ORIGINAL ARTICLE



Image recoloring for color vision deficiency compensation using Swin transformer

Ligeng Chen¹ · Zhenyang Zhu² · Wangkang Huang¹ · Kentaro Go² · Xiaodiao Chen³ · Xiaoyang Mao²

Received: 15 June 2023 / Accepted: 7 December 2023 / Published online: 18 January 2024 $\ensuremath{\mathbb{C}}$ The Author(s) 2024

Abstract

People with color vision deficiency (CVD) have difficulty in distinguishing differences between colors. To compensate for the loss of color contrast experienced by CVD individuals, a lot of image recoloring approaches have been proposed. However, the state-of-the-art methods suffer from the failures of simultaneously enhancing color contrast and preserving naturalness of colors [without reducing the Quality of Vision (QOV)], high computational cost, etc. In this paper, we propose an image recoloring method using deep neural network, whose loss function takes into consideration the naturalness and contrast, and the network is trained in an unsupervised manner. Moreover, Swin transformer layer, which has long-range dependency mechanism, is adopted in the proposed method. At the same time, a dataset, which contains confusing color pairs to CVD individuals, is newly collected in this study. To evaluate the performance of the proposed method, quantitative and subjective experiments have been conducted. The experimental results showed that the proposed method is competitive to the state-of-the-art methods in contrast enhancement and naturalness preservation and has a real-time advantage. The code and model will be made available at https://github.com/Ligeng-c/CVD_swin.

Keywords Color vision deficiency \cdot Unsupervised \cdot Swin transformer \cdot Recoloring \cdot Deep neural network \cdot Contrast \cdot Naturalness

⊠ Xiaoyang Mao mao@yamanashi.ac.jp

> Ligeng Chen ligengchen422@gmail.com

Zhenyang Zhu zzhu@yamanashi.ac.jp

Wangkang Huang huang10669668281@gmail.com

Kentaro Go go@yamanashi.ac.jp

Xiaodiao Chen xiaodiao@hdu.edu.cn

- ¹ Graduate School of Engineering, University of Yamanashi, 4-4-37 Takeda, Kofu, Yamanashi 400-8510, Japan
- ² Department of Computer Science and Engineering, University of Yamanashi, 4-4-37 Takeda, Kofu, Yamanashi 400-8510, Japan
- ³ School of Computer Science, Hangzhou Dianzi University, No.1, the No.2 Baiyang Street, Hangzhou 310018, Zhejiang, China

1 Introduction

There are three kinds of cone cells in human retinas, the L, M, and S cones, and they are sensitive to long-, medium-, and short-wavelength visible light, respectively, enabling people to perceive distinct colors. Anomalies in cone cells lead to color vision deficiency (CVD) in individuals. Specifically, when an anomaly occurs in the L, M, or S cones, CVD is classified as protanopia, deuteranopia, or tritanopia, respectively. It is estimated that approximately 8% of men and 0.5% of women in the world are affected by various degrees of CVD [1, 2]. People with CVD may experience inconvenience in their lives, such as the inability to read red characters written on a black background, due to the reduced capability to discriminate colors. CVD is an inherited deficiency for which there is currently no cure.

Over the years, numerous image processing approaches [3–7] have been proposed to support people with CVD. In [8], Ribeiro et al. summarized CVD compensation studies conducted in the 2 decades before 2019. Further, Zhu et al.

[9] surveyed the latest image recoloring studies, as well as quantitative and subjective evaluation metrics for CVD compensation methods. In recent years, image recoloring approaches have become the mainstream for CVD compensation, aimed at enhancing the color contrast perceived by CVD individuals by changing a part of the colors in the images. Considering CVD individuals may feel uncomfortable if significant changes are applied to the original color appearance, which can reduce quality of life (QOL), state-of-the-art studies [10-16] further considered naturalness preservation, which put constraints on deviations from the original image colors. The state-of-the-art studies [12-14] proposed recoloring images using optimization models, which take contradictory constraints, i.e., contrast enhancement and naturalness preservation, into a unified energy function and achieve balanced recoloring results; however, the GPU-based implementation, which delivers the best performance, still requires approximately 7 s to process an image sized 256×256 pixels [14]. In other words, the poor time efficiency limits their usage in real applications.

Recently, deep neural network (DNN)-based machine learning techniques have achieved remarkable success in the image processing and image synthesis fields. Specifically, Isola et al. proposed a Pix2Pix network [17] to transfer images from a source domain to a target domain in real time, and it showed the potential of using a DNN to recolor images in real time. In [18], Li et al. demonstrated the feasibility of utilizing Pix2Pix to generate compensation images for CVD by training the network using images generated with their own recoloring algorithms. Because their recoloring algorithm does not consider naturalness preservation for CVD and because the Pix2pix model used was based on a convolutional neural network (CNN), which has a limited receptive field, leading to a poor ability to learn explicit global and long-range information, the effectiveness of the resulting images is limited. Hu et al. [19] proposed a method to decompose an original image into 27 basis images and then to use a CNN to predict the coefficients for blending the basis images into image pairs for use with stereo displays, enabling content sharing between audiences with CVD and with normal vision. However, Hu et al. [19] were also subject to CNN limitations and did not consider naturalness preservation for CVD either.

In this paper, we propose a new DNN-based image recoloring method for CVD compensation. The proposed method aims to enhance contrast and preserve naturalness simultaneously in recolored images. The architecture of the proposed network is similar to that of U-Net [20], which is comprised of an encoder and a decoder. To improve the performance of U-Net, we replace its CNN layers with Swin transformer blocks [21]. Transformers, also known as

self-attention (SA) mechanisms, have gained significant attention in the fields of natural language processing and computer vision [22, 23]. The Swin transformer [21] has the same performance as a standard transformer, but it requires a much lower computational cost. To overcome the limitation caused by using ground-truth images generated by existing methods, we design an unsupervised training manner to guide the training of the proposed network; concretely, the training phase only requires input images and is guided by a newly designed loss function that considers both contrast enhancement and naturalness preservation. To enhance the generalization capability of the trained model, we built an image dataset containing color pairs that are confusing to CVD individuals, and we use these images as the input to the proposed model during the training stage. The dataset consists of 131,000 color images, including 31,000 photographs of natural scenes and 100,000 artificially created images. To evaluate the performance of the proposed method, we conducted quantitative and subjective experiments. The experimental results show that our method can generate recoloring results competitive to those of state-of-the-art methods but executed in real time. Our contributions can be summarized as follows:

- A novel DNN-based image recoloring method for CVD compensation. By using the U-net structure of Swin transformer blocks, the method succeeded in capturing both local and long-range color dependency.
- A novel unsupervised training approach guided by a newly designed loss function considering both global and local contrast enhancement, as well as naturalness preservation.
- A new dataset that can be used to train a DNN model in an unsupervised manner to generate CVD compensation images. The database consists of photographs and artificial images with color pairs that are confusing to CVD individuals.

2 Related work

2.1 Conventional recoloring method for CVD

With the popularization of digital devices, recoloring algorithms are dominating the approaches to CVD compensation. As mentioned above, Ribeiro et al. [8] published a survey paper regarding image processing approaches for CVD compensation, covering studies conducted in the most recent 2 decades, including 59 papers on image recoloring approaches. Further, Zhu et al. [9] conducted survey research focusing on image recoloring approaches, adding 17 studies on the topic. Early studies focused on enhancing color contrast, while preserving naturalness and enhancing contrast simultaneously became the trend in the most recent decade.

