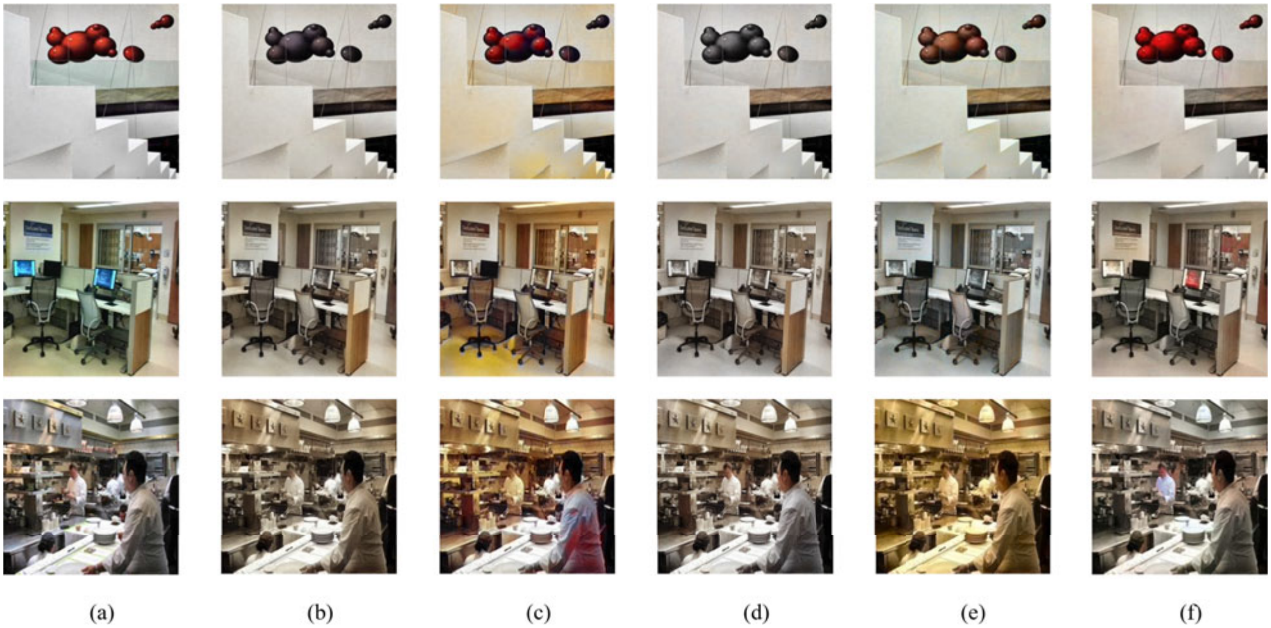


Fig. 10. Comparisons of different colorization methods used on indoor images. (a) Ground truth. (b) [23]. (c) [25]. (d) [8]. (e) [32]. (f) Proposed method.

scene classes such as indoor scenery and urban scenery, with each scene containing a variety of objects such as humans, animals, and buildings that can help our models colorize diverse grayscale image contents. The training and validation sets were used to train the LR-CMGN and the RCN, and all images in the dataset have a resolution of 256×256 pixels.

We excluded grayscale images from the dataset to avoid the machine from learning incorrect information. To predict the H and S components of the HSV color space, the color space of the training images is converted from RGB to HSV, and the value ranges of all channels are normalized between -1 and 1. During training, the V and HS components are employed as input and ground truth, respectively.

Given the fact that the RCN input contains both the grayscale image and the LP-CMGN input, the output of LP-CMGN will affect the RCN prediction result. As a result, when training the colorization model, we first trained LP-CMGN to near-convergence, then trained the RCN while fine-tuning the LP-CMGN parameters at the same time. Since the LR-CMGN prediction is conducted in low-scale, which is 1/16 of original chroma maps, we utilized low-scale ground truth created by shrinking the ground truth using bilinear interpolation to train this network.

