IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2021, accepted January 13, 2021, date of publication February 3, 2021, date of current version February 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3056144

Adversarial Edge-Aware Image Colorization With Semantic Segmentation

GUANGQIAN KONG, HUAN TIAN, XUN DUAN, AND HUIYUN LONG
School of Computer Science and Technology, Guizhou University, Guiyang 550025, China

Corresponding author: Huan Tian (huan.tian22@qq.com)

This work was supported in part by the National Natural Science Foundation of China [2018] under Grant 61741124, and in part by the Science Planning Project of Guizhou Province, Guizhou Science and Technology Cooperation Platform Talent [2018] under Grant 5781.

ABSTRACT It has become a trend in recent years to use deep neural networks for colorization. However, previous methods often encounter problems with edge color leakage and difficulties in obtaining a plausible color output from the Euclidean distance. To solve these problems, we propose a new adversarial edge-aware image colorization method with multitask output combined with semantic segmentation. The system uses a generator with a deep semantic fusion structure to infer semantic clues in a given grayscale image under chroma conditions and learns colorization by simultaneously predicting color information and semantic information. In addition, we also use a specific color difference loss with characteristics of human visual observation that is combined with semantic segmentation loss and adversarial loss for training. The experimental results show that our method is superior to existing methods in terms of different quality metrics and achieves good results in image colorization.

INDEX TERMS Colorization, semantic segmentation, multitask training, generative adversarial networks.

I. INTRODUCTION

In regard to coloring black and white photos, the first thought that comes to mind is the work of an artist named Marina Amaral. She used postprocessing to fill in the color of many famous historical photos, and the works were realistic and did not contain any holes. Most people do not have such skills, and it is very difficult to achieve color image. However, in recent years, a large number of automatic coloring methods have emerged, allowing people to easily add color to black and white pictures that provides a unique and strong visual experience.

From an aesthetic and artificial intelligence point of view, automatic coloring has a wide range of practical applications, and colorization is a promising approach in the field of self-supervised visual learning. Image colorization is the mapping of a real value gray image to a three-dimensional color image. In the early stage, users marked colors on the gray image in different areas and then colored the image through local diffusion. These methods [1]–[4] require the user to draw colored strokes on a grayscale image. Then, an optimization program generates a color image to match the user's scribble. The colored result can largely depend on how the color is chosen, so the result depends on the user's skills and experience. To reduce the complexity of use, the latter method [5] was subsequently designed with a better similarity measure, and then methods based on learning mechanisms, such as boosting [6], local linear embedding [8], feature extraction [7], etc., were employed. In recent years, some user-guided methods combined with deep neural networks have emerged, such as [9], [10]. These methods can achieve impressive results but often require intensive user interaction because each method has a different color and requires precise labeling.

To remove these limitations, researchers have explored more automated colorization methods from a data-driven perspective. The model of data-driven method involves learning the parametric mapping relation from a gray image to a color image through large-scale data. In this respect, the most mentioned methods are [11]–[17], [35]. Although these data-driven methods all have good performance, we find that the existing models do not substantively take into account the effective colorization of the object edge in the image and are limited in the selection of the loss function.

In this paper, we propose a new framework of an edge-aware colorized deep neural network with semantic segmentation to solve the above problems. Our main focus is to make

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S Raval.

FIGURE 1. Adversarial edge-aware image colorization with semantic segmentation. We propose an object edge-aware coloring method that can produce natural and colorful results in scenes with multiple objects.

the different objects in the colored image have clear coloring boundaries, which can effectively achieve the image colorization of edge perception. The above effect can be achieved in our work mainly because of the following three reasons. First, our network is based on the architecture of a generative adversarial network, and the generator of the network has the structure of deep semantic fusion. The addition of adversarial loss can generate more vivid results. Second, image colorization is the first task of our network, and the other task is semantic segmentation. That is, our work realizes multitask output [22] and constrains the network output by adding a semantic segmentation task. Basically, the two tasks share the same goal, which is to acquire as many image features as possible, so in addition to obtaining more information about the edge of the object, the color task can also be assisted. Third, we adopted a new color difference loss L_{CMC}, that makes the color difference calculation more in line with the characteristics of human visual observation. We trained and verified the proposed method on two public datasets, namely, the PASCAL VOC 2012 augmented dataset and the ADE20K dataset [23]. Figure 1 shows the results of image colorization using our method.

Our major contributions are as follows:

- A new automatic image colorization method combined with the semantic segmentation task is presented.
- A novel generative adversarial network with a deep semantic fusion structure and multitask output is proposed.
- We use a specific color difference loss with human visual observation characteristics and combine semantic segmentation loss and adversarial loss to form a multivariate loss function.

II. RELATED RESEARCH

At present, image colorization methods are mainly divided into two categories: user-guided colorization methods that require user participation, and data-driven automatic colorization methods that are completely end-to-end without human intervention.

User-guided colorization: Due to the multimodal problem of image colorization, early approaches relied on additional advanced user doodles (for example, dots or strokes) to guide the coloring process. Levin *et al.* [1] suggested assigning similar colors to adjacent pixels with similar brightness. Horiuchi and Kotera [4] proposed a new texture colorization algorithm based on mixed seed color that can accurately color the image with texture. Huang *et al.* [2] proposed a general fast coloring method based on adaptive edge detection to prevent the coloring process from overflowing the target edge. Yatziv and Sapiro [3] proposed a luminance-weighted chromaticity blend to reduce the dependence on the position of the scribble. Because strokes are propagated using low-level similarity measures, such as spatial offset and strength differences, it usually takes substantial user editing to obtain the real results. To reduce the workload of users, the latter methods use boosting [6], local linear embedding [8], feature extraction [7] and other learning mechanism methods to design better similarity measurements. These methods can yield convincing results when the user provides detailed guidance hints. However, this process requires a large amount

of human intervention (labor-intensive). Zhang *et al.* [9] partially alleviate the manual workload by combining color hints with deep neural networks.