Iaccarino et al. [4] enhanced the color contrast between text and backgrounds on web pages by simulating recolored colors falling in two different half-planes. Jefferson et al. [24] developed an interface to allow CVD users to adjust a single parameter of a recoloring algorithm. To maximize the contrast between colors on the gamut of dichromacy, Machado et al. [6] proposed first projecting all colors in the input image onto a two-dimensional plane, where the distribution of projected colors can be maximally dispersed, and then rotating the plane to match the CVD gamut. Lin et al. [7] proposed a color-warping method that utilizes the orientation of the eigenvectors of the CVD simulation results for recoloring images in the opponent color space [25]. These studies did not consider naturalness preservation, thus changing the color appearance of the original image significantly, which may make CVD users feel uncomfortable. Therefore, later researchers focused on color appearance preservation, that is, naturalness preservation. Kuhn et al. [26] enhanced color contrast and preserved naturalness based on mass-spring optimization, but the quality of the recoloring result depends on the quality of the quantization. In [27], Huang et al. proposed to extract key colors from an input image and to recolor them to enhance their distances, but not too far from the original image. Hassan et al. [10] compensated for the contrast loss by increasing the blue component of confusing colors, because CVD individuals with red-green deficiency, which is most individuals, have difficulty discriminating red and green, but they can perceive blue well. To achieve sufficient contrast enhancement, in [11], they extended their work so users could exaggerate the blue channel. However, contrast loss occurs in the recolored images due to the saturation of the blue channels. Zhu et al. [14] proposed an optimization model by using an object function considering both contrast enhancement and naturalness preservation, and they demonstrated that their method achieves the best compensation effect among all existing methods through a subjective experiment involving volunteers with CVD. However, their method is time-consuming. To reduce the computation time, they proposed to extract dominant colors from the original image and to recolor the extracted dominant colors using an optimization model. Finally, the recolored dominant colors are propagated to the whole image. Despite the acceleration, their method still requires more than 7 s to recolor an image sized 256×256 pixels on a desktop with an Intel(R) Core (TM) i7-9800X CPU@3.80GHz, 64-GB memory, and GeForce RTX 2080 Ti.

To accelerate processing, Wang et al. [15] proposed a fast recoloring method by computing a single optimal map

from the 3D color space to the CVD color gamut. Huang et al. [16] fitted a curved surface to represent the gamut of dichromacy in Lab space and considered a luminance channel in their compensation model. However, their methods still require a few seconds to generate an image. Ebelin et al. [28] introduced a real-time algorithm to enhance the visual experience for color vision-deficient observers by recoloring images. The proposed algorithm prioritizes temporal stability and luminance preservation. However, their method falls short in addressing the issue of naturalness for individuals with CVD.

2.2 Recoloring method for CVD individuals based on DNNs

Because image recoloring can be regarded as transferring an image from a source domain to a target domain, it is possible to apply a DNN to image recoloring for CVD compensation. For instance, Isola et al. [17] proposed a network, Pix2Pix, based on conditional GANs (generative adversarial networks) to carry out the image-to-image translation task. A standard GAN is comprised of a generator and a discriminator. Taking white noise as the input, the generator aims to generate fake images whose appearance is similar to real images, while the discriminator tries to distinguish real and fake images. Setting the input image as a condition, Pix2Pix maps the image from the source domain to the target domain. In [18], Li et al. trained the Pix2Pix network with thousands of ground-truth image pairs generated using their recoloring algorithm. However, their recoloring algorithm did not consider naturalness preservation; thus, the network has no ability to preserve naturalness. In [19], Hu et al. proposed a model for sharing the same contents to normal vision and CVD audiences. For CVD audiences who wear stereo glasses, the model generates a pair of contrast-enhanced images from the input image for left and right eyes, respectively. For normal color vision audiences who do not wear stereo glasses, they perceive the blending of the paired images. For generating the image pair, the original image is first decomposed into 27 basis images; then, their CNN-based model predicts a set of coefficients; finally, the output image pair is computed as the linear combination of the basis images with the predicted coefficients. Though the blending of the image pair, which synthesizes the image for people with normal vision, has been constrained to be similar to the input image, Hu et al. [19] did not consider naturalness preservation for CVD. Besides, both methods used a CNN as the backbone for feature extraction in their network model, which has a limited receptive field leading to a poor ability to learn explicit global and long-range information; hence, the effectiveness of the resulting images is limited. Jiang et al. [29] proposed a method based on

StyleGAN-ada[30] for generating compensation images across various degrees through latent traversal. However, the limitation lies in the fact that their method relies on latent codes as input, making it challenging to apply in real-time to real images.

2.3 Vision transformer

Driven by transformers' success in the natural language processing domain [22], which process all tokens at once and calculate relationships between them, more researchers are paying attention to the use of transformer-based models in the computer vision domain. The pioneering vision transformer (ViT) in [23] flattens 2D image patches in a vector and then feeds it to the transformer. It achieved an impressive performance in an image recognition task, but it requires a large-scale training dataset, such as JFT-300 M. The Swin transformer [21], an efficient and effective hierarchical vision transformer, has linear computational complexity from adopting a shifted windows scheme. With the hierarchical architecture, it has been adopted as a vision backbone, and it achieved a state-of-the-art performance in various vision tasks, including image classification, object detection, and semantic segmentation. In this work, we adopt a Swin transformer block as a basic unit to build an encoder-decoder model for generating compensation images for CVD people.

3 Proposed method

3.1 Overview

As illustrated in Fig. 1, given an input image I that contains color pairs confusing to CVD individuals, the proposed model can output a compensated image with enhanced contrast but still natural to CVD individuals. The proposed model uses an auto-encoder framework whose structure is similar to that of U-Net [20]. It is comprised of two parts: (1) an encoder and (2) a decoder; the encoder extracts features from the input image, while the decoder generates a recolored result from the extracted features. Particularly, the Swin transformer is adopted as a feature extractor.

The input image is first fed to the encoder for hierarchical feature extraction, and then the compensated image is generated by the decoder by referring to the different feature layers of the encoder using skip connection. In [20], the encoder adopted CNN layers for feature extraction. In this study, we construct our encoder with patch partitioning and patch embedding and with a series of Swin transformer blocks and patch merging. The decoder consists of Swin transformer blocks, up-sample layer patch expanding, and convolution block layers.



Fig. 1 Architecture of the proposed model: encoder, decoder, and skip connections

More details of the proposed model will be introduced in Sect. 3.2. In Sect. 3.3, we will introduce the unsupervised training approach, as well as the loss function, which aims used to guide mode training in an unsupervised manner.