In our experiment, both subnetworks use the same loss function, as shown in (1), where S_{Loss} and H_{Loss} represent the errors of S and H components predicted by the model, respectively, and these errors are calculated using the MAE (mean absolute error), and the weighting λ in the losses S_{Loss} and H_{Loss} is set to 1.

$$\text{Loss} = S_{\text{Loss}} + \lambda \times H_{\text{Loss}}, \qquad (1)$$

H is a color ring that represents color hue via angular dimension in the range $[0°, 359°]$ and is normalized to $[-1,1]$. Since the color hue is specified in a circular ring and has dis-continuality at ± 1, there is a risk that the model will learn the incorrect color hue when using MAE to calculate the error. For instance, if the model predicts H to be -0.9, the loss should be minor if the ground truth H is 0.9, but the error computed by MAE between these two values is high

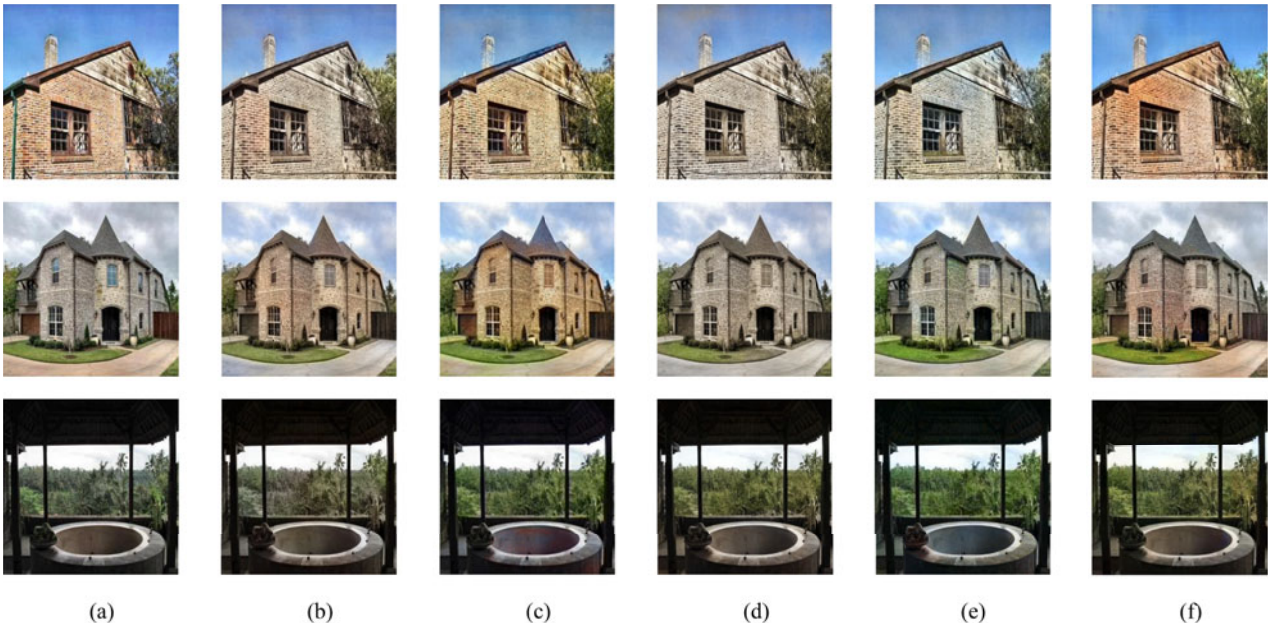


Fig. 11. Comparisons of different colorization methods used on outdoor images. (a) Ground truth. (b) [23]. (c) [25]. (d) [8]. (e) [32]. (f) Proposed method.

and does not accurately represent the actual difference. To address this issue, we modified the MAE as (2), where p and g denote the outcome of model prediction and ground truth, respectively. As the fact that the difference between two angles of H components should never surpass 180°, or 1 after normalization. Thus, we design the loss function in (2) based on this principle. If the difference is more than 1, the output of the loss function must be adjusted; otherwise, the output remains intact. By applying our design loss function (2) to the above example, we can calculate the H loss accurately by changing the error from 1.8 to 0.2.

$$H_{\text{Loss}} = \begin{cases} |H_p - H_g|, & \text{if } |H_p - H_g| \leq 1 \\ 2 - |H_P - H_g|, & \text{if } |H_p - H_g| > 1 \end{cases}. \quad (2)$$

The Adam optimizer was used to train the model, which has the advantage of fast convergence, and the learning rate is set to 2×10^{-4} across 10 epochs. Since weight initialization has a substantial impact on the convergence speed and the performance of the model, the He initialization [40] was employed to initialize the weights because it provides better training results of the model with the ReLU activation function than other techniques.