Data-driven automatic colorization: In the existing work, deep neural network-based methods have become the mainstream learning methods, learning color prediction from large-scale datasets (such as ImageNet [24]). In this method, a large number of gray/color images are used to train the neural network, and the mapping relationship between the gray image and color channel is modeled. In terms of network structure, some approaches choose generative nonadversarial networks to predict the color channel of the image, such as Iizuka *et al.* [13] and Zhao *et al.* [25], all of which propose local image characteristics and global fusion prior information of the double branch network architecture, presuming that the color is the best way to consider details of gray images on many levels of abstraction. Zhang *et al.* [15] adopted a cross-channel coding scheme to provide semantic interpretability, learned the color distribution of each pixel and the network's training and polynomial cross-entropy loss, and allowed unusual colors to appear by rebalancing the rare class. Larsson *et al.* [14] also achieved this by pretraining their network for a classification task and working on a system that could learn the color histogram (distribution) of a given grayscale pixel. There are also methods based on generative adversarial networks(GANs), which take advantage of GANs' ability to learn the probability distribution of higher-dimensional spatial data (such as color images) to produce very good color results. For example, Isola *et al.* [16] proposed the use of conditional generative adversarial networks [26] (cGans) based on the generator of U-Net [33] to map the input image to the output image and combined $L1$ loss and adversarial loss to train the network. Nazeri *et al.* [35] extended the method of Isola *et al.* [16] by extending the coloring process to high-resolution images and proposed training strategies to speed up the process and greatly stabilize it. Cao *et al.* [27] also used cGANs and obtained various possible colors by sampling the input noise several times. Deoldify [28] used the NoGAN training method to enable the network to generate realistic color transformation, and Vitoria *et al.* [17] proposed an automatic end-to-end adversarial method combining GANs and semantic class distribution learning. It is worth noting that none of these GAN-based approaches use additional information such as semantic segmentation, while our colorization approach based on the generative adversarial network combines the distribution of semantic segmentation with color regression.

III. OVERVIEW

Our work is carried out in CIELAB color space. The network uses a $256 \times 256 \times 1$ gray image $X \in R^{H \times W \times 1}$ as input (L channel in CIELAB color space) and then predicts two color channels $\hat{Y} \in R^{H \times W \times 2}$ corresponding to the gray image (AB channels in CIELAB color space). In our work, the network is based on GANs, which enhances the effect of colorization by adding the semantic segmentation task (Section IV-A). The network architecture consists of three parts (Section IV-B), which are the feature extraction module (Section IV-B2), the reconstruction module (Section IV-B1), and the multitask output module (Section IV-B3). In addition to combining semantic segmentation loss, we also use a chromatic aberration calculation that is more consistent with human visual observation, enhancing the consistency between visual assessment and measurement of chromatic aberrations (Section IV-C).

IV. METHOD

A. SEMANTIC SEGMENTATION

Numerous works [18]–[20] in computer vision exist that integrate semantic segmentation tasks with great results. The task of image colorization has a strong correlation with semantic segmentation. Essentially, they all achieve pixel-level classification. Our backbone network adopts the U-Net-like network, which was originally used for medical image segmentation. Inspired by these works, we use the semantic segmentation task to enhance the shading ability of the network. On the one hand, the network is constrained to extract object contours; on the other hand, intensive prediction and fine-grained reasoning are developed for each pixel to assist the network to extract features. We accomplish these goals by adding a module to the output layer of the network to output semantic labels.

B. NETWORK ARCHITECTURE

Referring to our previous work [21], we use PatchGAN| [16] as a discriminator, which is based on the Markov discriminator architecture. Different from the common GAN discriminator that maps the input to a real number, PatchGAN maps the input to the $N \times N$ patch (we output a 16×16 patch). The final result is the average value of the patch, which can take into account the influence of different parts of the image. As shown in Figure 2, our network consists of three parts: a feature extraction module (green and blue), a reconstruction module (purple), and finally a multitask output module (yellow).

1) FEATURE EXTRACTION MODULE

The main network of the generator takes our previous work and uses a U-Net-like network to extract the local features (colored dark green in Figure 2). Then, a network pretrained on ImageNet is used to extract global features; however, any feature extraction network can be used, such as VGG16 [34], ResNet [29], ResNeXt [30], etc. We choose the ResNeXt50 network with the SE [31] module as the branch network (the light green part in Figure 2). In contrast to previous work, we fully utilize the pretrained branch network, which not only extracts high-level features but also contributes low-level features with higher resolution, including more location details, and then integrates features of the same size that are extracted from the trunk network to improve the network's perception

