






Fine-grained semantic ethnic costume high-resolution image colorization with conditional GAN

Di Wu¹  | Jianhou Gan^{1,2}  | Juxiang Zhou^{1,2}  |
Jun Wang^{1,2}  | Wei Gao^{1,3} 

¹Key Laboratory of Education Informatization for Nationalities, Ministry of Education, Yunnan Normal University, Kunming, China

²Yunnan Key Laboratory of Smart Education, Yunnan Normal University, Kunming, China

³School of Information, Yunnan Normal University, Kunming, China

Correspondence

Jianhou Gan, Key Laboratory of Education Informatization for Nationalities, Ministry of Education, Yunnan Normal University, 650500 Kunming, China.
Email: ganjh@ynnu.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 61862068, 6216020621

Abstract

Grayscale image colorization, especially for ethnic costume images, is highly challenging due to its rich and complex color features. The existing image colorization methods usually take the costume image as a whole in practical applications that lead to the ignorance of the semantic information of different parts of the costume. It is known that each part's color distribution of the ethnic costume is different. So, the color mapping of other parts is also diverse, which is determined by distinctive ethnic characteristics. This study introduces fine-grained level semantic information and proposes a high-resolution image colorization model for ethnic costumes targeting enhancement, inspired by semantic-level colorization. The semantic information of different regions of ethnic costumes has a significant impact on the performance of the coloring task. Using Pix2PixHD as the backbone network, we create a new network architecture that maintains color distribution correspondence and spatial consistency of costume images using fine-grained semantic information. In our network, we take the splice result of fine-grained semantic for ethnic costume and grayscale image as the conditions and then feed them into the generative adversarial networks. We also discuss and analyze the influences of the grayscale channel and

fine-grained semantics on discriminator. Extensive experiments demonstrate that our method performs well compared with other state-of-the-art automatic colorization methods.

KEYWORDS

ethnic costume image, fine-grained semantic, generative adversarial networks, image colorization

1 | INTRODUCTION

Grayscale image colorization is a popular topic in computer and image processing and has a range of applications in aspects, such as image encoding, color correction, image restoration,^{1–3} and so forth. In the CIE *Lab*⁴ color space, the automatic grayscale image coloring process predicts the values of the other two color channels based on the single-channel grayscale channel information. The colorization of grayscale images is fundamentally ill-posed, but it may also be multimodal.⁵ In other words, multiple trustworthy colorization outcomes can be chosen (e.g., the T-shirt in grayscale can be either blue or red). Automatic grayscale image colorization methods based on deep learning have been widely used, including References [6–11], which rely heavily on training data and confront limited semantic understanding.

The ethnic costume is the vivid symbol of each ethnic group, with great diversity and independent development. Numerous ethnic costumes have formed their distinctive cultural characteristics, change to of which the most prominent aspect is the color of the costume. Ethnic costumes represent the different cultures of each ethnic group, and color is a unique way of expression. The primary colors of Dai costumes are red, light yellow, and light green. The Hani costumes' primary color tones are green, blue, red, and so forth. Minority cultures can thus be better protected and passed down through the use of grayscale image colorization technology.

However, because of the complex design and rich color distribution of ethnic costumes, it is difficult to achieve better results when compared with typical costumes when using general coloring methods. Although additional instance-level semantic conditions work well for natural images, colorized models perform poorly, particularly for ethnic costumes.

This paper takes ethnic costumes as the research object and explores the impact of the different regional color distributions on the automatic coloring technique of ethnic costume grayscale images. We propose a fine-grained level semantic ethnic costume high-resolution grayscale image coloring method based on a conditional generative adversarial network (GAN) to address the above issues. Conditional GAN (cGAN) has an excellent performance in image translation, converting source images into target images. The grayscale image is the source image for the grayscale image coloring task. Because the target image is a color image, coloring grayscale images of ethnic costumes can be thought of as image translation. To improve the colorization of ethnic costume grayscale images, we use the semantics of costumes in different regions. Automatic coloring methods typically ignore semantic information. When it comes to ethnic costumes, there are significant differences between the various areas of the costume. Previous automatic grayscale image coloring methods did not consider the color distribution of

different areas of the costume, only grayscale data. As input generator conditions, we use fine-grained level semantic information and grayscale image information in our approach. Among them, the most important thing is to solve problems such as color distribution correspondence and spatial consistency of costume images through the semantic information of various costumes. The results of coloring under different conditions are shown in Figure 1. “W/O Semantic” means that only grayscale images are used as input. “With Semantic” means using the semantic mask of the background and costume regions as additional input. Our method shows a significant advantage from the comparative effectiveness, especially in the red box area. The coloring without using fine-grained semantic information as additional input conditions is ineffective. The treatment of local details of ethnic costumes is not good, and spatial consistency is not considered.



FIGURE 1 The first line shows the colorization results with no semantic information as the coloring condition input. The second line shows the colorization results with single semantic information as the coloring condition input. The third line shows the colorization results with fine-grained semantic information as the coloring input [Color figure can be viewed at wileyonlinelibrary.com]

The main contributions can be summarized as follows:

1. We propose utilizing fine-grained level semantics in the ethnic costume grayscale image coloring task, which adds additional input conditions to guide the colorization of ethnic costume grayscale images.
2. For the task of coloring grayscale images of ethnic costumes, we propose a novel coloring model based on Pix2PixHD. Compared with traditional coloring methods, it is more effective for the task of coloring ethnic costumes with complex design and rich color.
3. We constructed an ethnic costume data set consisting of four Chinese minority groups and applied fine-grained level semantic annotation to each image in the data set.
4. We demonstrate the effectiveness of the proposed coloring model, and our model performs well in the task of coloring grayscale images of ethnic costumes compared with other mainstream coloring methods.