III. EXPERIMENTAL RESULTS

Our experiments were carried out using a PC with an Intel I7-4750 K at 4.00 GHz processor and an Nvidia GTX 1080 graphics card. Previous research in the literature has utilized the PSNR and the SSIM as objective image quality assessment metrics to evaluate the image quality of colorization, but since some objects might have multiple colors, and as [41] mentioned, these methods are unable to represent the true human perception, as illustrated in Fig. 5. Our result is clearly more natural than the other methods in Fig. 5, but our performance as measured by the PSNR and the SSIM is the lowest. As a result, we only subjectively assess the colorizing results.

A) Comparisons of various model designs

To stress the necessity of our model architecture, we explain and compare the performance impact of our network components in this section. We initially investigate the differences in model outcomes with and without refinement network. We scale up the first-stage result by bilinear interpolation, and then concatenate the grayscale image to obtain the colorized image, as shown in Fig. 6(a).

Although these images are fairly close to natural images, they have blocky color distortion in some areas. In contrast, Fig. 6(b) shows the results of applying the refinement network restoration after the first-stage result. The restored results are more realistic than Fig. 6(a), and the refinement network is necessary as it overcomes the problem of blocky color artifacts.

Next, we discuss the structural design of the refinement network. Figure 7 depicts our initial refinement network architecture, which is not pyramidal and composed of 13 layers with 3 × 3 kernels, with no decreasing dimensionality of convolutional layers. The inputs are a grayscale image and scaled low-resolution chroma maps, and some results are shown in Fig. 8(a).

In comparison to our proposed pyramidal structure, which achieves very similar results, the parameters for the original design and pyramidal structure are 2.2 and 1.5 M, respectively. We can observe that the parameters of our pyramidal structure are only 0.68 times those of the original design, thus we use the pyramidal structure for our refinement network.