3.2 Network design

3.2.1 Swin transformer block

Taking the original image as the input, the patch partition of the encoder divides the input image into small patches. As shown in Fig. 2, the area partitioned by the gray square denotes a patch, and each patch can be regarded as a token. The big squares in brown represent local windows, and tokens exchange information with all other tokens in the



Fig. 2 An illustration of the W-MSA and SW-MSA modules. In layer l (left), self-attention is computed within each regular window. In the next layer l + 1 (right), the window partitioning is shifted, resulting in new windows. Within each new window, the self-attention computation leads to connections with the previous windows

same window. The standard transformer contains only one big window, which includes all tokens, and each token must be compared with all other tokens using the multihead self-attention (MSA) module; thus, the standard transformer is time-consuming. In [21], Liu et al. proposed dividing the whole image into several windows; thus, the MSA operation can be conducted within each window (Fig. 2[layer l]). Such a module is called window-based MSA (W-MSA). To enable a token to exchange information with tokens in other windows, when it proceeds to the next layer, the windows are shifted, as illustrated in Fig. 2[layer l + 1]; that is, the Swin transformer block of the next layer uses the shifted W-MSA (SW-MSA) module from the second layer. Because the number of token pairs in each layer is reduced significantly, the computational cost of the Swin transformer is much smaller than that of a standard transformer. Besides the W-MSA and SW-MSA modules, a token must also proceed through the Layer-Norm (LN) module, residual connections, and multi-layer perceptron (MLP) modules. As illustrated in Fig. 3, the l layer takes the output of the previous layer, z^{l-1} , as the input, and the input proceeds through the LN, W-MSA, LN, and MLP modules in sequence, while the output of the l layer, z^l , proceeds through the SW-MSA instead of the W-MSA module. Such a procedure can be represented using a formula as follows:

where \hat{z}^l and z^l denote the output features of the (S)W-MSA module and the MLP module for Swin transformer block *l*, respectively.

3.2.2 Encoder

As mentioned in Sect. 3.2.1, the input image is divided into patches, and each patch is regarded as a token. In this study, the input image *I* is supposed to consist of $H \times W$ pixels, and it is equally divided into non-overlapping pixel



Fig. 3 The architecture of two Swin transformer blocks (notation presented with Eq. 1)

patches. Then, the patch embedding layer converts the pixel patches into tokens, which is followed by a series of Swin transformer blocks. By going through the first layer of Swin transformer blocks, feature representations can be obtained from the patch tokens. Then, the patch merging layer reduces the feature resolution while increasing feature dimensions, which produces a hierarchical representation. Then, the downsized feature representations are further processed by the other two layers of the Swin transformer blocks and patch merging layers; as a result, hierarchical feature representations of the input image can be obtained.

In this study, each pixel patch can be defined as $P \in \mathbb{R}^{4 \times 4 \times 3}$, and each token converted by patch embedding can be denoted as $v \in \mathbb{R}^{1 \times 1 \times C}(C$ is set to 96). In the patch merging layer, the feature resolution is down-sampled to a quarter of the input size, and the output feature dimension is enlarged to two times the values of the input features. In summary, the encoder has three layers of Swin transformer blocks, which have 8, 4, and 4 blocks, respectively. In addition, the input resolutions of three Swin transformer layers are $H/4 \times W/4$, $H/8 \times W/8$, and $H/16 \times W/16$, and their dimensions are C, 2C, and 4C, respectively.

3.2.3 Decoder

Given the hierarchical feature representations extracted by the encoder, our decoder outputs a recolored image of the same resolution as the input image for CVD individuals. The structure of the decoder is shown in Fig. 1. In our decoder, two layers of Swin transformer blocks are adopted to reconstruct the hierarchical features. The first layer of Swin transformer blocks (at the lower part of the decoder shown in Fig. 1) takes the final feature from the encoder as the input; the output is then fed into an up-sampling layer, which enlarges the feature resolution and reduces the feature dimension. Before the enlarged feature is fed into the upper layer of Swin transformer blocks, it is combined with the middle layer feature of the Swin transformer blocks in the encoder through skip connections. The output feature of the upper layer of Swin transformer blocks is further enlarged by the second up-sampling layer. Combined with the token produced by the patch embedding layer in the encoder, the enlarged feature is fed into the patch expanding layer, and the resulting image can be obtained through a final CNN block, which reduces the number of channels to three.

In summary, our decoder has two layers of Swin transformer blocks, both of which consist of four blocks, respectively. In addition, the resolutions of the output feature maps of the up-sampling layers are $H/8 \times W/8$ and

 $H/4 \times W/4$, and their dimensions are 2C and C, respectively.

3.3 Loss function

In this study, the training of the proposed model is guided by a loss function only; in other words, no ground-truth results are used to train the model. Considering the two goals of the proposed model, that is, contrast enhancement and naturalness preservation, we expect the contrast in the CVD simulation for the resulting image, O, can be enhanced to the same level as that in the input image, I, while changes to the color appearance are minimized. To do so, we introduce contrast and naturalness terms into our loss function. At the same time, we further divide the contrast terms into global and local, given that both can be reduced due to CVD.

To calculate the contrast loss, the distance between CVD-simulated colors on two pixels in *O* is compared with that between the colors of the corresponding pixels in *I*. The contrast loss is calculated in the lab color space, considering that lab color space is more consistent with the human color perception system, and it can be represented as follows:

$$CL(x, y) = ||\widehat{c}'_x - \widehat{c}'_y| - |c_x - c_y||$$

$$\widehat{c}'_x = CVD(c'_x), \widehat{c}'_y = CVD(c'_y)$$
(2)

where c_x, c_y, c'_x, c'_y denote the colors of the homologous pixels x, y in *I*, *O*, respectively; $CVD(\cdot)$ stands for the CVD simulation for a color and $|\cdot|$ represents the L1 norm of a vector. We adopt the model proposed by [31] for CVD simulation. Because the model in [31] can simulate varying degrees of CVD, we adopt the most severe degree in this study for the dichromacy simulation.

For the loss of local contrast, each pixel x in the recolored image O is required to be compared with a set of neighboring pixels of x. The local contrast loss term is defined as follows:

$$L_l = \sum_{x=1}^{N} \sum_{y \in \omega_x} \frac{\operatorname{CL}(x, y)}{||\omega_x||}$$
(3)

where ω_x stands for a set of pixels in a small window, which is centered on pixel x, and ω_x represents the number of pixels in ω_x ; N is the number of pixels in an image. In this study, the window size for local contrast loss is set to 11×11 .

For global contrast, the contrast loss of any pixel pair $\langle x, y \rangle$ randomly selected from the image for evaluation is calculated as follows:

$$L_g = \frac{1}{||\omega||} \sum_{\langle x, y \rangle \in \omega} \operatorname{CL}(x, y)$$
(4)

where ω stands for a set of randomly selected pixel pairs, and ω is the number of selected pixel pairs. ω is set to 3000 according to a validation experiment conducted in this study, which will be discussed in Sect. 5.3. As a result, the integrated contrast loss term is defined as follows:

$$L_c = \beta L_g + (2 - \beta) L_l \tag{5}$$

where β is the weight of the global contrast, we set the $\beta = 1$ in the proposed method, which will be discussed in Sect. 6.4. To preserve the naturalness of the original image, the recolored image is required to be as similar to the original image as possible. To measure the similarity between the recolored and original images, we adopt the structural similarity (SSIM) metric [32], which differs from the color difference metric used in [14], as SSIM is known to be able to measure the perceptual similarity of images.