The remaining parts of our paper are as follows. In Section 2, we first introduce the related work. Then, the details of our proposed method are presented in Section 3. The experiment setup consists of four parts in Section 4. Finally, we conclude the paper in Section 5.

2 | RELATED WORK

This section briefly reviews the main related techniques and works of grayscale image colorization.

- *Condition GANs*: The GANs¹² have been widely used in image translation¹³ and image generation. GANs are made up of two components: a generator and a discriminator. The generator is in charge of creating the image, making the output as deceptive as possible to the discriminator, which is used to determine whether the image is real or fake, to assess the quality of the image generated by the generator. During the training process, the two subnetworks are constantly optimized. The generator's final image effect performs well in many fields due to the design concept of adversarial loss.¹³ The Pix2Pix model applies GAN to supervise image translation. The CycleGAN¹⁴ model uses two mirrored GANs to transform the source image and the target image to each other. The DiscoGAN¹⁵ model shares the same design idea as CycleGAN and is important in image translation. On the basis of cGAN, various grayscale image coloring schemes using semantic conditions from coarse to fine are proposed. Hensman and Aizawa¹⁶ used Single Training Image to color Manga based on cGAN. Vitoria et al.¹⁷ designed an adversarial image coloring method incorporating semantic-level information, using grayscale images with additional semantic conditions as the input to the GAN. Su et al.¹⁸ further guided image colorization by employing instance-level object semantics. On the basis of the object classification of grayscale images, its colorization results are superior to the former methods. Moreover, an image colorization method presented by Zhao et al.¹⁹ can be achieved by pixel-level semantic embedding and pixel-level semantic generators.
- *Automatic colorization*: With the great progress of deep learning in image processing, deep neural networks' training and the mapping of grayscale images to color images are achieved by using large-scale grayscale images and corresponding color images, to complete the grayscale image colorization. Grayscale image coloring methods are generally divided into the following three categories: (i) scribble-based, (ii) exemplar-based, and (iii) learning-based.²⁰ Automatic

colorization is based on learning and was first proposed by Cheng et al.¹ Zhang et al.²¹ developed a user-guided coloring with learned deep priors, which can be colorized in real-time according to user requirements in the interaction stage. Messaoud et al.²² established a conditional random field-based variational auto-encoder algorithm to solve structural inconsistencies occurring in coloring tasks. It not only achieves diversity but also ensures the structural consistency of the results. Yoo et al.²³ proposed a novel colorization model of the memory augmented networks method. This method points out that high-quality colorizing results can be obtained from a small sample of training data, and the training of the memory network can be unsupervised through the threshold triplet loss without using a classification tag.

- *Fine-grained Semantic Colorization:* With the spurt of progress in computer vision, more and more researchers have paid attention to coarse-to-fine semantics in recent years. Recent studies have shown that the method of completing coloring tasks based on semantic information is adequate. In ChromaGAN, GAN, and semantic classification distribution are used as conditions the colorization,²⁴ and the model training strategy is fully self-supervised. Su et al.¹⁸ and Zhao et al.¹⁹ utilized object-level semantic information for image colorization. The former proposed architecture for achieving instance-aware colorization that used target detection and an instance coloring network to extract object-level features, then combined them with image-level features to complete colorization. To find the target object while completing the target coloring, the latter proposed two types of merging object semantics into the coloring model, pixel-level semantic features and pixel-level semantic generators.

Each Chinese ethnic minority has developed its cultural characteristics since ancient times, which are most visibly reflected in the color of their costumes. Zhao et al.²⁵ proposed a retrieval method of ethnic costume based on integrated regional matching, using regional colors for ethnic costume retrieval. Zhenrong et al.²⁶ proposed Zhuang ethnic costume images generation method based on GAN. Liu et al.²⁷ proposed a GAN-based method for coloring ethnic costume sketches. Currently, there are no specific research results on the subject of grayscale image coloring of ethnic costumes. We pioneered an ethnic costume image data set and applied it to grayscale image coloring of ethnic costumes. A conditional adversarial network (pix2pixHD) approached for an image to image conversion based on coarse-to-fine generators and multiscale discriminator structure. This paper combines the costume grayscale information and fine-grained level semantic information as conditional inputs to the network model. Then, high-resolution color ethnic costume images are generated.

3 | METHOD

In this paper, we propose a grayscale image coloring method for ethnic costumes. The grayscale image colorization model for ethnic costumes is proposed, which adopts the main network structure of Pix2PixHD. Mainly, the costume fine-grained level semantic information is used as one of the generator network input conditions and applied for the first time to the ethnic costume grayscale image coloring task. We take the grayscale image $X \in \mathbb{R}^{H \times W \times 1}$ and the fine-grained semantics mask $M \in \mathbb{R}^{H \times W \times k}$ as input conditions into our generator model, where H and W are the height and width of the images. k is the number of categories of fine-grained semantics in our data set; the size of k shows the degree of costume semantic fine-grainedness. In other words, the more fine-grained semantic categories there are in the data set, the more detailed the semantic fine-grainedness will be.