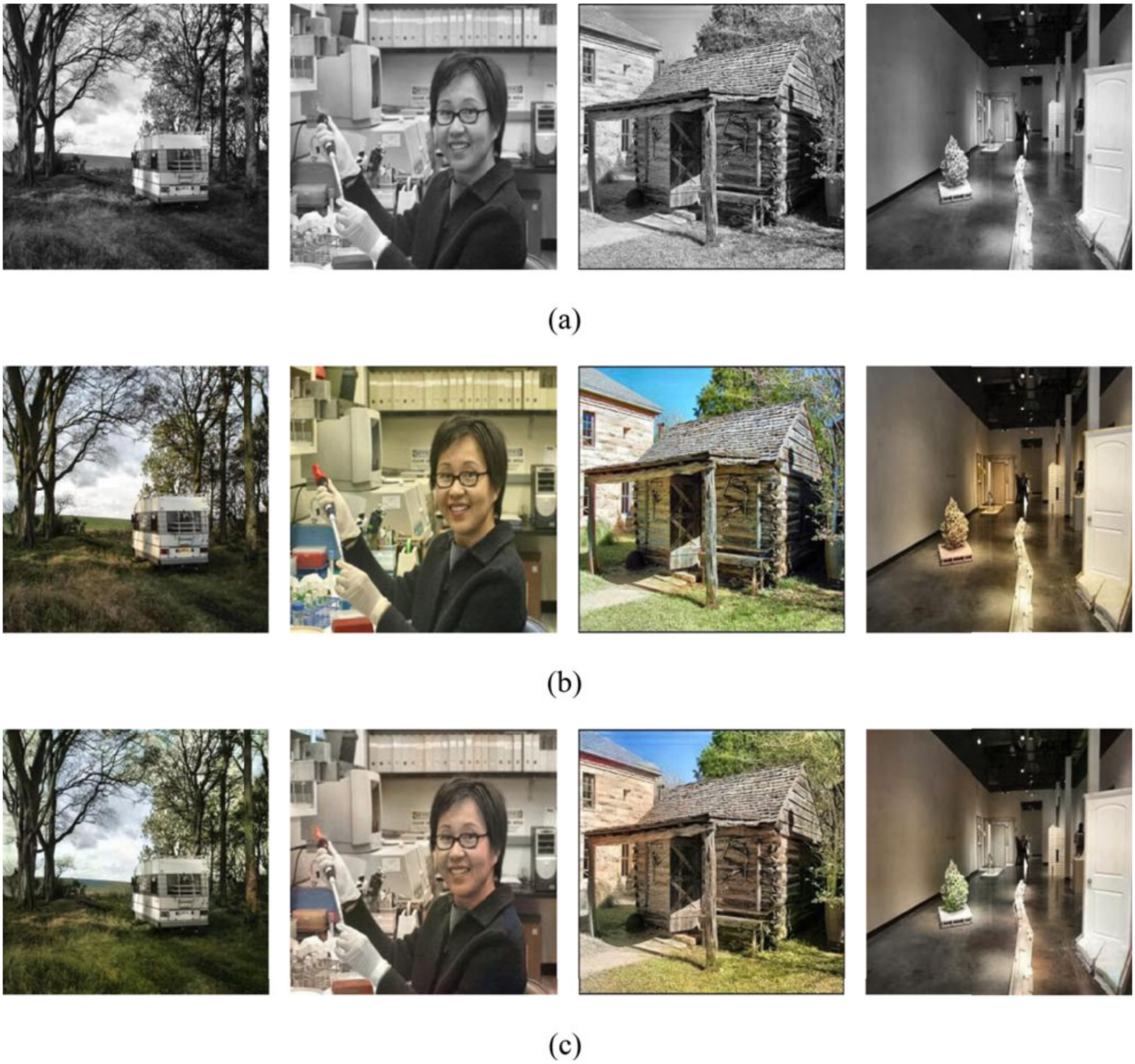


Fig. 12. More results by applying our method to grayscale images. (a) Grayscale image. (b) Ground truth. (c) Proposed method.

Table 1. Comparisons of the number of models parameters.

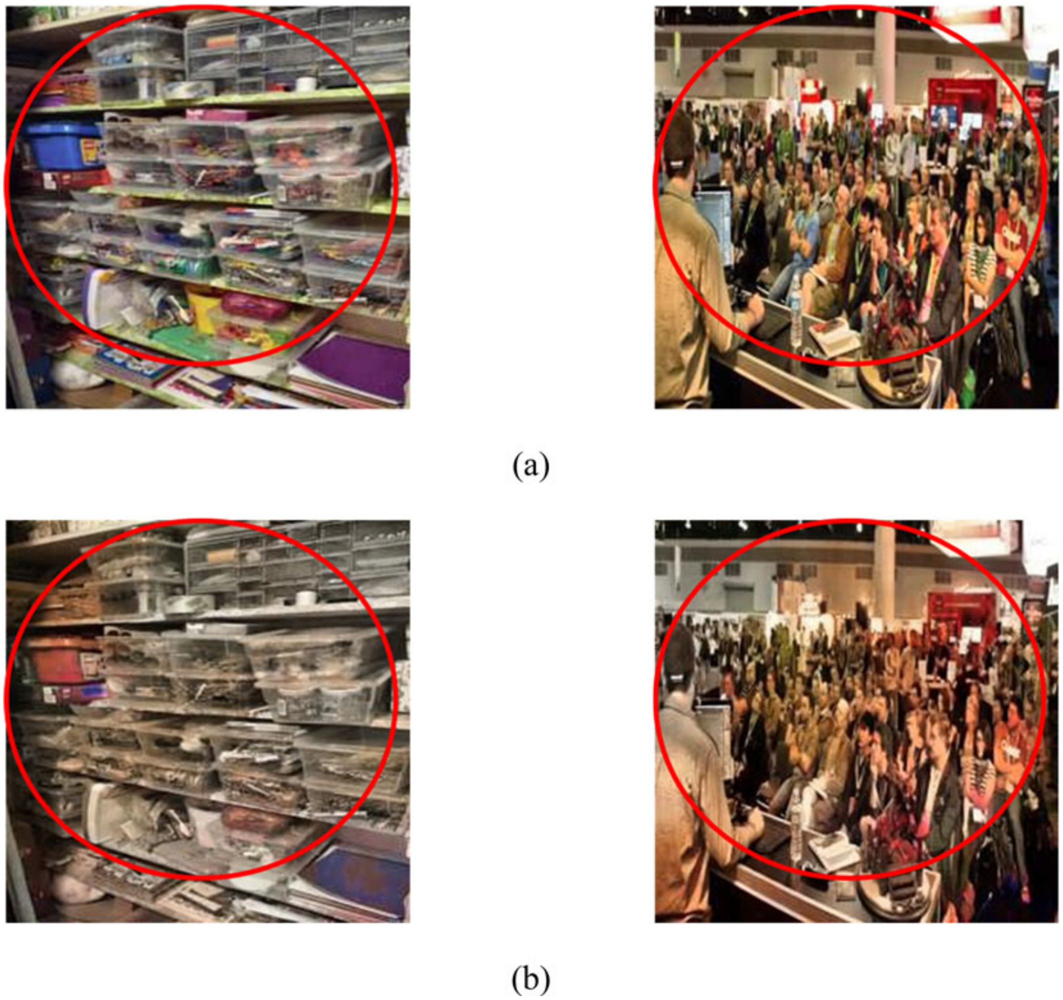
[23]	[25]	[8]	[32]	Proposed method (original design)	Proposed method (pyramidal structure)
44.5 M	32.2 M	34 M	171 M	11 M	10.3 M