Given the pixels x, y in the test and reference images, the SSIM metric is calculated among local windows X, Y centered on x, y, respectively. The formula of SSIM is written as follows:

$$SSIM(X, Y) = \frac{(2u_x u_y + c_1)(2\sigma_{xy} + c_2)}{(u_x^2 + u_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$
(6)

where u_x and σ_x are the mean value and standard deviation of the pixels in X, and u_y and σ_y are the mean value and standard deviation of the pixels in Y; σ_{xy} is the covariance of x and y, c_1 and c_2 are small values for avoiding zero division set to be 0.0001 and 0.0009, respectively. In this study, the naturalness loss term is defined as follows:

$$L_n = 1 - \text{SSIM}(O, I) \tag{7}$$

Finally, the loss function for guiding training is defined as follows:

$$L = \alpha L_n + (1 - \alpha) L_c \tag{8}$$

where α denotes the parameter to control the weight of naturalness preservation over the contrast enhancement in the recolored image. The higher the value of α , the more natural the generated image will be. In this study, to clarify the effect of α , we trained various models with different α values, which will be elaborated in Sect. 4.

4 Dataset collection

To improve the capability of the proposed model to enhance the contrast between CVD-indistinguishable color pairs, in this study, we created a new dataset consisting of pictures of natural scenes and artificial images containing CVD-confusing colors.

4.1 Natural image

We first collected images from the Places365-Standard dataset [33]. However, not all images in the dataset [33] contain colors confusing to CVD individuals. To improve the training efficiency, we filter out images in which CVD individuals have no difficulty discriminating colors. To do so, we assess the degree of contrast loss in the CVD simulation of the original image using the color contrast-preserving ratio (CCPR) metric introduced in [34], which is calculated as follows:

$$\frac{||\langle \hat{x}, \hat{y} \rangle| \langle x, y \rangle \in \Phi \text{ and } D(\hat{x}, \hat{y}) > \tau||}{||\Phi||}$$
(9)

where Φ represents a set of pixel pairs in the original image; for an arbitrary pixel pair $\langle x, y \rangle$ in Φ , the contrast D(x, y) is greater than a threshold τ . $\langle x, y \rangle$ and $\langle \hat{x}, \hat{y} \rangle$ indicate a homologous pixel pair in I and the CVD-simulated result of I; $||\Phi||$ stands for the number of pixel pairs in this Φ set. In this paper, the contrast D between a pixel pair refers to the distance in the Lab color space, and the threshold τ is empirically set to 6.

The lower the CCPR value, the severer the degree of contrast loss. In our implementation, we adopted images whose CCPR value is smaller than 0.8 in our dataset. As a result, we collected 31,000 natural images from the Places365-Standard dataset.

4.2 Artificial image

To make the proposed network more general and to learn how to recolor the most confusing colors, we synthesized artificial images to enrich the dataset. If the distance of a color pair in the Lab color space is greater than the threshold C_{ν} , while the distance of its CVD simulation result is smaller than the threshold C_{ν} , then this color pair is called a CVD-confusing color pair. And the threshold C_{v} is empirically set to 2.3, same as [35]. To obtain confusing color pairs, we traversed all colors in the RGB color space at a step length of 8 for each channel and then transferred the colors to the Lab color space for evaluation. Finally, we collected about 770,000 CVD-confusing color pairs. Then, we randomly selected 4 to 10 pairs of confusing colors and put them on a white canvas; finally, we filled the whole image using the nearest-neighbor interpolation method. Examples of artificial images are shown in Fig. 4a, c, and their CVD simulation results are shown in Fig. 4b, d, respectively. In total, we generated 100,000 artificial images to enrich our training dataset.



Fig. 4 Examples of artificially generated images for our dataset. **a** and **c** are artificial images; **b** and **d** are the CVD simulations of **a** and **c**, respectively

5 Implementation and validation

5.1 Implementation details

In this study, our model was implemented and trained on a PC with an Intel(R) Core(TM) i7-9800X CPU@3.80GHz, 64-GB memory, and GeForce RTX 2080 Ti. Before the input image are fed into the neural network, it will be resized to 256×256 , and the patch size is set to 4. For the training phase, the Adam optimizer was adopted, whose hyperparameters were set as the learning rate lr = 0.0002, $\beta 1 = 0.5$, and $\beta 2 = 0.999$. To explore the effect of the weight α for the naturalness term, we trained our model with three different naturalness weights $\alpha = 0.25, 0.5, and$ 0.75, and the obtained models were denoted as Our₀₂₅, Our₀₅₀, Our₀₇₅, respectively. To shorten the time for training, we pretrained a model without a naturalness constraint by setting $\alpha = 0$, and the model is denoted as Our_{000} . Then, we fine-tuned the model Our_{000} using three kinds of naturalness weights to obtain the models Our₀₂₅, Our₀₅₀, Our₀₇₅, respectively.

To determine the best number of pixel pairs for global contrast enhancement and to verify the effectiveness of the Swin transformer block layers, we conducted two experiments on a subset of our dataset that consisted of 10,000 artificial images and 1000 natural images, and we tested the trained models on 20,000 natural images.

5.2 Effect of Swin transformer block layer compared with CNN

To validate the effect of the Swin transformer layer, we compared the proposed model constructed using Swin transformer layers with the model based on CNN layers, which was used in the existing DNN-based recoloring method [18, 19].

First, for a fair comparison of the effect of the network structure, we excluded the global contrast term while training the models. Table 1 shows the average contrast loss of 20,000 recolored images by the two models. It can be observed that the average contrast loss of images recolored by the Swin transformer network is smaller than that by the CNN network. An example of the recoloring

 $\label{eq:contrast} \begin{array}{c} \textbf{Table 1} \\ \textbf{Contrast loss between of CNN network and transformer network} \end{array}$

Network	Net _{CNN}	Net _{Transformer}
Contrast loss	4.66e-1	3.08e-1

results by the two networks is shown in Fig. 5. The red square in the original image (Fig. 5a) is hard to distinguish from the brown background in the CVD simulation. The recoloring results of the Swin transformer-based network and CNN-based network are shown in Fig. 5b, c, respectively. Squares in both recoloring results can be easily distinguished. In Fig. 5a, the triangle has the same red color as the square; in Fig. 5b, the triangle and the square are both recolored to blue. It is important to avoid recoloring the same colors differently, as colors are usually linked to some particular semantic meanings. However, in Fig. 5c, the triangle and the square are recolored differently, demonstrating that the Swin transformer-based network could model long-range dependency successfully, while the CNN-based network failed to do so.

5.3 Number of pixel pairs for capturing global contrast

In Zhu et al.'s study [14], to enhance the global contrast, the contrast loss of all color pairs was considered, which led to a high computation cost. Given the color coherence within a particular area, we assume that a relatively small number of randomly selected pixel pairs can also work well to capture the global color contrast. To determine empirically an appropriate value for the number of pixels in ω , we



Fig. 5 An example showing that the proposed Swin transformer network-based model outperforms the existing CNN network-based one in modeling long-range dependency. **a** original image; **b** result of the Swin transformer network; **c** result of the CNN network

conducted an experiment in which we trained our Our_{000} model with ω being set to 0, 1000, 3000, 5000, and 7000, respectively. To evaluate the performance of the model trained with different values of ω , we calculated the contrast loss of the output images using Eq. 5, and the result is shown in Fig. 6. The smaller the value, the better the recoloring result. As shown in Fig. 6, the contrast loss decreases when the global points increase from 0 to 3000. However, when the global points are greater than 3000, the contrast loss increases with an increase in global points. The model trained with the parameter $\omega = 3000$ obtained the best score; thus, all remaining models in this paper are trained with the parameter $\omega = 3000$

6 Evaluation experiments

We conducted experiments to compare the results of the proposed method with two state-of-the-art recoloring methods [11, 14]. According to a recent survey [9], these two methods have the best performance among all existing methods considering both contrast enhancement and naturalness preservation. For the two existing CNN-based methods [18, 19], we already demonstrated in Sect. 5.2 that our Swin transformer-based structure could achieve better results in contrast enhancement. These two methods do not consider naturalness preservation; therefore, we excluded them from further evaluation experiments.