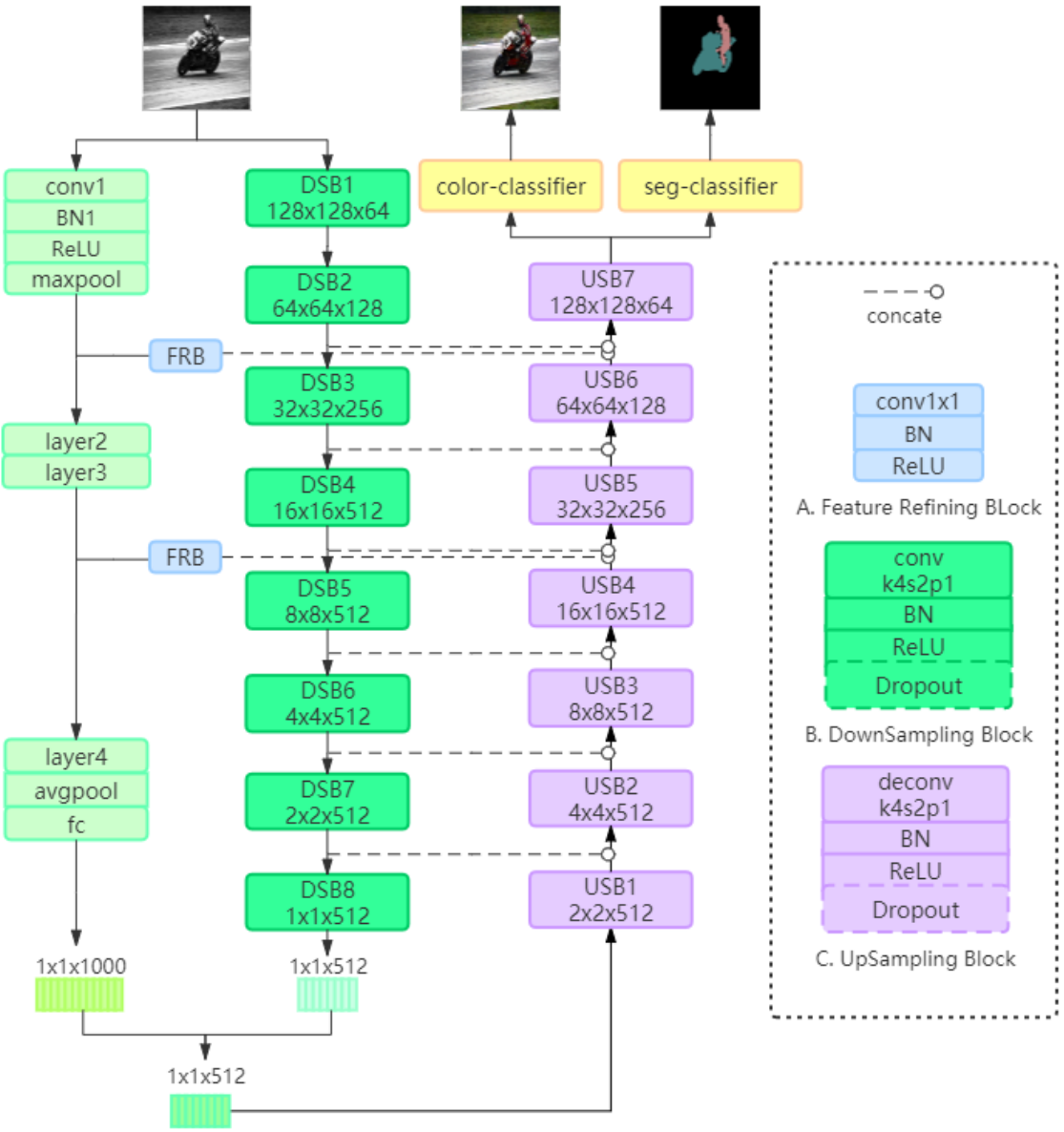

FIGURE 2. Overview of the generator. All the green and blue parts belong to the feature extraction module, the purple part belongs to the reconstruction module, and the yellow part belongs to the multitask output module. The input of the model is a gray image with a size of 256 × 256 × 1, and the output of the prediction result of the AB channel has a size of 256 × 256 × 2 and a semantic label with a size of 256 × 256 × *C*. The right side of the figure shows the specific structure of the corresponding color block. The content in the solid line box is required, and the content in the dotted line box is optional.

of details. The grayscale image X with the size of $256 \times 256 \times 1$ is taken as the input of both feature extraction networks at the same time. In the local feature extraction network, the size of the image is reduced to $\frac{1}{2}$ of the input every time it passes through a downsampling module. After 8 downsampling operations, features with a size of $1 \times 1 \times 512$ are obtained. The final output of the pretrained global feature extraction network is $1 \times 1 \times 1000$. In addition, we extract the features of the middle layer whose two feature sizes are $64 \times 64 \times 256$ and $16 \times 16 \times 512$. After the feature refining block (FRB, colored blue in Figure 2), the number of channels is unified to 512.

2) RECONSTRUCTION MODULE

After the features of the low and high levels are extracted, the multilayer features are fused through early fusion, as shown in the purple part of Figure 2. We used deconvolution to carry out upsampling of features with an upsampling factor of 2. In particular, before the fifth and seventh upsampling modules were carried out, features passing through FRB were fused with trunk features of the same size through the concat operation. This structure, which integrates more low-level semantic features, contributes more spatial information and object details to the network and provides support for network prediction.

3) MULTITASK OUTPUT MODULE

Our generator has two kinds of outputs, so there are two branches in the output module (colored yellow in Figure 2). The first output branch outputs the prediction results of the AB channel of $256 \times 256 \times 2$, which calculates the color difference loss with the GT's AB channels. The second output branch outputs the prediction of the semantic label with the size of $265 \times 256 \times C$ (C refers to the number of categories of semantic labels), and this part obtains the segmentation loss from the semantic label.