B) Comparisons of different color spaces

To demonstrate that HSV is a more suitable color space for colorization tasks, we compared the performance of the YUV, Lab, and HSV color spaces using the same network architecture and training conditions. We converted the training data into multiple color spaces and then used these training data to train the colorization models, resulting in different sets of parameters for each color space. Figure 9 compares and displays the colorized results of three models. We can notice that the HSV results are more realistic than the other color spaces, especially the higher saturation color on leaves, proving that HSV is the best choice.

C) Comparisons of different methods

We compare our proposed method with four popular or recent learning-based colorization methods [8, 23, 25, 32]. Figures 10 and 11 show the results of these and our methods when colorizing indoor scenes and outdoor scenes, respectively. In comparison to the literature methods, our results for both scenes of images look more natural, have higher color saturation, and are closer to the ground truth. The


(a)

(b)

Fig. 13. Failure cases of our colorization method. (a) Ground truth. (b) Proposed method.

color of colorized images created using the [23] method is a bit dull, with more gray and brown. The results of [25] have color bleeding artifacts in the indoor scene, such as the wall, floor, and garments, and in the outdoor scene, such as the chimney, roof. The colorized images by [8] tend to have lower saturation than ours. The result of [32] has colorization defects in texture parts, such as the chimney of the house and the bricks near the roof of the castle. More colorized images generated by our method are shown in Fig. 12.

Table 1 shows the results of comparing the number of parameters of models. Comparing to others, our method uses a far smaller number of network parameters, while producing more realistic images.

D) Analysis of failure cases

Finally, we examine the failure cases of our method. We discover in the experiment that the ground truth images with much higher details than normal can sometimes cause colorization issues for our method. It is because that we begin by predicting the chroma channel at a lower scale, and it can be sometimes difficult to predict the appropriate chroma value when the targets are too small in size. This could lead to the failure in such places, such as the red-circled items in Fig. 13.

IV. CONCLUSION

In this research, we present a two-stage colorization model based on the CNN with a pyramidal structure that allows us to minimize our model parameters. To generate a colorized image, our method first generates low-scale chroma components by the LR-CMGN and then analyzes multi-scale information using the RCN. Our method addresses the issue that existing colorization methods tend to generate subdued and poor saturation results. We investigate how to create better colorized grayscale images using the HSV color space. To tackle the problem caused by the *H* component of the HSV color space represented by a color ring, we design a loss function that enables our model to predict the chroma components of HSV accurately. Our experiments with Places365-standard datasets validate that our outcomes are more natural and closer to the ground truth than previous methods.

FINANCIAL SUPPORT

This work was partially supported by the Ministry of Science and Technology under the grant number MOST 110-2221-E-027-040-.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Limmer, M.; Lensch, H.P.: Infrared colorization using deep convolutional neural networks, *in 2016 15th IEEE Int. Conf. on Machine Learning and Applications (ICMLA)*, 2016, 61–68: IEEE.

[2] Qayynm, U.; Ahsan, Q.; Mahmood, Z.; Chcmdary, M.A.: Thermal colorization using deep neural network, in *2018 15th Int. Bhurban Conf. on Applied Sciences and Technology (IBCAST)*, 2018, 325–329: IEEE.