The fine-grained semantic division of the data set should fully reflect the ethnic cultural characteristics for colorizing grayscale images of ethnic costumes. As a result, our ethnic costume data set has fine-grained semantics for seven categories, sleeves, coats, belts, dresses, pants, leg guards, and accessories are all available. Furthermore, to these seven ethnic costume fine-grained categories, the background part of the image is also considered, so a total of eight fine-grained categories. We view the task of colorizing grayscale images of ethnic costumes as using grayscale image features to predict the values of the color features of the corresponding pixels. So the output of the generator model is set to two color channels $Y \in \mathbb{R}^{H \times W \times 2}$ in the CIE *Lab* color space. The general case is divided into two inputs for the discriminator model, real or fake images. We set a single channel of a gray image as one of the discriminator inputs, while the results of both color channels generated by the generator and the corresponding fine-grained semantic masks of the image are used as discriminator inputs. The three are sequentially concatenated as fake images. Meanwhile, the color image and the corresponding fine-grained semantic mask are also concatenated together as the real image.

3.1 | Colorization networks

Our network is divided into two parts, and the final objective function is defined as follows:

$$G^* = \arg \min_G \max_D \mathcal{L}_{\text{CGAN}}(G, D) + \lambda \mathcal{L}_{L1}(G), \quad (1)$$

where G is the generator network, D is the discriminator network, and λ is set to 0.5.

Figure 2 shows the specifics of our network structure in detail. First, the grayscale image with each fine-grained semantic mask is used as the generator's condition. Grayscale image information is a single-channel L in the CIE *Lab* color space. And the masks are fine-grained semantics that corresponds to various parts of the costume, representing various parts of the ethnic costume masks (the visualization effect is as shown in Figure 5). It is worth noting that the number of mask channels is the same as the number of fine-grained semantic categories in the data set. Each channel comprises 0 and 1 values that describe the semantic information of costumes in different regions. The image's content determines the number of fine-grained semantics contained in any image. Then, the grayscale images and fine-grained semantic masks are concatenated, and the results are fed into the generator. The bottleneck layer of the generator here uses Resnet Blocks and has jump connections. The generator finally predicts the values of both channels a and b under the CIE *Lab* color space based on the L channel. We then combine the features of the a and b channels output by the generator with the L channel, and finally output a colorized image based on the CIE *Lab* color space, also called fake images. The input of our discriminator in this paper is not only the fake image or real image but also their corresponding fine-grained semantic masks are combined with their corresponding fine-grained semantic masks as the discriminator's input.

3.2 | Semantic generator

Currently for grayscale image color tasks, the input and output images of the general model are low-resolution, with a common size of 256×256 pixels or 512×512 pixels.^{28–30} Our model uses high-definition resolution input and output images, and we have made improvements and

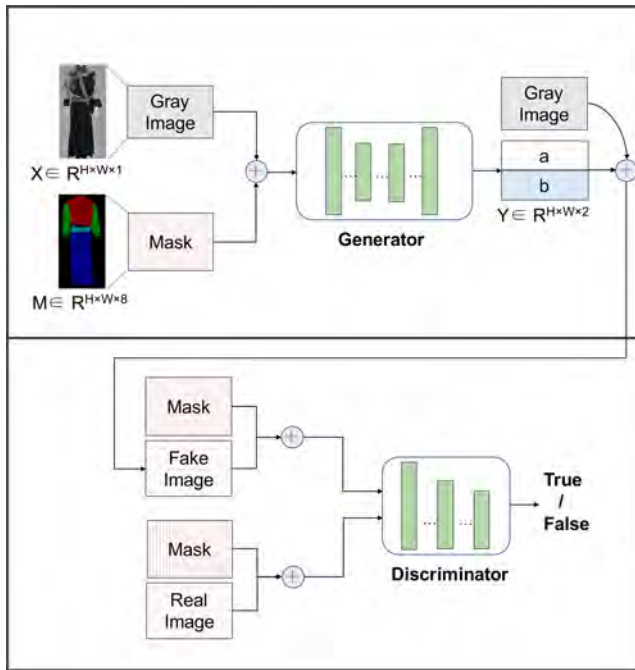


FIGURE 2 The framework of the proposed network architecture. The input of the generator consists of two parts: grayscale image and semantic mask. In CIE *Lab* color space, we convert the gray image into a single-channel gray information. The discriminator receives the real and false images generated by the generator, and the corresponding costume semantic mask is attached [Color figure can be viewed at wileyonlinelibrary.com]

innovations in the model structure.³¹ This paper designs the model with input and output images of size 512×1024 pixels up to 1024×2048 pixels. Some attempts have been made to generate high-definition color images more accurately from grayscale images and to preserve the details of ethnic costumes better. The network structure was designed by referring to the idea of Pix2Pix's generator using the coarse-to-fine approach, which allows the generator model to generate good-quality images. Then, an end-to-end deep GAN structure is used for the coloring task, and costume fine-grained level semantic information is used as an additional condition to assist the coloring task. As for the objective function, the generative adversarial loss function can make the generated images more realistic.^{32,33}