Figures 7 and 8 show the recoloring results of the proposed and existing methods for protan and deutan CVD, respectively. In each figure there two examples; for each example, the first row shows the input images (Figs. 7a and 8a), the recolored images of Hassan et al. [11] (Figs. 7b and 8b), the recoloring results of Zhu et al. [14] (Figs. 7c and 8c), and the proposed methods $Our_{025},Our_{050},Our_{075}$ with different naturalness weights, i.e., $\alpha = 0.25$ (Figs. 7d and 8d), $\alpha = 0.5$ (Figs. 7e and 8e), and $\alpha = 0.75$ (Figs. 7f and 8f), respectively. The corresponding CVD simulation images are aligned in the second row to help readers with normal vision understand intuitively how the original and recolored images are perceived by CVD individuals. The



Fig. 6 Contrast loss with different numbers of global points



Fig. 7 Results for the existing methods and proposed method (Protan). **a** Input image. **b** Result of Hassen. **c** Result of Zhu. **d** Result of Our_{025} . **e** Result of Our_{050} . **f** Result of Our_{075}

CVD simulation images are obtained using the CVD simulation model proposed in [31].

For Example 1 in Fig. 7, the flowers in the input image (Fig. 7a) are almost indistinguishable from the ground by individuals with protanopia. For contrast enhancement, the result by Hassan et al. [11] (Fig. 7b) added too much blue component, leading to an unnatural appearance. The recolored image by Our_{075} (Fig. 7f) is similar to that by Zhu et al. [14] (Fig. 7c), which preserved naturalness well while failing to enhance the contrast between the flowers and the background. The proposed models Our_{025} (Fig. 7d) and Our_{050} (Fig. 7e) succeeded in enhancing the contrast in recolored images, where the flowers are easily distinguishable and their color appearance is natural. For Example 2 in Fig. 7, the sun in the original image (Fig. 7a) is almost unnoticeable from the perception of CVD individuals. The result by Hassan et al. [11] (Fig. 7b) changes the color of the sun to blue, which is still difficult to distinguish from the blue sky. It seems Our_{075} (Fig. 7f) put so much weight on naturalness that it failed to enhance the contrast between the sun and the sky. The recoloring results by Zhu et al. [14] (Fig. 7c), Our₀₂₅ (Fig. 7d), and Our₀₅₀ (Fig. 7e) succeeded in preserving naturalness and enhancing contrast, that is, making the sun distinguishable. For Example 1 in Fig. 8, the contrast between the lotus leaves and the flowers in the input image (Fig. 8a) is weakened for those with deuteranopia. The result by Hassan et al. [11] added too much value to the blue channel, resulting in the loss of naturalness, while their method also failed to enhance the color contrast. For the recolored image by Zhu et al. [14], there is almost no contrast enhancement effect. Meanwhile, in the recolored images by Our₀₂₅, Our₀₅₀, and Our₀₇₅, the flowers and leaves are easily distinguished. For Example 2 in Fig. 8, the contrast between the flower and the background is almost unnoticeable in the simulation of the original image (Fig. 8a). The result by Hassan et al. [11] (Fig. 8b) significantly changed the colors of the flower, but it added too much blue component, leading to an unnatural appearance. Further, the recolored images by Zhu et al. [14] (Fig. 8c) and Our₀₇₅ (Fig. 8f) failed to enhance the contrast. Meanwhile, the recoloring result of Our_{025} (Fig. 8d) and Our_{050} (Fig. 8e) succeeded in enhancing the contrast. But Our₀₂₅ (Fig. 8d) changed the color too much, while Our₀₅₀ (Fig. 8e) succeeded in preserving naturalness, keeping the flowers as same as the original one.



Fig. 8 Results for the existing methods and proposed method (Deutan). a Input image. b Result of Hassen. c Result of Zhu. d Result of Our_{025} . e Result of Our_{050} . f Result of Our_{075}

6.1 Quantitative evaluation

To assess the proposed methods objectively in terms of both naturalness and contrast, we conducted a quantitative experiment using total color contrast (TCC) and SSIM metrics to compare state-of-the-art studies on contrast enhancement and naturalness preservation, respectively. The TCC metric calculates the absolute contrast in an image, and it is defined as follows:

$$TCC = \frac{1}{n_1} \sum_{(i,j)\in\Omega_1} (|x_i - x_j|) + \frac{1}{N * n_2} \sum_{i=1}^N \sum_{j\in\Omega_2} (|x_i - x_j|)$$
(10)

where Ω_1 denotes a set of randomly selected global pixel pairs, and n_1 stands for the number of pixel pairs in Ω_1 ; Ω_2 represents a set of pixels located at the window centered at an arbitrary pixel x_i , and n_2 indicates the number of pixels in Ω_2 ; finally, N represents the number of pixels in the whole image. In our experiment, n_1 was set to 20,000, and the window size for Ω_i was set to 11×11 .

To validate the effectiveness of compensating for protanopia and deuteranopia, 10 natural scene, plant, artificial object, and painting images were selected for each type of color vision deficiency in this study. Table 2 shows the TCC results. The larger the value, the better the contrastenhancing effect. The top three scores are indicated in ascending order by ***, **, and *, respectively. This shows that for protan CVD, Hassan et al. [11] achieved the best score, while the proposed models Our_{025} and Our_{050} achieved higher scores than the model proposed by Zhu et al.[14]; for deutan CVD, the proposed models Our_{025} achieved the best score, and all proposed models outper-form that in [14].

We used chromatic difference(CD) metric as the metric for evaluating naturalness preservation. CD is computed in the CIE Lab color space, and it can be computed as:

$$CD(i) = \sqrt{\lambda(l'_i - l_i)^2 + (a'_i - a_i)^2 + (b'_i - b_i)^2}$$
(11)

Table 2 Result for the total color contrast metric (contrast)

CVD Type	Hassan	Zhu	Our ₀₂₅	Our ₀₅₀	Our ₀₇₅
Protan	0.725***	0.668	0.711**	0.670*	0.636
Deutan	0.879**	0.776	0.898***	0.851*	0.792

where i ranges over the pixels in the image; $l'_i, l_i, a'_i, a_i, b'_i, b_i$ are the L,a,b value of homologous pixel i in the test image and reference image. By setting $\lambda = 0$, same metric is adopted in the [14]. The same images used in the quantitative evaluation for contrast are used to evaluate the naturalness metric. The average scores are shown in Table 3. The bigger the value, the better the result. The first-, second-, and third-place scores are indicated by ***, **, and *, respectively. This shows that for protan, images recolored by Our₀₅₀ and Our₀₇₅ achieved better scores than those by [11, 14]; for deutan, images recolored by Our₀₇₅ achieved a better score than those by [11, 14].

The result of the quantitative experiment will be further discussed in the Discussion section.

