Globally and Locally Semantic Colorization via Exemplar-Based Broad-GAN

Haoxuan Li, Bin Sheng, *Member, IEEE*, Ping Li, *Member, IEEE*, Riaz Ali, and C. L. Philip Chen, *Fellow, IEEE*

***Abstract*—Given a target grayscale image and a reference color image, exemplar-based image colorization aims to generate a visually natural-looking color image by transforming meaningful color information from the reference image to the target image. It remains a challenging problem due to the differences in semantic content between the target image and the reference image. In this paper, we present a novel globally and locally semantic colorization method called exemplar-based conditional broad-GAN, a broad generative adversarial network (GAN) framework, to deal with this limitation. Our colorization framework is composed of two sub-networks: the match sub-net and the colorization sub-net. We reconstruct the target image with a dictionary-based sparse representation in the match sub-net, where the dictionary consists of features extracted from the reference image. To enforce global-semantic and local-structure self-similarity constraints, global-local affinity energy is explored to constrain the sparse representation for matching consistency. Then, the matching information of the match sub-net is fed into the colorization sub-net as the perceptual information of the conditional broad-GAN to facilitate the personalized results. Finally, inspired by the observation that a broad learning system is able to extract semantic features efficiently, we further introduce a broad learning system into the conditional GAN and propose a novel loss, which substantially improves the training stability and the semantic similarity between the target image and the ground truth. Extensive experiments have shown that our colorization approach outperforms the state-of-the-art methods, both perceptually and semantically.**

***Index Terms*—Image colorization, image manipulation, adversarial generative networks, example-based, broad learning.**

I. Introduction

IMAGE colorization refers to assigning colors to a grayscale image in such a way that it looks natural. Specifically, given a grayscale image as the input image, the purpose of image colorization is to output a chromatic image that is visually realistic and perceptually appealing. Image colorization is a valuable technique with many practical applications like automatically colorizing the black-and-white films or cartoon scenes in a meaningful manner. Therefore, optimizing the image colorization method is a worthwhile pursuit. However, this problem is time-consuming and intrinsically equivocal since there are feasibly many different colors that can be allocated to the gray pixels of a target image (for instance, the color of the same tree leaves may vary in different seasons). Therefore, there is no single right solution, and human intervention usually plays a significant role in the coloring process.

Prevailing colorization approaches are classified into the following three categories: scribble-based colorization [1]–[5], example-based colorization [6]–[13], and learning-based colorization [14]–[20]. Scribble-based methods require the user to assign suitable colors to specific pixels according to the patch's semantic and luminance to achieve a plausible result. These approaches are not only time-consuming and labor-intensive but also require the users to have a strong artistic sensibility. Hence, using them is a challenge for most users. Regarding the example-based methods, given a grayscale image and a similar reference image, these techniques will output a colorized image in the light of the provided chrominance information. These methods also take a lot of time to find a suitable reference image. More seriously, the quality of colorized results particularly depends on the selection of the reference image. To further reduce the burden of users, learning-based colorization algorithms have been proposed. In these approaches, a large-scale image database is leveraged to train the neural network in order to predict the appropriate colors of the target image. Unfortunately, the produced result is uncontrollable precisely because the whole colorization process is completely automatic. Hence, none of the learning-based methods allows customization. Besides, we cannot get satisfactory results when similar objects are not contained in the reference image database.

Due to the considerable capability of learning image distribution, GANs [21] can be applied to generate the synthetic color images [22]. Nevertheless, unfortunately, the original GANs are often subjected to the model collapse problem, which directly affects the training stability and makes it difficult to ensure the image quality. The conditional GAN [23], one variant of GANs, provides a new way to overcome these obstacles. Extra information can be utilized to constrain the generator by guiding the color generation process, and assist the discriminator by providing hints for the discrimination process. Substantial works based on the above-mentioned methods have been successfully used in style transfer [22],