[3] Song, Q.; Xu, F.; Jin, Y.-Q.: Radar image colorization: converting single-polarization to fully polarimetric using deep neural networks. *IEEE Access*, **6** (2017), 1647–1661.

[4] Lo, T.; Sim, K.; Tso, C.P.; Nia, M.E.: Improvement to the scanning electron microscope image adaptive Canny optimization colorization by pseudo-mapping. *Scanning*, **36** (5) (2014), 530–539.

[5] Levin, A.; Lischinski, D.; Weiss, Y.: Colorization using optimization. *ACM Trans. Graph. (TOG)*, **23** (3) (2004), 689–694.

[6] Huang, Y.-C.; Tung, Y.-S.; Chen, J.-C.; Wang, S.-W.; Wu, J.-L.: An adaptive edge detection based colorization algorithm and its applications, in *Proc. of the 13th Annual ACM Int. Conf. on Multimedia*, 2005, 351–354: ACM.

[7] Yatziv, L.; Sapiro, G.: Fast image and video colorization using chrominance blending. *IEEE Trans. Image Process.*, **15** (5) (2006), 1120–1129.

[8] Zhang, R. *et al.*: Real-time user-guided image colorization with learned deep priors, arXiv preprint arXiv:1705.02999, 2017.

[9] Xiao, Y.; Zhou, P.; Zheng, Y.: Interactive deep colorization with simultaneous global and local inputs, arXiv preprint arXiv:1801.09083, 2018.

[10] Zhang, L.; Li, C.; Wong, T.-T.; Ji, Y.; Liu, C.: Two-stage sketch colorization. *ACM Trans. Graph. (TOG)*, **37** (6) (2018), 1–14.

[11] Welsh, T.; Ashikhmin, M.; Mueller, K.: Transferring color to greyscale images. *ACM Trans. Graph. (TOG)*, **21** (3) (2002), 277–280.

[12] Gupta, R.K.; Chia, A.Y.-S.; Rajan, D.; Ng, E.S.; Zhiyong, H.: Image colorization using similar images, in *Proc. of the 20th ACM Int. Conf. on Multimedia*, 2012, 369–378: ACM.

[13] Gupta, R.K.; Chia, A.Y.-S.; Rajan, D.; Zhiyong, H.: A learning-based approach for automatic image and video colorization, arXiv preprint arXiv:1704.04610, 2017.

[14] He, M.; Chen, D.; Liao, J.; Sander, P.V.; Yuan, L.: Deep exemplar-based colorization. *ACM Trans. Graph. (TOG)*, **37** (4) (2018), 47.

[15] Lee, J.; Kim, E.; Lee, Y.; Kim, D.; Chang, J.; Choo, J.: Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence, in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, 5801–5810.

[16] Kuo, T.-Y.; Wei, Y.-J.; You, B.-Y.: Chroma component generation of gray images using multi-scale convolutional neural network, in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. (APSIPA ASC)*, 2020, 1243–1246: IEEE.

[17] Varga, D.; Szirányi, T.: Fully automatic image colorization based on Convolutional Neural Network, in *2016 23rd Int. Conf. on Pattern Recognition (ICPR)*, 2016, 3691–3696: IEEE.

[18] Larsson, G.; Maire, M.; Shakhnarovich, G.: Learning representations for automatic colorization, in *European Conf. on Computer Vision*, 2016, 577–593: Springer.

[19] Cheng, Z.; Yang, Q.; Sheng, B.: Deep colorization, in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2015, 415–423.

[20] Qin, P.; Cheng, Z.; Cui, Y.; Zhang, J.; Miao, Q.: Research on image colorization algorithm based on residual neural network, in *CCF Chinese Conf. on Computer Vision*, 2017, 608–621: Springer.

[21] Zhao, J.; Liu, L.; Snoek, C.G.; Han, J.; Shao, L.: Pixel-level semantics guided image colorization, arXiv preprint arXiv:1808.01597, 2018.

[22] Zhao, J.; Han, J.; Shao, L.; Snoek, C.G.: Pixelated semantic colorization, arXiv preprint arXiv:1901.10889, 2019.

[23] Iizuka, S.; Simo-Serra, E.; Ishikawa, H.: Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph. (TOG)*, **35** (4) (2016), 110.