6.2 Subjective evaluation

To evaluate subjectively the performance of the proposed method in comparison with state-of-the-art methods, we invited 18 CVD volunteers aged from 18 to 50 years. First, we used the Ishihara test and Farnsworth Panel D-15 test to diagnose their CVD types, where 5 volunteers were diagnosed with protan CVD and 13 with deutan CVD. CVD volunteers were asked to evaluate recolored images according to three aspects: contrast, naturalness, and preference. The same images used in the quantitative experiment were used in the subjective evaluation experiment. In this study, all images were presented on an Iiyama ProLite 23-inch display calibrated using the X-Rite i1Display 2 under ambient illuminance below 10 lux, and all participants were siting half a meter from the display. Given one input image, the recoloring results produced by five different methods are presented. In other words, six images in total, including one input image and five recolored images, were shown to the participants at a time. And the participants were notified which one is input image. We use the boxplot to display the score by all subjects on all images for contrast, naturalness, and preference, respectively, and use the solid square to present the median. We also conducted the Mann Whitney Test to analyze the results. The results are shown in Figs. 9 and 10. In the following descriptions, if we say A is significantly different from B, it means that A has a significantly better score than B at a significance level of 5%.

Table 3 Result for the CD metric (naturalness)

CVD type	Hassan	Zhu	Our ₀₂₅	Our ₀₅₀	Our ₀₇₅
Protan	0.923	0.992*	0.972	0.995**	0.999***
Deutan	0.899	0.995**	0.963	0.983*	0.998***





Fig. 9 The result of the subjective experiment. (Protan)



Fig. 10 The result of the subjective experiment. (Deutan)

Contrast Participants were required to compare the recolored images with the original image and to rate the degree of contrast in the recolored image on a scale of 1 to 5, where "1: contrast decreased significantly," "2: contrast decreased slightly," "3: almost the same with the original image," "4: contrast enhanced slightly," and "5: contrast enhanced significantly." As shown in Fig. 9, for protan CVD, Our₀₂₅, Zhu et al. [14], and Our₀₅₀ are in the top three, and Our₀₂₅ is significantly different from Zhu et al. [14], while Zhu et al. [14] is significantly different from Our₀₇₅. For deutan CVD, as shown in Fig. 10, Our₀₂₅, Our₀₅₀, and Hassan et al. [11] are in the top three. Meanwhile, Our₀₂₅ and Our₀₅₀ are significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [11] is significantly different from Zhu et al. [14], and Hassan et al. [14] is significantly different from Zhu et al. [14], and Hassan et al. [14] is significantly different from Zhu et al. [14] is significantly different from Zhu et a

Naturalness Similar to the contrast evaluation, all participants were asked to compare the recolored images with the original image and to rate the statement, "The color appearance of the recolored image is similar to the original image," on a scale of 1 to 5, where "1: totally disagree," "2: disagree," "3: neutral," "4: agree," and "5: totally agree." As shown in Fig. 9, for protan CVD, Our_{075} , Our_{050} , and Zhu et al. [14] are in the top three. Our_{075} , Our_{050} and Our_{025} are all significantly different from Hassan et al. [11]. Our_{075} and Our_{050} are significantly different from Zhu et al. [14], while Zhu et al. [14] is only significantly from Our_{025} . For deutan CVD, as shown in Fig. 10, Our_{075} , Zhu et al. [14], and Our_{050} are in the top three. Our_{075} , Our_{050} and Our_{025} are all significantly different from Hassan et al. [14]. And Our_{050} are in the top three. Our_{075} , Our_{050} and Our_{025} are all significantly different from Hassan et al. [11]. And Our_{075} is significantly different from Zhu et al. [14]. Zhu et al. [14] is significantly from Our_{025} and Our_{050} . For the proposed models, which were trained with different α values, that is, weights of naturalness, the scores become higher with an increasing α .

Preference As a comprehensive evaluation of both contrast and naturalness, CVD participants were asked to evaluate recolored images according to their preference and to rate the statement, "The recolored image is preferable from the aspects of both contrast and naturalness," on a scale of 1 to 5, where "1: totally disagree," "2: disagree," "3: neutral," "4: agree," and "5: totally agree." As can be observed from Figs. 9 and 10, for protan CVD, Our₀₇₅, Our₀₅₀, and Zhu et al. [14] are in the top three. Our₀₇₅,Our₀₅₀ and Our₀₂₅ are all significantly different from Hassan et al. [11]. Our₀₇₅ and Our₀₅₀ are significantly different from Zhu et al. [14], while Zhu et al. [14] is only significantly from Our₀₂₅; for deutan CVD, Our₀₇₅, Zhu et al. [14], and Our_{050} are in the top three. Our_{075} , Our_{050} and Our₀₂₅ are all significantly different from Hassan et al. [11]. And Our_{075} is significantly different from Zhu et al. [14]. Zhu et al. [14] is significantly from Our_{025} and Our_{050} .

From the contrast and naturalness evaluation results, we can see a tradeoff between the two, as shown in Fig. 11. Nevertheless, all scores of the Our_{050} model are higher than 3.0, meaning it can improve the contrast but also maintain naturalness for both protan and deutan users. Moreover, its preference score is also larger than 3.0 for both protan and deutan cases, meaning the recolored images are preferred by all users. The result of the subjective experiment will be further considered in the Discussion section.

6.3 Visualization experiment

Besides the experiment with natural images, we also conducted an evaluation experiment using images without the issue of naturalness preservation. In such a case, a recoloring method is usually required to enhance the perceivability of the information contained in the image, which is visible to individuals with normal vision and almost invisible to people with CVD. In this study, we took figures from the Ishihara test as input images and compared the recoloring results of the proposed model Our_{000} with those of the methods in [11] and [14]. In [14], they set a



Fig. 11 The subjective experiment result of the proposed method. a the result for Protan; b the result for Deutan

smaller weight, 0.01, for naturalness in the objective function when applying the model to the Ishihara images. We used the same weight in our experiment. The results of the two examples are shown in Figs. 12 and 13. The first row of each figure shows the input image (Figs. 12a, and 13a) and the recoloring results of Hassan et al. [11] (Figs. 12b and 13b), Zhu et al. [14] (Figs. 12c and 13c), and Our_{000} (Figs. 12d and 13d), while the second row shows the CVD simulation results for the images in the first row.

Individuals with protanopia have difficulty recognizing the figure "5" in Fig. 12a; given the recoloring result of



Fig. 12 Recoloring results of Ishihara images using existing methods and the proposed method (Protan). **a** original image. **b** result of Hassan et al. [11] **c** result of Zhu et al. [14]. **d** Result of Our₀₀₀



Fig. 13 Recoloring results of Ishihara images using existing methods and the proposed method (Deutan). \mathbf{a} original image. \mathbf{b} result of Hassan et al. [11] \mathbf{c} result of Zhu et al. [14]. \mathbf{d} Result of Our_{000}

[11] (Fig. 12b), it is still very difficult for people with protanopia to recognize the correct figure, as depicted by the simulation images. With the recoloring result of [14] (Fig. 12c) or Our₀₀₀ (Fig. 12d), CVD individuals can read the figure "5" correctly. Individuals with deuteranopia have difficulty recognizing the figure "97" in Fig. 13a; given the recoloring result of [11] (Fig. 13b) or [14] (Fig. 13c), it is still difficult for people with deuteranopia to recognize the figure, as depicted by the simulation images. With the recoloring result of Our₀₀₀, though the color appearance is changed significantly, CVD individuals do not have any problem reading the figure. In this experiment, 16 Ishihara images in total were selected and presented to the participants, together with the recoloring results, and the evaluation results are shown in Table 4. Each value in Table 4 indicates the number of recolored images that could be correctly identified by the participant, and the best score of each participant is in bold. In contrast to the recoloring results of [11] and [14], which only partially assisted CVD individuals in passing the Ishihara test, Our₀₀₀ enabled CVD individuals to recognize all the images. Hassan et al.'s [11] method may result in images with a saturated blue channel, and it may fail to enhance the contrast in the areas containing colors consisting of blue components. Zhu et al. [14] recolored key colors and then propagated the results to all colors. If the color of the figure does not take up a substantial proportion, their method will fail to enhance the contrast between the figure and the background.