Figure 3 shows the details of our generator network structure. The generator network structure is based on the Pix2PixHD structure using residual block groups, and the network structure consists of three downsampling layers, nine residual block groups, and three upsampling layers. The downsampling layer convolution kernel parameter is set to 3, the stride parameter to 2, and the padding parameter to 1. For each downsampling layer that is used, the feature dimension is doubled. Simultaneously, we use the structure of residual blocks rather than skip connections, and the residual block is made up of two convolutional layers. After passing through two convolutional layers, the input features are connected to themselves. The residual block's input and output feature dimensions are the same. The most important component of the upsampling layer is the deconvolution layer. The deconvolution kernel parameter is set to 3, the stride parameter to 2, and the padding parameter to 1. Each upsampling layer feature dimension is changed to half of the original feature size. All layers are

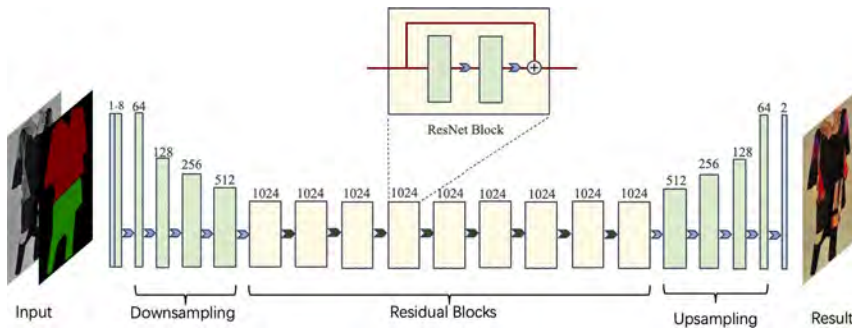


FIGURE 3 Generator network [Color figure can be viewed at wileyonlinelibrary.com]

added with batch normalization and the activation function is *ReLU*, except for the last layer where the activation function is *TanH*.

3.3 | Semantic discriminator

It is a great challenge for GANs to generate high-definition images, and discriminate between real and fake images. The discriminator plays an important role in GAN, and the discriminator can guide the generator to generate more realistic images. In this paper, a multiscale discriminator is used to discriminate high-definition images, while PatchGAN¹³ is used to discriminate high-definition images in a chunking manner. Multiscale discriminator discriminates the images at different scales, which is beneficial for the generator to consider the global consistency generating images fully and facilitates the generation of high-quality images. The design idea of PatchGAN is to discriminate the images in chunks, and each chunked image will output a discriminant result. Finally, the final result will be output by combining the discriminant results of all the chunks. Combining multiscale discriminator and PatchGAN's method can limit the local details of the generator-generated images, thus achieving a better image quality.

Figure 4 shows the details of our discriminator network structure. The discriminator network structure is mainly based on multiscale discriminator and PatchGAN. the final discriminator structure is obtained by combining the two. The discriminator in this paper has two scales and the number of chunks in each scale is 512. As Figure 4 shown, Lower scales are obtained by down-sampling from higher scales, and the downsampling is averaged pooling. The downsampling kernel parameter is set to 3, the stride parameter to 2, and the padding parameter to 1.

4 | EXPERIMENTS AND DISCUSSION

For a more comprehensive validation of our proposed grayscale image coloring method for ethnic costumes, our experiments consist of the following parts:

- By setting different input conditions for the generator network, we compare nonsemantic information with different levels of semantic information.
- For the input of the discriminator network, we compared the effect of the presence or absence of grayscale information channels on the coloring task.

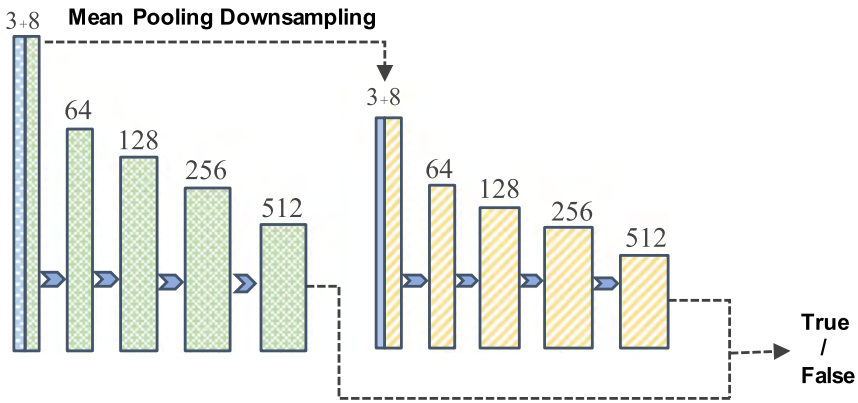


FIGURE 4 Discriminator network [Color figure can be viewed at wileyonlinelibrary.com]

- Compare the effect of the presence or absence of fine-grained semantic mask information as an input condition on the discriminator.
- The comparison experiments between four state-of-the-art grayscale image coloring algorithms of Iizuka et al.,³⁴ Larsson et al.,⁷ Antic,³⁵ Su et al.,¹⁸ and our algorithm are designed.