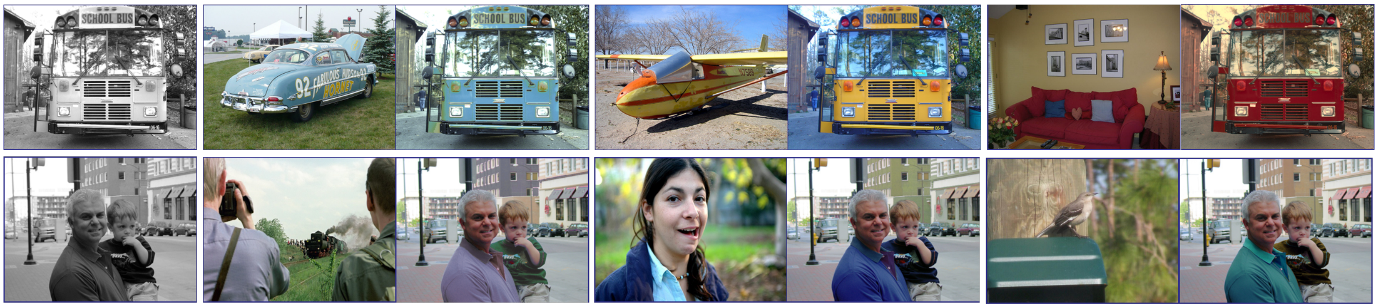
Manuscript received January 14, 2020; revised July 01, 2021; accepted September 22, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61872241 and Grant 61572316, in part by the National Key Research and Development Program of China under Grant 2019YFB1703600, and in part by The Hong Kong Polytechnic University under Grant P0030419, Grant P0030929, and Grant P0035358.

H. Li and B. Sheng are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (Email: shengbin@sjtu.edu.cn).

P. Li is with the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong (Email: p.li@polyu.edu.hk).

R. Ali is with the Department of Computer Science, Sukkur IBA University, Sukkur 65200, Sindh, Pakistan (Email: riaz.khp@iba-suk.edu.pk).

C. L. P. Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China; with the Navigation College, Dalian Maritime University, Dalian 116026, China; and also with the Faculty of Science and Technology, University of Macau, Macau 999078, China (Email: philip.chen@ieee.org).



(a) Target (b) Reference 1 (c) Result 1 (d) Reference 2 (e) Result 2 (f) Reference 3 (g) Result 3

Fig. 1: Image colorization results using our method with different reference images. By using different reference images, the same target grayscale image is effectively colorized to have different yet meaningful colors even if the semantic content of the reference image is significantly different from target images (f) and (g).

[24] and image inpainting [25]. However, they have seldom been used for colorization to realize personalization within the GANs' literature.

Our recent work attempts to achieve the optimal solution based on two valuable aspects: interaction and robustness. Inspired by the recent success in GANs, which have achieved plausible results in style transfer [26], this paper reformulates the colorization problem to address the limitations discussed previously. We define a sparse representation learning method to reconstruct the input target image. And then, this reconstructed information will be fed into conditional GANs as a 'hint' to tackle the issue of training and ensure the personalized colorized results. Finally, a broad learning system is added to conditional GANs to improve the problem of model collapse. In addition, we propose a simple yet effective loss function in our GAN's architecture to enforce the semantic similarity even if there is substantial color variation between the reference image and the target image. The proposed network realizes the chromaticity transfer between multiple images, even if the contents are quite different. As shown in Fig. 1, one can change the image's chrominance information while keeping the image visually meaningful by using various reference images. Note that the luminance information of the target image always remains unchanged during the whole process.

To achieve these goals, we divide our framework into two sub-networks. First, the match sub-net is used as a preprocessing step of the colorization network. Compared to the previous example-based colorization methods [6]–[8], which estimate possible colors based on semantic similarity, our approach considers both global semantic similarity and local structure constraint simultaneously. In many instances, the reference image may be unrelated to the input grayscale image. Hence, it may be difficult to achieve the desired results by dense correspondence. We establish the relationship between the reference and the target image through feature extraction and sparse representation. We first utilize the pre-trained VGG network [27] to extract the features of the target image and the reference image in the luminance channel only. And then, we reconstruct the target grayscale image through the sparse representation method based on the dictionary, which is composed of the reference image features. Inspired by the observation [28] that patches with similar semantics within the whole image and pixels with close spatial distance will have similar sparse representations, we define global-local affinity energy that leverages the global-semantic and local-structure self-similarity to constrain the sparse matrix. The self-similarity constrained sparse representation dramatically decreases the noise in image reconstruction, thus improving the proposed algorithm's robustness.