C. LOSS FUNCTION

The color difference can be calculated by the Euclidean distance in color space; however, different colors are in fact not linearly separable. In other words, in our visual system, the color difference is very large, when, in fact, the Euclidean distance between colors may be very small. Therefore, we cannot simply use the Euclidean distance to judge the color difference. In addition, the sensitivity of human eyes to hue, saturation and lightness is different. We first observe the difference of hue, then saturation, and finally the difference of lightness. Consequently, the true visually acceptable tolerance is not equal to the hue, saturation and lightness. The CMC(l:c) color difference loss can achieve better consistency between visual evaluation and the measured color difference. Therefore, according to this characteristic, we adopt a color difference loss L_{CMC} that is closer to human visual observation, as follows:

$$\mathcal{L}_{CMC} = \sqrt{(\frac{L_2^* - L_1^*}{lS_L})^2 + (\frac{C_2^* - C_1^*}{cS_C})^2 + (\frac{H_2^* - H_1^*}{S_H})}. \quad (1)$$

$$S_L = \begin{cases} 0.511, & L_1^* \leq 16, \\ \frac{0.040975L_1^*}{1+0.01765L_1^*}, & L_1^* \geq 16. \end{cases} \quad (2)$$

$$S_C = \frac{0.0638C_1^*}{1 + 0.0131C_1^*} + 0.638. \quad (3)$$

$$S_H = S_C(FT + 1 - F). \quad (4)$$

$$F = \sqrt{\frac{C_1^{*4}}{C_1^{*4} + 1900}}. \quad (5)$$

$$T = \begin{cases} 0.56 + |0.2\cos(h_1 + 168°)|, & 164° \leq h_1 \leq 345°, \\ 0.36 + |0.4\cos(h_1 + 35°)|, & otherwise. \end{cases} \quad (6)$$

The CMC(l:c) color difference formula [32] defines the ellipsoid mathematically around the half-axis of the standard color and luminance (L^*), chromaticity (C^*) and hue (H^*). In the above formula, L_1^*, C_1^*, and H_1^* are the chromaticity parameters of the standard color, but in the image colorization task, the chromaticity parameters of the standard color refer to the chromaticity parameters of the GT. S_L, S_C and S_H are the half-axis of the ellipse, and $l : c$ is the ratio of lightness to saturation. The larger the ratio is, the larger the tolerance range. The length of the relative half-axis can be changed by two parameters l and c (we set $l : c$ as $2 : 1$), and then the relative tolerance of ΔL^*, ΔC^*, and ΔH^*.

In our network, semantic segmentation task is used to improve the ability of edge division of the object on the network. At the same time, it can assist the network to learn the color and contour of objects. The loss function of semantic segmentation is $\mathcal{L}_{seg}$ as follows (where p represents the segmentation label and q represents the network segmentation result):

$$\mathcal{L}_{seg} = -\sum_C p_i \log q_i. \quad (7)$$

The general objective function of a GAN is as follows:

$$\mathcal{L}_g(G, D) = \mathbb{E}_{y \sim p_{data}}[\log D(y)] + \mathbb{E}_{x \sim p_g}[\log(1 - D(G(x))]. \quad (8)$$

The generator G tries to minimize $\mathcal{L}_g(G, D)$, and the discriminator D iterates to maximize $\mathcal{L}_g(G, D)$. However, considering that our network follows multi-output and multitask forms, the final objective is as follows:

$$G^* = (1-\lambda) \arg\min_G \max_D \mathcal{L}_g(G, D) + \lambda \mathcal{L}_{CMC} + \mathcal{L}_{seg}. \quad (9)$$

λ is the weighted hyperparameter and set to 0.99 in the experiment to ensure the two losses of $\mathcal{L}_g(G, D)$ and $\mathcal{L}_{CMC}$ have a consistent order of magnitude.

V. EXPERIMENT

In this part, we will show the experimental results to verify the effectiveness of the proposed edge-aware image colorization method. In Section V-A, the datasets and evaluation metrics used in the experiment and some details of model training are described. In Section V-B, the color images obtained by our method and other methods are compared and analyzed quantitatively. Finally, the effectiveness of the model is studied in Section V-C, and qualitative analysis is carried out.

A. EXPERIMENTAL SETTING

1) DATASETS

We used two kinds of data to train and verify the model, namely, the segmented PASCAL VOC 2012 augmented dataset and the ADE20K dataset.

PASCAL VOC 2012 Augmented Dataset: the object categories in the dataset are divided into 21 categories, including background. There are 11,355 images, with 10,582 for training, 1449 for evaluation and 1456 for testing.

ADE20K Dataset: this is a scene dataset containing 150 categories and a total of 22,210 images, of which 20,120 are the training set and 2000 are the validation set.

2) EVALUATION METRICS

We use a variety of comprehensive evaluation indicators to evaluate the prediction results. The selected indicators are peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and *Image Entropy*.

TABLE 1. Quantitative comparison on two datasets.

Method	Pascal VOC 2012 augmentation validation split			ADE20K validation split		
	PSNR	SSIM	*Image Entropy*	PSNR	SSIM	*Image Entropy*
Iizuka *et al.* [13]	29.851	0.955	7.960	29.88	0.979	7.882
Larsson *et al.* [14]	29.521	0.953	7.934	29.88	0.979	7.882
Zhang *et al.* [15]	30.249	0.951	7.845	**30.846**	0.986	7.701
Nazeri *et al.* [35]	29.415	0.928	7.918	28.878	0.896	**7.939**
DeOldify [28]	29.953	0.939	7.955	29.428	0.979	7.915
ChromaGAN [17]	29.842	0.953	7.936	28.842	0.964	7.902
Ours	**31.359**	**0.972**	**7.972**	30.650	**0.987**	7.778

3) TRAINING DETAILS

In our experiment, one single GTX 1080ti GPU was used to train the model, and when the input size was set to 256×256, the FLOPs of the network was 37.596GFLOPS, the parameters consumed 328.29MB of memory, and forward/backward pass consumed 919.02MB of memory. Before inputting a single-channel image, we enhance the data, including histogram equalization and random flipping. Histogram equalization can enhance the contrast of the image and compensate for the difference of gray levels which is difficult to distinguish by visual entropy. In all the training processes, we use the ADAM optimizer with $\beta 1$=0.5 and $\beta 2$=0.999.