4.1 | Data sets

The main research of this paper is the task of ethnic costumes grayscale image coloring. Currently there is no publicly available data set for ethnic costumes. We constructed a costume data set consisting of four Chinese ethnic minority costumes, including Dai, Hani, Wa, and Yi. Each image in the data set is of high quality and has a uniform standard, with an aspect ratio of 2:1 and a resolution of up to 2048×1024 pixels. Our group collected and compiled all of the images in various ethnic museums and gathering places. The data set contains 340 costume images, and each costume set includes 3 or 4 images, including front, back, and side angles, totaling approximately 1200 images. As shown in Figure 5, each image was manually semantically annotated.³⁶ The semantics of the data set can be divided into eight categories, sleeve, coat, belt, skirt, pant, leg guard, accessories, and background, according to the characteristics of ethnic costumes. In total, more than 4000 fine-grained semantics were manually annotated in our data set (Figure 6).

4.2 | Evaluation metric

In this section, we use two universal evaluation metrics, including peak signal-to-noise ratio (PSNR) and structural consistency (SSIM), to evaluate the coloring quality of our method. Furthermore, the PSNR and SSIM are calculated separately for each of the three channels in RGB color space. The evaluation results are obtained by calculating the average of the three channels.

Given a size $M \times N$ of the original image I and noise image Y , the mean square error is defined as

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i, j) - Y(i, j)]^2. \quad (2)$$



FIGURE 5 Here, we present samples of ethnic costumes and their corresponding semantic images from four perspectives: front, left front, right front, and back [Color figure can be viewed at wileyonlinelibrary.com]



FIGURE 6 This picture is mainly divided into the image part and the label part. The display is divided into four groups. The image part includes three images. The original image, semantic and image combination graph, and semantic information graph are displayed from left to right. The tag section is the semantic tag that corresponds to the photo group [Color figure can be viewed at wileyonlinelibrary.com]

And then PSNR can be expressed as

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right), \quad (3)$$

where MAX_I^2 indicates the maximum possible pixel value of a current image, and the unit is dB. Note that, each pixel is represented by an 8-bit binary and MAX_I is 255, because the experimental result is an RGB three-channel image (Figure 7).

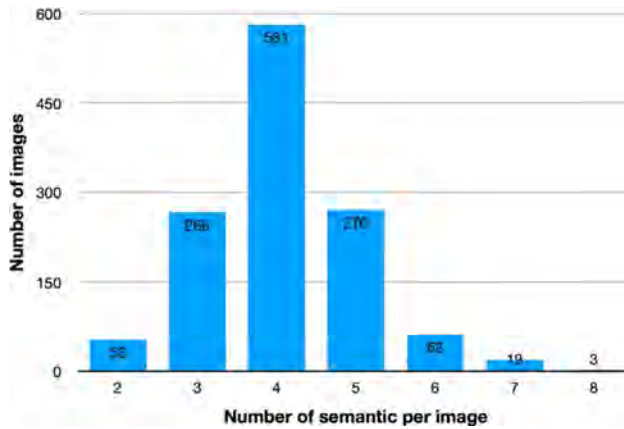


FIGURE 7 The distribution of semantic tags in a data set [Color figure can be viewed at wileyonlinelibrary.com]

In SSIM, image similarity is measured by brightness, contrast, and structure (see Equations 4–6):

$$l(X, Y) = \frac{2\mu_X\mu_Y + c_1}{\mu_X^2 + \mu_Y^2 + c_1}, \quad (4)$$

$$c(X, Y) = \frac{2\sigma_X\sigma_Y + c_2}{\sigma_X^2 + \sigma_Y^2 + c_2}, \quad (5)$$

$$s(X, Y) = \frac{\sigma_{XY} + c_3}{\sigma_X\sigma_Y + c_3}, \quad (6)$$

where μ_X and μ_Y represent the mean values of X and Y , σ_X and σ_Y represent the variance of X and Y , and σ_{XY} represents the covariance of X and Y . $c_1 = (k_1 * L)^2$ and $c_2 = (k_2 * L)^2$ are constants to avoid the situation where the formula is 0. In general, $k_1 = 0.01$, $k_2 = 0.03$, and $L = 255$. The SSIM is defined as

$$SSIM(X, Y) = l(X, Y) * c(X, Y) * s(X, Y). \quad (7)$$

The range of SSIM is $[0, 1]$. The larger the SSIM value is, the smaller the degree of image distortion will be.

4.3 | Experimental and setting

The experimental software configuration is as follows:

1. *Operating system*: Ubuntu16.04.
2. *Programming language*: Python3.7.6.
3. *Image processing library*: Opencv-python4.3.0.36.
4. *Deep learning framework*: Pytorch1.5.1.
5. *Other related libraries*: Cudnn, CUDA11.0, and numpy1.16.0.

The experimental hardware configuration is as follows.

1. *Central processing unit*: Intel Xeon Silver, 4210.
2. *Graphics processing unit (GPU)*: 2*NVIDIA GeForce RTX 2080 Ti.
3. *Memory*: 128 GB.

Our network structure is mainly divided into two parts: generator and discriminator. The inputs of these two parts, including grayscale image and semantic mask, have been improved. For λ in Equation (1), we set it as 0.5, and the resolution of the input and output images is 512×1024 . During the training process, our generator and discriminator use Adam Optimizer³⁷ with $\beta_1 = 0.99$ and $\beta_2 = 0.999$. Since the training process in Pix2Pix¹³ does not require a large amount of data, we train about 1200 images, run about 20,000 iterations and 200 epochs.