Then the colorization sub-net realizes the color transformation and prediction by an exemplar-based conditional broad-GAN. We take a set of chrominance images and grayscale images for training. To reduce restrictive demands on reference images while achieving high-quality results, we do not require a high similarity between a reference and a target image. The matching information that was generated by the match sub-net is input into the GAN to facilitate customization. In addition to transforming correct colors between the reference image and target image, another important goal of image colorization is to ensure that the output color images contain semantic features of the input images. Considering the high efficiency of the broad learning system (BLS) [29], we have introduced the BLS into the discriminator to extract the semantic features of colorized images and reference images, thereby guaranteeing the presence of the target image's semantic features. Therefore, two diverse loss functions suitable for the multi-task learning framework are devised: 1) Chrominance loss, which enforces the sub-net to transfer the correct reference colors. 2) Perceptual loss, which is formulated as a regularization in the semantic feature maps of the BLS network to enforce semantic similarity between the ground truth image and the target image. In sum, our approach has the following contributions:

- **Global-local affinity energy is designed for matching consistency and robustness.** We propose a novel match sub-net architecture for feature matching and reconstruction. Global-local affinity energy is introduced to the target grayscale image to constrain the sparse representation for the global-semantic and local-structure self-similarity constraints, which dramatically improves the matching consistency and robustness.
- **Supplemental matching information is used as an extra condition of the conditional GAN for user-**

guided colorization. We propose to utilize the matching information between the target image and reference image as an additional condition and feed it into both the generator and the discriminator. This extra information, as a form of perceptual loss or as a 'hint', enforces the generation of customization results and helps the network converge to a good state efficiently.

- **Exemplar-based conditional broad-GAN is proposed for semantic-aware image colorization.** We propose a novel semantic-aware network architecture, called exemplar-base conditional broad-GAN, for jointly learning correct color propagation and prediction based on the reference image. Furthermore, we introduce a perceptual loss formulated as an L_2 sparse regularization in the semantic-level feature maps of the BLS network for keeping the semantic content of the target image.

II. Related Work

In this section, we introduce related colorization works. We classify image colorization methods into four categories: scribble-based colorization, example-based colorization, learning-based colorization, and hybrid colorization.

A. Scribble-Based Colorization

These methods require users to describe the picture with a few color points or strokes, and then the specified colors will automatically propagate to the entire target image. The earliest interactive work [1] propagated colors through Markov Random Field based on a prerequisite that the colors of neighboring pixels depend on their intensity similarity metrics. Huang et al. [2] proposed a novel adaptive edge detection algorithm that applied the Sobel filter with a high threshold to detect edges and then extended them to protect from bleeding over region boundaries in the colorization process. In order to process mangas, Qu et al. [3] applied the Gabor wavelet to measure the pattern continuity and the level set method to maintain the pattern continuity. Inspired by the ideas of luminance-weighted chrominance blending and fast intrinsic distance computations, Yatziv and Sapiro [4] generated high-quality colorization results. Luan et al. [5] further extended texture-similarity constraints through a color labeling scheme. However, scribble-based methods require intensive manual operation and cannot provide rich enough color information.

B. Example-Based Colorization

Compared with the scribble-based methods, the example-based colorization reduces the intensity of manual operation substantially. These methods transfer color information from a reference image, which can be supplied by users or taken from the web, to the target grayscale image. Welsh et al. [10] transferred color information by matching luminance and texture information between the images. Charpiat et al. [6] defined a nonuniform spatial coherency criterion to estimate the color probability distribution of each pixel. To let the method be robust to any illumination differences between grayscale image and reference images, Liu et al. [9] reconstructed the illumination-independent intrinsic reflectance image of the target from the reference image. Chia et al. [7] required users to provide a semantic text label and segmentation cues for finding a suitable reference image from the websites. Gupta et al. [8] developed a superpixel representation approach to encourage spatial coherence. It is further improved in [13] by taking into account the intensity, texture, and semantic features. Each superpixel's descriptor is constructed by concatenating the extracted low-level intensity features, mid-level texture features, and high-level semantic features. Then, a dictionary is made up of feature vectors of a reference image's all superpixels. The authors also include a regularization component in the energy function that emphasizes the locality to enhance the output. Bugeau et al. [11] exploited specific energy to solve the color selection and the spatial constraint problems. He et al. [12] proposed a fully automatic image colorization system that employs an end-to-end network that calculates the similarity between the reference image and the target image before the color transfer. Besides, to further reduce manual work, their image retrieval algorithm automatically suggests reference images by analyzing the luminance and semantic features. However, their technique is unable to compensate incorrect colors in less semantically significant areas or differentiate less semantic portions having identical local textures [12]. Compared with these approaches, our method further considers the semantic similarity between the target image and the reference image due to employing the match sub-network. Instead of using simple color mapping, we formulate the matching between the target and the reference as a sparse representation problem. We further introduce global-local affinity energy to constrain the global-semantic and local-structure self-similarity for the target image.