Table 4 The number of correctly recognized Ishihara test images

CVD type	Original	Hassan	Zhu	Our ₀₀₀
Protan	1.4	7.4	11	16
Deutan	1.8	7.6	11.2	16

6.4 Time efficiency evaluation

To evaluate the time efficiency, we run all the methods on the same machine on which we implemented our model in the experiment. On average, our method takes less than 0.02s to process a 256×256 image, which is faster than the 0.08s of Hassan et al.'s method [11] and the 7.16s of Zhu et al.'s method [14].

6.5 Different global contrast weight

In this study, we initially set the weights for global contrast and local contrast in a 1:1 ratio. We conducted an evaluation of the model's performance across varying ratios of global contrast to local contrast to determine the appropriateness of this configuration. To maintain the contrast and naturalness ratio as defined in Eq. 5, we ensured that the sum of the contrast weights equaled 2. Assuming the global weight is denoted as β , the weight of the local contrast was calculated as 2 - β . And we adopt the Our₀₅₀ model, and the results are presented in Table 5. As observed, increasing the weight of global contrast leads to an increase in the contrast evaluation value but a decrease in the naturalness evaluation value. Therefore, for a balanced trade-off between contrast and naturalness, a setting of $\beta = 1$ is deemed reasonable.

6.6 Ablation study

In our study, there are global contrast(GC), local contrast(LC) and naturalness(N) components in the loss function Eq. 8. To test the impact of each component, we conducted an ablation study where we removed GC, LC, N component from Eq. 8 respectively. And we use the $\alpha =$ 0.5 in Eq. 8. The evaluation result is shown in Table 6. When focusing solely on the contrast component, we observe that the contrast value is the best, but the naturalness value is the lowest. When considering the contrast component separately - global contrast alone yields a considerably high contrast value, whereas local contrast yields a significantly lower contrast value. Remarkably, the combination of both global and local contrast results in a satisfactory contrast value, while preserving a high level of naturalness.

 Table 5
 Contrast evaluation value and naturalness evaluation value with different global weight

Global weight (β)	0	0.5	1	1.5	2
Contrast	0.732	0.759	0.76	0.765	0.772
Naturalness	0.999	0.996	0.996	0.992	0.988

GC	LC	Ν	Contrast value	Naturalness value
			0.760	0.996
	1		0.819	0.935
			0.772	0.988
	1		0.732	0.999

Table 6 Ablation study with different loss combination

7 Discussion

In this study, the resolution of image is 256×256 , so if we want to deal with larger images, we can reduce the image to 256×256 first, and then use the state-of-the-art superresolution method (e.g., Wang et al. [36]) to enlarge the recolored image. We proposed a DNN-based image recoloring model, which was trained with different naturalness weights, namely, Our₀₂₅, Our₀₅₀, and Our₀₇₅. We also conducted qualitative, quantitative, and subjective experiments to evaluate these models and compare them with state-of-the-art studies. The results of quantitative and subjective evaluation experiments show that with the increase in the naturalness weight, the scores of naturalness and preference increased, while the contrast score decreased. To further investigate how contrast and naturalness impact the preference, we computed two correlation coefficients, one between contrast and preference and the other between naturalness and preference. Figures 14 and 15 show the plots of all subjects, with the x-axis representing the coefficient between contrast and preference and the y-axis representing the coefficient between naturalness and preference We can clearly see that for naturalness, there is a high correlation with preference for almost all subjects. However, for contrast, correlations vary largely by subject. First, this means that preserving naturalness is crucial when considering user preferences. Nevertheless, the requirements of CVD compensation may vary by



Fig. 14 The correlation coefficient of contrast and naturalness with the preference of each CVD individual with protanopia



Fig. 15 The correlation coefficient of contrast and naturalness with the preference of each CVD individual with deuteranopia

situation, and contrast enhancement may become the most important characteristic if recognizing the detailed information in an image is critical, such as in the visualization experiment shown in Sect. 6.3. One fact is that the subjective contrast experiment did not use such a task, and the images presented are pictures or paintings of natural scenes and still life, where the color appearance is likely a more dominant factor in affecting subjects' preferences. Therefore, pre-training a set of recoloring models with different contrast and naturalness weights and then selecting an appropriate model according to the particular situation and user preferences is a promising way to use our method in real applications.

Figure 16 shows an example where all methods failed to enhance the color contrast between the purple ball and blue balls. Our current models consider the relationships between sample pixels, and it may be possible to solve such a problem by considering the relationships between objects. In the current study, contrasts in all areas of an input image are treated with the same importance; however, contrast enhancement to the RoI (region of interest) in the image should be much more important to CVD users. Thus, incorporating some RoI prediction approach, such as the salient detection (e.g., Zhao et al. [37]) technique, to the recoloring model can be another important future work. By relaxing the constraints on contrast enhancement in lesssalient areas, it becomes easier to restrict the contrast enhancement to salient areas while better preserving naturalness in less-salient areas.

8 Conclusion

In this paper, we proposed a DNN-based image recoloring method for generating compensation images for CVD individuals that considers contrast enhancement and naturalness preservation. Considering long-range color dependency, we adopted Swin transformer blocks in the



Fig. 16 The correlation coefficient of contrast and naturalness with the preference of each CVD individual with deuteranopia

proposed network. The evaluation experiment showed that the proposed model achieved image qualities that are competitive with state-of-the-art studies while improving time efficiency drastically. The proposed unsupervised training approach, as well as the dataset containing images with confusing colors for CVD individuals, can be used to develop and evaluate other image synthesis technologies for assisting CVD users. Currently, we set people with dichromacy as our compensation target; however, CVD individuals' perceptions differ from person to person, and the recoloring results may be unsuitable for individuals with lower degrees of CVD. Therefore, adapting the recoloring model to the degree of CVD can be important future work. Additionally, we could expand our research from images to videos in the future. In videos, we can use a visual tracking method described in [38] to follow objects and ensure their colors remain consistent from one frame to the next.

Author contributions CL, ZZ, and MX designed the algorithm and the network architecture. CL was responsible for the implementation and for conducting the experiments. HW, GK, and CX contributed to the design of the experiments and the analysis of the results. CL served as the primary author of the paper. ZZ and CX reviewed the paper and provided valuable feedback, while MX played a crucial role in enhancing the manuscript's quality. The funding for this project was provided by MX and ZZ.