4.4 | Experimental comparison and analysis

We detailed conducted qualitative and quantitative experiments with our proposed model. In Section 4.4.1, we first conduct several ablation experiments. Set different conditional input levels to the generator, and the second set different levels of input to the discriminator. In Section 4.4.2, we compare our method with the mainstream grayscale image coloring methods.

4.4.1 | The semantic information of different levels of grayscale image colorization is compared as conditions

Different semantic levels of inputs may produce different coloring effects, so we conducted the following experiments. Under the same generator network model, different input conditions of the generator are first set, including nonsemantic input, single-instance-level semantic input. Set the discriminator's input without grayscale channel information, in other words, the discriminator receives only the two channels of information generated by the generator in the CIE *Lab* color space. The discriminator's input is then set to include the color information from the three channels but not the semantic mask information. Finally, we instruct the generator to accept fine-grained level semantic masks as a condition. As a condition for the input, the discriminator employs fine-grained level semantics.

In the architecture of GAN, the generator is responsible for generating images and the discriminator is responsible for evaluating the images generated by the generator. The two subnetworks are continuously trained alternately to update the parameters of each subnetwork, which eventually makes the generator create images with better quality. In this paper, the generator finally generates the values of the two image channels and then concatenates them with the values of the *L* channel to form the color images of the three color channels under the CIE *Lab* color space. There are two advantages for the generator not to generate the three-channel values under CIE *Lab* color space directly, the first is to reduce the number of parameters of the generator model and improve the training efficiency. The second is to avoid the loss of grayscale information brought by the forward propagation of the model to ensure the coloring effect. The input conditions of the discriminator include the color information of the three channels with the corresponding fine-grained level semantic masks. We also performed ablation experiments on the discriminator's input and we only fed the discriminator the color features of the two channels generated by the generator.

The color information from the three channels and the corresponding fine-grained level semantic masks for ethnic costumes are used as discriminator inputs in the proposed method. We set up the corresponding comparison experiments to test the efficacy of fine-grained level semantic masks as the discriminator condition. Under the same conditions, we add fine-grained level semantic masks to the discriminator input as the condition in one group and the color features of the three channels in the other.

We conducted some experiments, and the comparison results are shown in Table 1. From Table 1, you can see the impact of different input conditions on the final generator during the training phase. The first is that the generator sets different input conditions. We use semantic-free input and instance-level semantic input for comparison. Then we compare the discriminator using different input conditions, comparing no grayscale channel information as input with nonsemantic mask information as input. The last part is the setup of our fine-grained level semantic grayscale image coloring method, using the fine-grained level semantic masks as additional inputs to the generator. Similarly, the discriminator uses semantic masks as one of its inputs. We quantify the coloring effect by two evaluation metrics: PSNR and SSIM. With these two quantifiers, our method outperforms the conditions of the other settings.

4.4.2 | A comparison of several commonly used gray image coloring algorithms

We compare some common gray image coloring algorithms. The effect is shown in Figure 8. Since the input of other automatic colorization methods is different from ours, we additionally input fine-grained semantics of ethnic costumes. We use the model parameters trained by the author in other methods because it is impossible to systematically compare each method under the same conditions.

We compare four automatic coloring methods: Iizuka et al.,³⁴ Larsson et al.,⁷ Antic,³⁵ Su et al.,¹⁸ and comparison results are shown in Table 2. We also evaluate the results of colorization from two evaluation indicators: PSNR and SSIM. Likewise, our results have better performance.

In Figure 8, we show the results from some recently grayscale image colorization methods and our model. The first column shows the grayscale input image, Columns 2–5 show different automatic coloring methods, including Iizuka et al.,³⁴ Larsson et al.,⁷ Antic,³⁵ and Su et al.¹⁸ Column 6 shows the semantic input label, and Column 7 shows the result from our model, semantic label, and the effects of our method. The final column displays the actual image. The colorization effect of the cuff in the first row is superior to other methods, and local consistency

TABLE 1 Comparison of different methods under different input conditions

Method	PSNR	SSIM
Generator W/O semantic mask	25.1	0.904
Generator with semantic mask	26.28	0.912
Discriminator W/O gray channel	26.13	0.905
Discriminator W/O semantic mask	26.41	0.913
Ours	26.45	0.914