B. QUANTITATIVE COMPARISONS

In this section, the proposed method is evaluated quantitatively. In Table 1, we report the quantitative comparison of the methods in the experiments on the two datasets. The table shows the average values of all test images on each evaluation metric. On the PASCAL VOC 2012 augmented dataset, our model achieves better values than the latest methods [13]–[15], [17], [28], [35]. However, we can observe that our method does not achieve the highest score of the three indicators on the ADE20K dataset, but compared to the method of Zhang *et al.* [15], which obtains the PSNR highest score, our method exceeds its corresponding scores in the other two indicators. Similarly, compared to Nazeri *et al.* [35], which obtains the highest *image entropy* score, our method exceeds its scores in the first two indicators. This is not surprising, since our network not only has a deep semantic fusion structure but also incorporates a multitask design that enhances the ability of the network to capture object boundaries and achieve a fine-grained reasoning ability. From the perspective of comparing the loss of this work and the loss of other methods, the loss function $\mathcal{L}_{CMC}$ is close to human visual observation, reduces the tolerance of color differences and gives more guidance for updating parameters.

Figure 3 shows our colorization results compared to the advanced methods. By using a public online demo, our results are compared to those obtained in [13]–[15], [17], [28], [35]. The first three row methods are the deep convolutional neural network without adversarial training, and the last four row methods are based on the generative adversarial network.

TABLE 2. Quantitative comparison of the images in Figure 4 that are ordered from top to bottom.

Method	PSNR	SSIM	*Image Entropy*
+seg	32.472	0.988	7.755
-seg	31.489	0.983	7.870

It can be observed that the method in Iizuka *et al.* [13], Larsson *et al.* [14] and Zhang *et al.* [15] can color the plants with obvious texture features. For some artificial objects, they tend to output soft colors. However, Zhang *et al.* [15] trained with polynomial cross-entropy and used classification to rebalance rare classes, especially allowing unusual colors in color images, such as colors with high saturation in the third image. DeOldify's [28] overall tone is uniform but tends to be desaturated. Compared to [13]–[15] in the test sample, Nazeri *et al.* [35] produced more colorful images, but they were uneven, while Deoldify's [28] better coloring effect on characters was not maintained in other aspects. ChromaGAN [17] and our approach can achieve a more lively appearance and are closer to the ground truth, especially in regard to the colored drinks in the third row that look very tempting. Nevertheless, ChromaGAN [17] has insufficiently dealt with some details. For example, in the picture of the second row of buses, the roadside signs are not accurately identified and are confused with the surrounding objects. In the sixth row of images, the outdoor scenery of the window is also given the same light brown yellow color as the surrounding environment. In contrast, given the influence of multitask learning, the generator with the deep multifeature fusion structure, and the color difference loss in line with human visual perception, our network can not only obtain rich semantic information but also emphasize the edge of objects. In these test cases, it shows excellent color ability. The helmets of motorcyclists appear beautifully red, the buses and landmarks beside them have natural and realistic colors, the overlapping drink bottles also have rich colors, and the dinner plate on the table appears very good with the meal while the table is different. The birds standing on the branches are almost the same as the real image, and the scenery outside the window has its own color.

C. ABLATION STUDY

We conduct ablation studies by evaluating a variant of our method and different color difference loss choices. There