Funding Open Access funding provided by University of Yamanashi. This work is supported by JSPS Grants-in-Aid for Scientific Research (Grant Nos. 20K20408, 22H00549, 22K21274, 23K16899). The authors would like to thank all participants for evaluating the proposed method and their constructive comments.

Data availability All data generated or analyzed during this study are included in this published article.

Code availability The code and model will be made available at https://github.com/Ligeng-c/CVD_swin.

Declarations

Conflict of interest The authors declare that they have no competing interests

Ethics approval The experimental protocol was established, according to the ethical guidelines of the Helsinki Declaration and was approved by the Ethics Committee of University of Yamanashi.

Consent to participate Written informed consent was obtained from individual or guardian participants.

Consent for publication Written informed consent was obtained from individual or guardian participants.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

- Sharpe LT, Stockman A, Jägle H, Nathans J (1999) Opsin genes, cone photopigments, color vision, and color blindness. Color Vis: Fenes Percept: 3–51
- 2. Hunt RWG (2005) The Reproduction of Colour. Wiley, New York
- Wakita K, Shimamura K. Smartcolor: disambiguation framework for the colorblind. In: Proceedings of the 7th international ACM SIGACCESS conference on computers and accessibility, pp 158–165
- Iaccarino G, Malandrino D, Del Percio M, Scarano V. Efficient edge-services for colorblind users. In: Proceedings of the 15th international conference on World Wide Web, pp 919–920

- Jefferson L, Harvey R. Accommodating color blind computer users. In: Proceedings of the 8th international ACM SIGACCESS conference on computers and accessibility, pp 40–47
- Machado GM, Oliveira MM. Real-time temporal-coherent color contrast enhancement for dichromats. In: Computer graphics forum, Wiley Online Library, vol. 29, pp 933–942
- Lin HY, Chen LQ, Wang ML (2019) Improving discrimination in color vision deficiency by image re-coloring. Sensors (Basel) 19(10):2250. https://doi.org/10.3390/s19102250
- Ribeiro M, Gomes AJP (2019) Recoloring algorithms for colorblind people: a survey. Acm Comput Surv 52(4):1–37. https://doi. org/10.1145/3329118
- Zhu Z, Mao X (2021) Image recoloring for color vision deficiency compensation: a survey. Vis Comput 37(12):2999–3018. https://doi.org/10.1007/s00371-021-02240-0
- Hassan MF, Paramesran R (2017) Naturalness preserving image recoloring method for people with red–green deficiency. Sign Process: Image Commun 57:126–133
- Hassan MF (2019) Flexible color contrast enhancement method for red–green deficiency. Multidimens Syst Sign Process 30(4):1975–1989. https://doi.org/10.1007/s11045-019-00638-7
- Zhu Z, Toyoura M, Go K, Fujishiro I, Kashiwagi K, Mao X (2019) Naturalness- and information-preserving image recoloring for red–green dichromats. Sign Process: Image Commun 76:68–80. https://doi.org/10.1016/j.image.2019.04.004
- Zhu Z, Toyoura M, Go K, Fujishiro I, Kashiwagi K, Mao X (2019) Processing images for red–green dichromats compensation via naturalness and information-preservation considered recoloring. Vis Comput 35(6–8):1053–1066. https://doi.org/10. 1007/s00371-019-01689-4
- Zhu Z, Toyoura M, Go K, Kashiwagi K, Fujishiro I, Wong T-T, Mao X (2022) Personalized image recoloring for color vision deficiency compensation. IEEE Trans Multimed 24:1721–1734. https://doi.org/10.1109/tmm.2021.3070108
- Wang X, Zhu Z, Chen X, Go K, Toyoura M, Mao X (2021) Fast contrast and naturalness preserving image recolouring for dichromats. Comput Graph 98:19–28. https://doi.org/10.1016/j. cag.2021.04.027
- Huang WK, Zhu ZY, Chen LG, Go K, Chen XD, Mao XY (2022) Image recoloring for red–green dichromats with compensation range-based naturalness preservation and refined dichromacy gamut. Vis Comput 38(9–10):3405–3418. https://doi.org/10. 1007/s00371-022-02549-4
- Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
- Li HS, Zhang L, Zhang XD, Zhang ML, Zhu GM, Shen PY, Li P, Bennamoun M, Shah SAA (2020) Color vision deficiency datasets and recoloring evaluation using gans. Multimed Tools Appl 79(37–38):27583–27614. https://doi.org/10.1007/s11042-020-09299-2
- Xinghong H, Xueting L, Zhuming Z, Menghan X, Chengze L, Tien-Tsin W (2019) Colorblind-shareable videos by synthesizing temporal-coherent polynomial coefficients. ACM Trans Graph 38(6):1–12. https://doi.org/10.1145/3355089.3356534
- Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 234–241
- Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:.14030

- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
- 23. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:.11929
- Jefferson L, Harvey R (2007) An interface to support color blind computer users. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 1535–1538
- Judd DB (1966) Fundamental studies of color vision from 1860 to 1960. Proc Natl Acad Sci U S A 55(6):1313–30. https://doi.org/ 10.1073/pnas.55.6.1313
- 26. Kuhn GR, Oliveira MM, Fernandes LA (2008) An efficient naturalness-preserving image-recoloring method for dichromats. IEEE Trans Vis Comput Graph 14(6):1747–54. https://doi.org/10. 1109/TVCG.2008.112
- Huang HB, Tseng YC, Wu SI, Wang SJ (2007) Information preserving color transformation for protanopia and deuteranopia. IEEE Sign Process Lett 14(10):711–714. https://doi.org/10.1109/ Lsp.2007.898333
- Ebelin P, Crassin C, Denes G, Oskarsson M, Åström K, Akenine-Möller T (2023) Luminance-preserving and temporally stable daltonization. In: EUROGRAPHICS 2023, the 44th annual conference of the European association for computer graphics. Eurographics-European association for computer graphics
- Jiang S, Liu D, Li D, Xu C (2023) Personalized image generation for color vision deficiency population. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 22571–22580
- 30. Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T (2020) Training generative adversarial networks with limited data. Adv Neural Inf Process Syst 33:12104–12114
- Machado GM, Oliveira MM, Fernandes LA (2009) A physiologically-based model for simulation of color vision deficiency. IEEE Trans Vis Comput Graph 15(6):1291–8. https://doi.org/10. 1109/TVCG.2009.113
- Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–12. https://doi.org/10.1109/ tip.2003.819861
- 33. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Places: a 10 million image database for scene recognition. IEEE Trans Patt Anal Mach Intell 40(6):1452–1464. https://doi.org/10. 1109/TPAMI.2017.2723009
- 34. Lu C, Xu L, Jia J. Contrast preserving decolorization. In: 2012 IEEE international conference on computational photography (iccp), IEEE, pp 1–7
- 35. Hu XH, Zhang ZM, Liu XT, Wong TS (2019) Deep visual sharing with colorblind. IEEE Trans Comput Imag 5(4):649–659. https://doi.org/10.1109/Tci.2019.2908291
- 36. Wang X, Xie L, Dong C, Shan Y. Real-esrgan: training realworld blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1905–1914
- Zhao T, Wu X. Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3085–3094
- Yang K, He Z, Pei W, Zhou Z, Li X, Yuan D, Zhang H (2021) Siamcorners: Siamese corner networks for visual tracking. IEEE Trans Multimed 24:1956–1967

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.