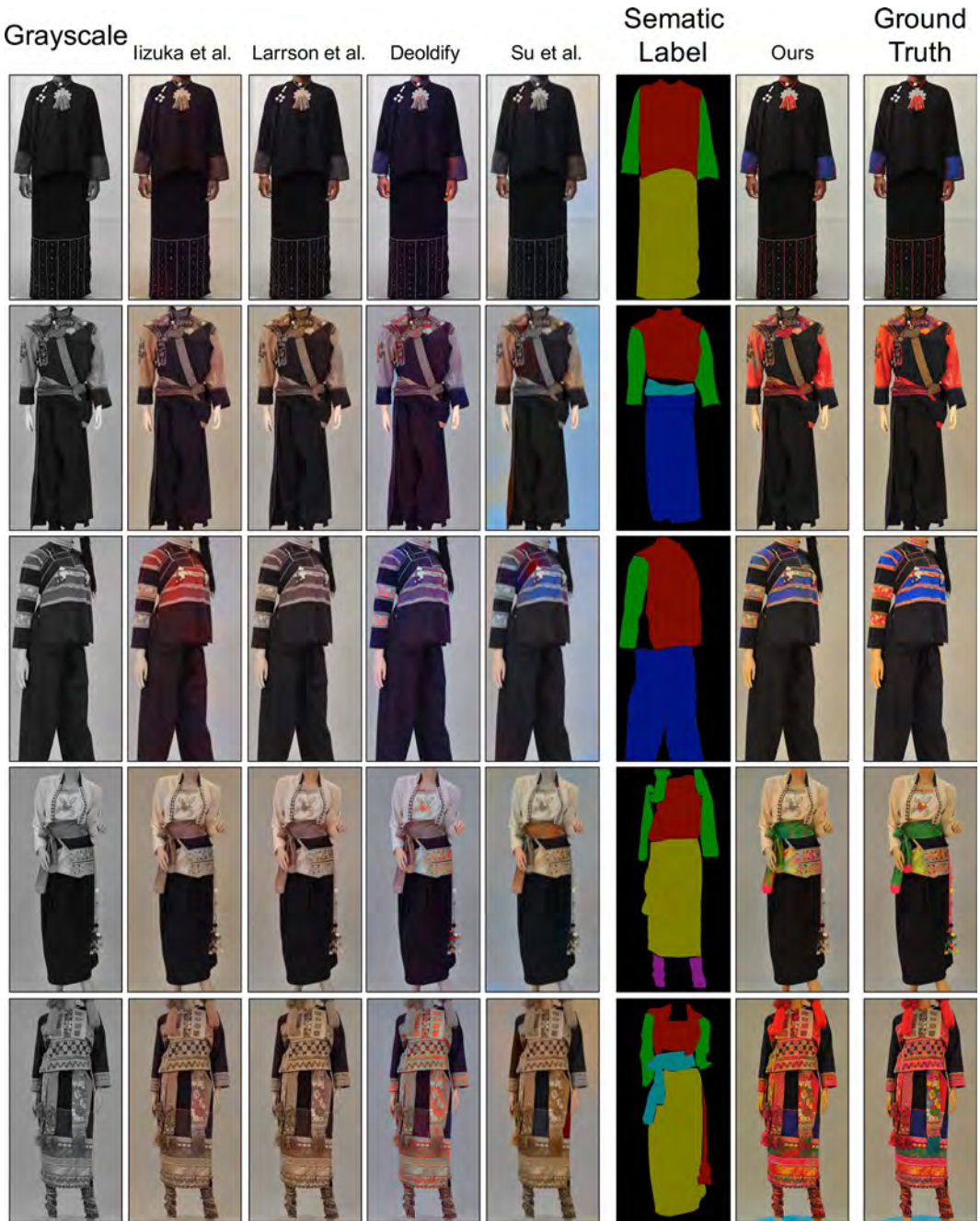


FIGURE 8 Visual comparison of different colorization methods [Color figure can be viewed at wileyonlinelibrary.com]

is satisfactory. The color feature details of ethnic costumes are more excellent in the other rows. In general, our coloring effect is more colorful and more accurate.

In Figure 9, we show the effects of colorization method on different grayscale images of ethnic costumes. Our method has a better performance for the grayscale image of ethnic costumes, and can show the colorful costume details.



FIGURE 9 Several sets of colorization examples of our proposed model [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Comparison with other coloring methods

Method	PSNR	SSIM
Iizuka et al. ³⁴	21.72	0.868
Larsson et al. ⁷	21.84	0.876
Antic ³⁵	20.02	0.862
Su et al. ¹⁸	20.20	0.913
Ours	26.45	0.914

5 | CONCLUSION AND FUTURE WORK

We propose a coloring method for ethnic costumes grayscale images using fine-grained level semantic information as a condition and construct and label an image data set consisting of four Chinese minority costumes. Then, we demonstrate that our method can better colorize ethnic costumes with complex color distributions through experimental comparison and analysis.

We use the fine-grained level semantics of various regions of ethnic costumes as additional conditions for coloring, and these additional fine-grained level semantics are derived from manually annotated data sets. A lot of current work investigates how to automate semantic segmentation to obtain fine-grained level semantic information automatically. Therefore, we

can first obtain semantic information through a semantic segmentation model and use this information for the automatic coloring process.

ACKNOWLEDGMENTS

This study is supported by the National Natural Science Foundation of China under Grant No. 61862068, Yunnan Innovation Team of Education Informatization for Nationalities, and Yunnan Expert Workstation of Xiaochun Cao.