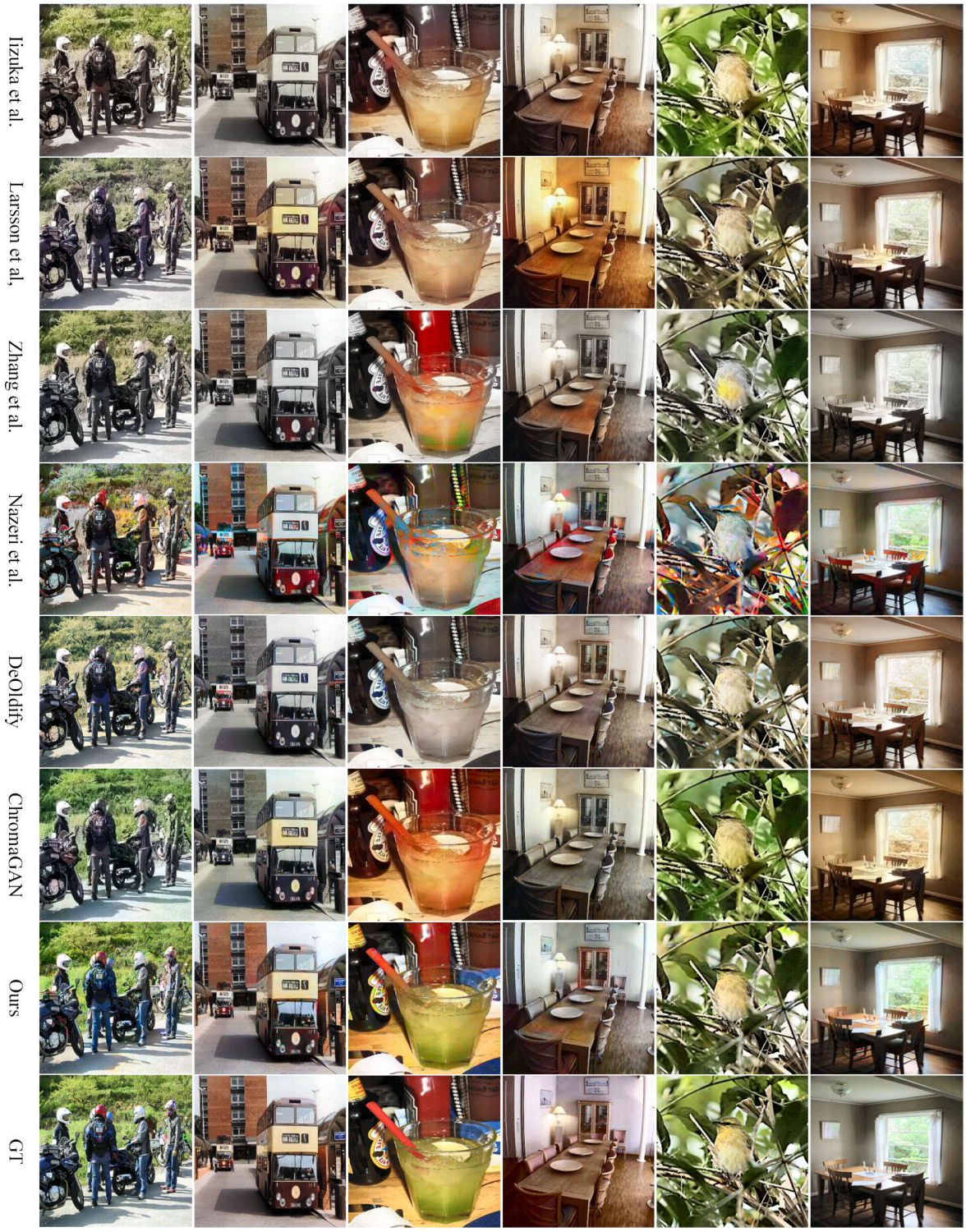

FIGURE 3. From left to right are several qualitative results of the methods of Iizuka *et al.* [13], Larsson *et al.* [14], Zhang *et al.* [15], Nazeri *et al.* [35], DeOldify [28], ChromaGAN [17] and our approach. The last column is the GT.

are two main components that are critical to our final performance: the visual perception color difference loss and the semantic segmentation edge-aware function. We conducted ablation studies to assess the effectiveness of each component.

In the first ablation experiment, the semantic segmentation edge-aware function was disabled (we denote our approach as Ours and its variants with $+$**seg** and $-$**seg**). We can observe that the red chair in the first row (a) of Figure 4 has a certain degree of color leakage without the constraint of the semantic

TABLE 3. Quantitative comparison of the images in Figure 5 from top to bottom, where each image is represented by *image-x* (x = 1, 2, 3).

Method	$\mathcal{L}_{CMC}$			L_2		
	PSNR	SSIM	*Image Entropy*	PSNR	SSIM	*Image Entropy*
image-1	32.200	0.988	7.789	30.825	0.950	7.753
image-2	32.460	0.991	7.867	29.774	0.942	7.818
image-3	30.667	0.968	7.814	30.208	0.934	7.864

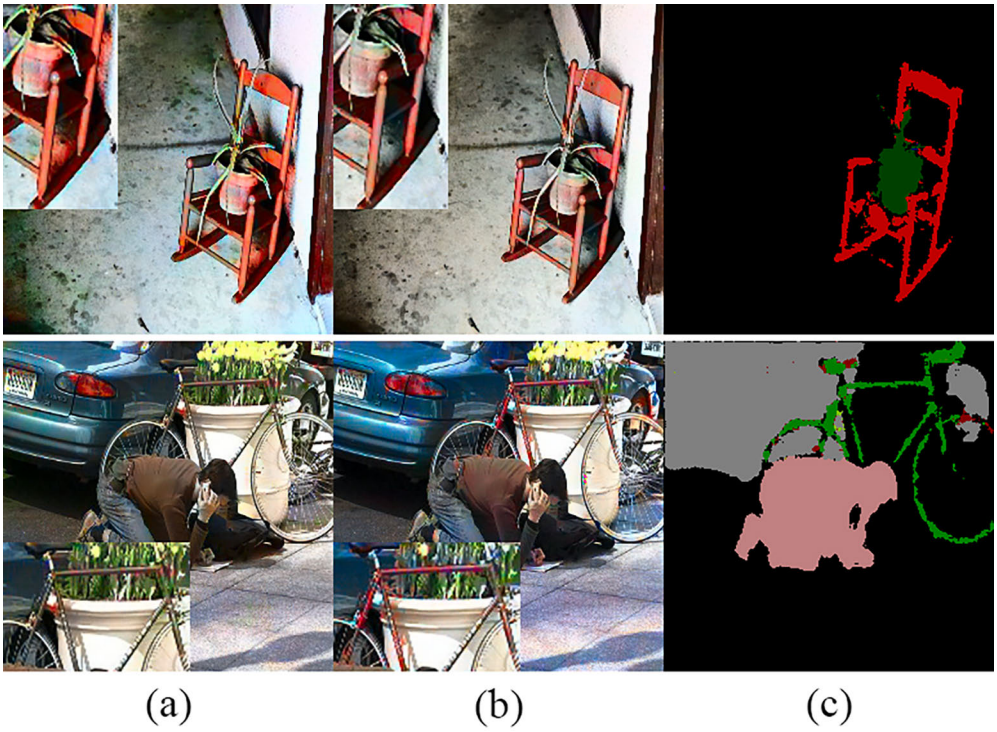

FIGURE 4. Visualization of the effectiveness of the semantic segmentation edge-aware function. Column (a) is the colorization results of variant (−seg); column (b) is the colorization result of our proposed method (+seg); and column (c) is the prediction result of the semantic label obtained simultaneously with column (b).

segmentation task. With the semantic segmentation module, the color of the chair is changed so that the color range is corrected, and the red chair in column (b) is the revised result. The bicycle in column (a) of the second row also has problems in color prediction. The body of the bicycle is not effectively colored. It can be observed that our method of multitask training enhances the network's ability to recognize objects and effectively segment the bicycles, people and traveling cars, so that the bicycle bodies in column (b) can be colored well. Therefore, we conclude that the combination that includes the semantic segmentation edge-aware function can effectively avoid the phenomena of unclear color and edge color leakage of these objects. For the images in Figure 4, we mainly focus on the reconstruction quality of the reconstructed images when the color of the images obtained by the two methods is very good. According to the data in Table 2, the method with the semantic segmentation edge-aware function can reconstruct the image better. The PSNR value is 5.3% higher than that of the method without the semantic segmentation edge-aware function, and the SSIM value is 3.1% higher.