[24] Cheng, Z.; Yang, Q.; Sheng, B.: Colorization using neural network ensemble. *IEEE Trans. Image Process.*, **26** (11) (2017), 5491–5505.

[25] Zhang, R.; Isola, P.; Efros, A.A.: Colorful image colorization, in European Conf. on Computer Vision, 2016, 649–666: Springer.

[26] Mouzon, T.; Pierre, F.; Berger, M.-O.: Joint CNN and variational model for fully-automatic image colorization, in *Int. Conf. on Scale Space and Variational Methods in Computer Vision*, 2019, 535–546: Springer.

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

[28] Deshpande, A.; Lu, J.; Yeh, M.-C.; Jin Chong, M.; Forsyth, D.: Learning diverse image colorization, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, 6837–6845.

[29] Lal, S.; Garg, V.; Verma, O.P.: Automatic image colorization using adversarial training, in *Proc. of the 9th Int. Conf. on Signal Processing Systems*, 2017, 84–88: ACM.

[30] Cao, Y.; Zhou, Z.; Zhang, W.; Yu, Y.: Unsupervised diverse colorization via generative adversarial networks, in *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*, 2017, 151–166: Springer.

[31] Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A.: Image-to-image translation with conditional adversarial networks, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, 1125–1134.

[32] Vitoria, P.; Raad, L.; Ballester, C.: Chromagan: adversarial picture colorization with semantic class distribution, in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, 20202445–2454.

[33] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.

[34] Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.: Conditional image generation with pixelcnn decoders, in *Advances In Neural Information Processing Systems*, 2016, 4790–4798.

[35] Goodfellow, I. *et al.*: Generative adversarial nets, in *Advances in Neural Information Processing Systems*, 2014, 2672–2680.

[36] Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S.: Feature pyramid networks for object detection, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, 2117–2125.

[37] Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J.: Pyramid scene parsing network, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, 2881–2890.

[38] Kinoshita, Y.; Kiya, H.J.A.T.O.S.; Processing, I.: Hue-correction scheme considering CIEDE2000 for color-image enhancement including deep-learning-based algorithms, vol. 9, 2020.

[39] Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A.: Places: a 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (6) (2017), 1452–1464.

[40] He, K.; Zhang, X.; Ren, S.; Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in *Proc. of the IEEE Int. Conf. on Computer Vision*, 2015, 1026–1034.

[41] Blau, Y.; Michaeli, T.: The perception-distortion tradeoff, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2018, 6228–6237.

Yu-Jen Wei received the B.S degree in the Department of Communication Engineering from the National Penghu University of Technology, Penghu, where he is currently pursuing the Ph.D. degree in electrical engineering from the National Taipei University of Technology, Taipei, Taiwan, R.O.C. He joined the Image and Video Processing Lab of the National Taipei University of Technology in 2016. His current research interests include image quality assessment, computer vision, and machine learning.

Tsu-Tsai Wei received the B.S. and M.S. degree in the Department of Electrical Engineering from the National Taipei University of Technology, Taipei, Taiwan, R.O.C in 2018 and 2021. He joined the Image and Video Processing Lab of the National Taipei University of Technology in 2017. His current research interests include convolution neural network and grayscale image colorization.

Tien-Ying Kuo received the B.S. degree from the National Taiwan University, Taiwan, R.O.C., in 1990 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1995 and 1998, respectively, all in electrical engineering. In the summer of 1996, he worked as an intern in the Department of Speech and Image Processing, AT&T Laboratories, Murray Hill, NJ. In 1999, he was the Member of Technical Staff in the Digital Video Department, Sharp Laboratories of America, Huntington Beach, CA. Since August 2000, he has been an Assistant Professor and is currently an Associate Professor with the Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan, R.O.C. He received the best paper award from the IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP) in 2008. His research interests are in the areas of digital signal and image processing, video coding, and multimedia technologies.

Po-Chyi Su was born in Taipei, Taiwan in 1973. He received the B.S. degree from the National Taiwan University, Taipei, Taiwan, in 1995 and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 1998 and 2003, respectively, all in Electrical Engineering. He then joined Industrial Technology Research Institute, Hsinchu, Taiwan, as an engineer. Since August 2004, he has been with the Department of Computer Science and Information Engineering, National Central University, Taiwan. He is now a Professor and the Dept. Chair. His research interests include multimedia security, compression, and digital image/video processing.