ORCID

Di Wu  <https://orcid.org/0000-0003-2373-2063>

Jianhou Gan  <https://orcid.org/0000-0002-1287-857X>

Juxiang Zhou  <https://orcid.org/0000-0003-2693-2204>

Jun Wang  <https://orcid.org/0000-0001-9878-4664>

Wei Gao  <https://orcid.org/0000-0001-7963-3502>

REFERENCES

1. Cheng Z, Yang Q, Sheng B. Deep colorization. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2015:415-423.
2. Levin A, Lischinski D, Weiss Y. Colorization using optimization. In: *ACM SIGGRAPH 2004 Papers*; 2004: 689-694.
3. Zoran D, Weiss Y. From learning models of natural image patches to whole image restoration. In: *2011 International Conference on Computer Vision*, November 2011. IEEE; 2011:479-486.
4. Zhang X, Wandell BA. A spatial extension of CIELAB for digital color-image reproduction. *J Soc Inf Inf Dis*. 7;5(1):61-63.
5. Charpiat G, Hofmann M, Schölkopf B. Automatic image colorization via multimodal predictions. In: *European Conference on Computer Vision*. Vol 5304. Springer; 2008:126-139.
6. Guadarrama S, Dahl R, Bieber D, Norouzi M, Shlens J, Murphy K. Pixcolor: pixel recursive colorization. arXiv preprint. 2017: arXiv:1705.07208.
7. Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization. In: *European Conference on Computer Vision*, October 2016. Springer; 2016:577-593.
8. Zhang R, Isola P, Efros AA. Colorful image colorization. In: *European Conference on Computer Vision*, October 2016. Springer; 2016:649-666.
9. Furusawa C, Hiroshiba K, Ogaki K, Odagiri Y. Comicolorization: semi-automatic manga colorization. In: *SIGGRAPH Asia 2017 Technical Briefs*; 2017:1-4.
10. Wan S, Xia Y, Qi L, Yang YH, Atiquzzaman M. Automated colorization of a grayscale image with seed points propagation. *IEEE Trans Multimedia*. 2020;22(7):1756-1768.
11. Zou C, Mo H, Gao C, Du R, Fu H. Language-based colorization of scene sketches. *ACM Trans Graphics (TOG)*. 2019;38(6):1-16.
12. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems 27*; 2014:2672-2680.
13. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017:1125-1134.
14. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017:2223-2232.
15. Kim T, Cha M, Kim H, Lee JK, Kim J. Learning to discover cross-domain relations with generative adversarial networks. In: *International Conference on Machine Learning*, July 2017. PMLR; 1857-1865.
16. Hensman P, Aizawa K. cGAN-based manga colorization using a single training image. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, November 2017. Vol 3. IEEE; 2017:72-77.

17. Vitoria P, Raad L, Ballester C. ChromaGAN: adversarial picture colorization with semantic class distribution. In: *The IEEE Winter Conference on Applications of Computer Vision*; 2020:2445-2454.
18. Su JW, Chu HK, Huang JB. Instance-aware image colorization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2020:7968-7977.
19. Zhao J, Han J, Shao L, Snoek CG. Pixelated semantic colorization. *Int J Comput Vision*. 2020;128(4): 818-834.
20. Anwar S, Tahir M, Li C, Mian A, Khan FS, Muzaffar AW. Image colorization: a survey and dataset. *arXiv e-prints*. 2020;arXiv:2008.10774
21. Zhang R, Zhu JY, Isola P, et al. Real-time user-guided image colorization with learned deep priors. *arXiv preprint*. 2017;arXiv:1705.02999.
22. Messaoud S, Forsyth D, Schwing AG. Structural consistency and controllability for diverse colorization. In: *Proceedings of the European Conference on Computer Vision (ECCV)*; 2018:596-612.
23. Yoo S, Bahng H, Chung S, Lee J, Chang J, Choo J. Coloring with limited data: Few-shot colorization via memory augmented networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2019:11283-11292.
24. Shamsabadi AS, Sanchez-Matilla R, Cavallaro A. Colorfool: semantic adversarial colorization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2020:1151-1160.
25. Zhao W, Zhou J, Xu T. National costume image retrieval based on integrated region matching. In: *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, August 2017. IEEE; 2017: 172-177.
26. Zhenrong D, Shanjin B, Fuxin M, Wenming H, Xiaonan L. Zhuang national costume images generation method based on deep convolutional generative adversarial network. In: *2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI)*, June 2019. IEEE; 2019:314-319.
27. Liu B, Gan J, Wen B, LiuFu Y, Gao W. An automatic coloring method for ethnic costume sketches based on generative adversarial networks. *Appl Soft Comput*. 2021;98:106786.
28. Lei C, Chen Q. Fully automatic video colorization with self-regularization and diversity. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2019:3753-3761.
29. Li B, Lai YK, John M, Rosin PL. Automatic example-based image colorization using location-aware cross-scale matching. *IEEE Trans Image Process*. 2019;28(9):4606-4619.
30. Fang F, Wang T, Zeng T, Zhang G. A superpixel-based variational model for image colorization. *IEEE Trans Visualization Comput Graphics*. 2019;26(10):2931-2943.
31. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2019:4401-4410.
32. Mirza M, Osindero S. Conditional generative adversarial nets. *arXiv preprint*. 2014;arXiv:1411.1784.
33. Xiao C, Li B, Zhu JY, He W, Liu M, Song D. Generating adversarial examples with adversarial networks. *arXiv preprint*. arXiv:1801.02610.
34. 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 Graphics (ToG)*. 2016; 35(4):1-11.
35. Antic J. A deep learning based project for colorizing and restoring old images. GitHub; 2019. <https://github.com/charlespwd/project-title>
36. Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: a database and web-based tool for image annotation. *Int J Comput Vision*. 2008;77(1-3):157-173.
37. Kingma DP, Ba JA. A method for stochastic optimization. arXiv 2014. *arXiv preprint*. 2014;arXiv:1412.6980.

How to cite this article: Wu D, Gan J, Zhou J, Wang J, Gao W. Fine-grained semantic ethnic costume high-resolution image colorization with conditional GAN. *Int J Intell Syst*. 2021;1-17. doi:10.1002/int.22726