In the second ablation experiment, Euclidean distance (L_2 loss) and $\mathcal{L}_{CMC}$ were used to calculate the color difference. The experimental results are shown in Figure 4. We can see that the colorization result using L_2 loss is very poor in predicting some bright colored objects, which is almost in the middle value of the color, while $\mathcal{L}_{CMC}$ predicts satisfactory results. The objects are more evenly and naturally colored, and the color difference was smaller than that using L_2 loss. Although there is still a slight difference in saturation with the GT, compared to the L_2 loss, it has been greatly improved and can achieve effective colorization. The results of their quantitative analyses are shown in Table 3. From the results of the three evaluation metrics, the evaluation results of $\mathcal{L}_{CMC}$ are also better than that of Euclidean distance. Although the value of *Image Entropy* in image-2 is lower than that of using L_2 loss, considering the other two metrics, $\mathcal{L}_{CMC}$ can obtain better results.

In general, $\mathcal{L}_{CMC}$ achieves an improvement that addresses the network's difficulty of tending to average values when the color is uncertain, which is caused by the traditional Euclidean distance.

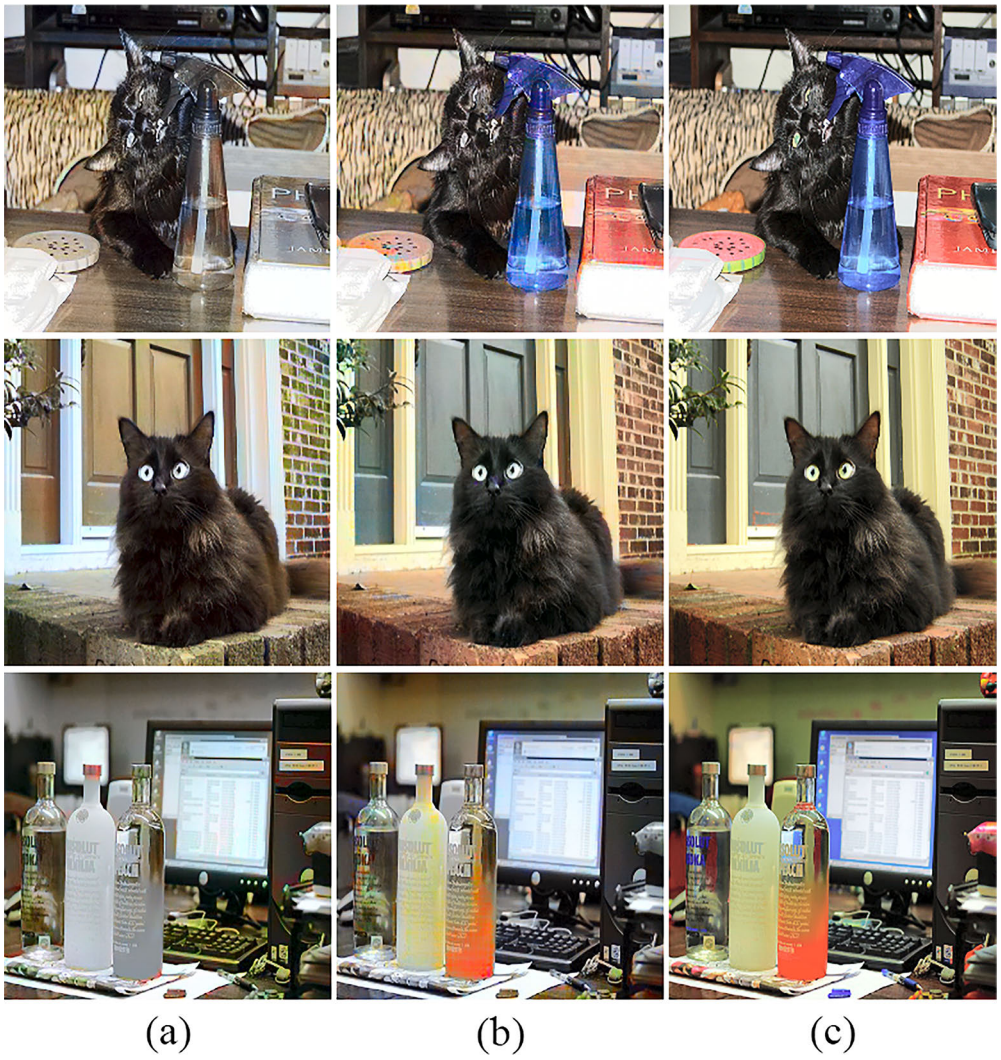

FIGURE 5. Visualization of the results obtained by different color difference losses. Column (a) is the colorization result of Euclidean distance L_2 loss, column (b) is colorization result of CMC(l:c) color difference loss $\mathcal{L}_{CMC}$, and column (c) is the real label image.

VI. CONCLUSION

In this paper, we propose a new method of image colorization based on a generative adversarial network combined with the semantic segmentation task. The network outputs color images and semantic segmentation tags at the same time. Our system can achieve better image colorization by using a color difference calculation loss function, a semantic segmentation loss function and a loss resistance loss function with the characteristics of human visual observation. Therefore, with the help of the semantic segmentation task, our method can better realize the color of the image and effectively improve the problem of graying color and color leakage. Compared to the advanced methods, our method is superior to them in PSNR, SSIM, and *Image Entropy*. At present, our system is trained in cases with fewer semantic categories. To obtain more accurate color results, semantic classification labels with greater accuracy are necessary, but this also increases the difficulty of semantic segmentation. In future work, we will continue to study the balance between color tasks and semantic segmentation/instance segmentation.

ACKNOWLEDGMENT

The authors would like to thank the Editor-in-Chief, an Associate Editor, and reviewers for their insightful comments and suggestions.

REFERENCES

[1] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, Aug. 2004.
[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 *Proc. 13th Annu. ACM Int. Conf. Multimedia (MULTIMEDIA)*, 2005, pp. 351–354.
[3] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1120–1129, May 2006.

[4] T. Horiuchi and H. Kotera, "Colorization for monochrome image with texture," in *Proc. 13th Color Imag. Conf.*, Jan. 2005, pp. 245–250.
[5] Q. Luan, W. Fang, D. Cohen-Or, L. Lin, and H. Y. Shum, "Natural image colorization," in *Proc. 18th Eurogr. Conf. Render. Tech. (EGSR)*, 2007, pp. 309–320.
[6] Y. Li, E. Adelson, and A. Agarwala, "ScribbleBoost: Adding classification to edge-aware interpolation of local image and video adjustments," in *Proc. 9th Eurogr. Conf. Render. Tech. (EGSR)*, 2008, pp. 1255–1264.
[7] L. Xu, Q. Yan, and J. Jia, "A sparse control model for image and video editing," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 197.1–197.10, 2013.
[8] X. Chen, D. Zou, Q. Zhao, and P. Tan, "Manifold preserving edit propagation," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–7, Nov. 2012.
[9] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–11, Jul. 2017.
[10] Y. Xiao, P. Zhou, Y. Zheng, and C.-S. Leung, "Interactive deep colorization using simultaneous global and local inputs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1887–1891.
[11] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 415–423.
[12] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 567–575.
[13] 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 Trans. Graph.*, vol. 35, no. 4, pp. 110.1–110.11, 2016.
[14] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 577–593.
[15] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 649–666.
[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[17] P. Vitoria, L. Raad, and C. Ballester, "ChromaGAN: Adversarial picture colorization with semantic class distribution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2445–2454.
[18] V. Pham, S. Ito, and T. Kozakaya, "BiSeg: Simultaneous instance segmentation and semantic segmentation with fully convolutional networks," in *Proc. Brit. Mach. Vis. Conf.*, London, U.K., Sep. 2017, pp. 1–12.
[19] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1328–1338.
[20] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[21] H. Tian, G. Kong, and Y. Wu, "Dual feature extractor generative adversarial network for colorization," *J. Electron. Imag.*, vol. 29, no. 4, pp. 1–14, 2020.
[22] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*. [Online]. Available: http://arxiv.org/abs/1706.05098
[23] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5122–5130.
[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[25] J. Zhao, L. Liu, C. G. M. Snoek, J. Han, and L. Shao, "Pixel-level semantics guided image colorization," 2018, *arXiv:1808.01597*. [Online]. Available: http://arxiv.org/abs/1808.01597
[26] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784
[27] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu, "Unsupervised diverse colorization via generative adversarial networks," in *Proc. Mach. Learn. Knowl. Discov. Databases (ECML PKDD)*, 2017, pp. 151–166.
[28] J. Antic, *Jantic/Deoldify: A Deep Learning Based Project for Colorizingand Restoring Old Images (and Video!)* Accessed: 2019. [Online]. Available: https://github.com/jantic/DeOldify
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[30] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.
[31] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[32] M. R. Luo and B. Rigg, "Uniform colour space based on the CMC(l: C) colour-difference formula," *J. Soc. Dyers Colourists*, vol. 102, nos. 5–6, pp. 164–171, Oct. 2008.
[33] O. Ronneberger and P. Fischer, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Comput. Comput. Assist. Interv. (MICCAI)*, 2015, pp. 234–241.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[35] K. Nazeri, E. Ng, and M. Ebrahimi, "Image colorization using generative adversarial networks," in *Proc. Int. Conf. Artucul. Motion Deform. Objects*, 2018, pp. 85–94.

GUANGQIAN KONG received the B.S. degree from Central South University, Hunan, China, in 1996, and the M.S. and Ph.D. degrees from Guizhou University, Guizhou, China, in 2002 and 2009, respectively. He is currently an Associate Professor and a Graduate Supervisor. His research interests include computer networks, big data, deep learning, and their applications. He is also a member of the China Computer Federation.

HUAN TIAN received the B.E. degree in computer science and technology from the School of Data Science, Tongren University, Tongren, in 2018. She is currently pursuing the M.E. degree in computer technology with the School of Computer Science and Technology, Guizhou University, Guizhou, China. Her research interests include deep learning, computer vision, image processing, and image colorization.

XUN DUAN received the B.S. degree from the Chongqing University of Posts and Telecommunications, in 1995, and the Ph.D. degree in computer science from Guizhou University, in 2007. Since 2007, he has been an Associate Professor with the School of Computer Science and Technology, Guizhou University. His current interests include big data, deep learning, and computer vision.

HUIYUN LONG received the Ph.D. degree in computer science from Guizhou University, Guiyang, China, in 2009. She is currently an Associate Professor and the Dean of the School of Computer Science and Technology, Guizhou University. Her main research interests include process algebra, formalization of web services, and deep learning.