C. Learning-Based Colorization

These methods rely on large-scale image data to enforce networks to learn color distributions. Cheng et al. [14] combined a single neural network with a joint bilateral filtering to realize automatic colorization. Deshpande et al. [15] developed a quadratic objective function to train the network. Recently, several works based on Convolutional Neural Network (CNN) have achieved realistic colorization results. Iizuka et al. [16] merged local information with global priors, and Zhang et al. [17] solved this problem as a classification task to increase the diversity of the results. Larsson et al. [18] predicted each pixel color histogram via low-level and semantic representations. In addition, GANs can also be applied to multi-modal colorization. Isola et al. [22] learned a loss function to train the mapping from the target image to the resultant image. Yang et al. [20] utilized the GANs for 3D colorization, which learned a model to transfer a latent color parameter space to color space by a shape collection. The common shortcoming of the learn-based methods is that they only produce a single realistic color image for each target image and lack user interaction.

D. Hybrid Colorization

To further improve the quality of colorization, several works proposed hybrid frameworks which inherit the interactivity

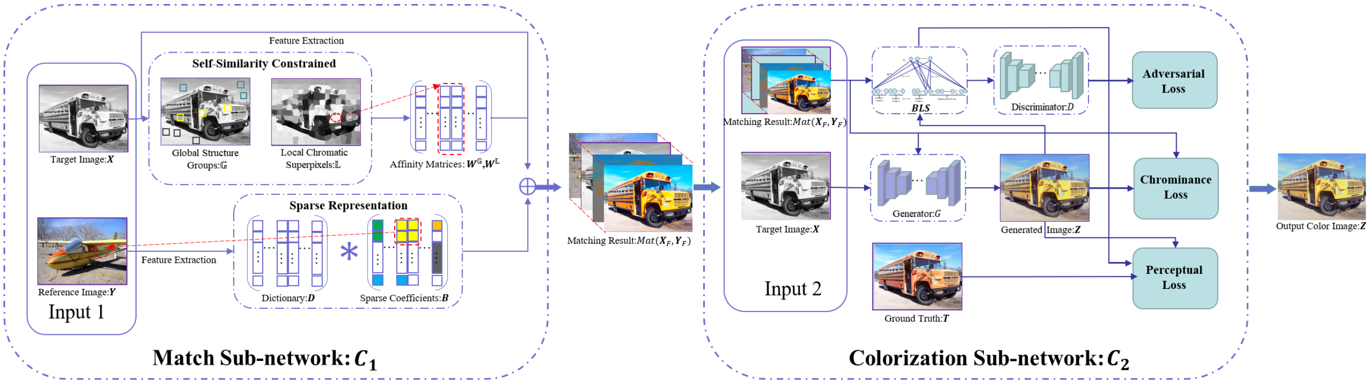


Fig. 2: The pipeline of the proposed colorization algorithm. The system consists of two sub-networks. The match sub-net $C1$ works as a pre-processing step. Taking the target grayscale image $\mathbf{X}$ and the reference image $\mathbf{Y}$ as inputs, it constructs the mappings from the reference image to the target image. The colorization sub-net is an end-to-end conditional broad-GAN that predicts the chrominance channels of the target image. It takes the target image $\mathbf{X}$, the reference image $\mathbf{Y}$ and the matching information as inputs, and outputs the resultant color image.

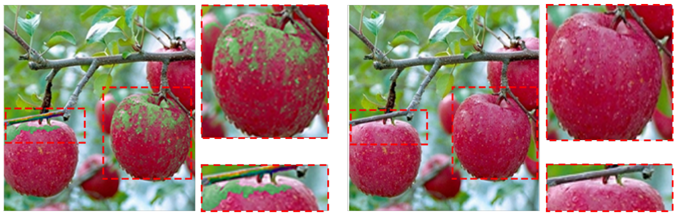
of scribble-based or example-based method and robustness of the learn-based method. Zhang et al. [19] utilized low-level hints and high-level semantic information to guide color transference. Sangkloy et al. [30] designed a feed-forward neural network that allows users to control the result in real-time. In [31], image colorization is defined as a multinomial classification problem. Xiao et al. [31] estimated the colors of the target grayscale image by analyzing the color distribution of the reference image. He et al. [12] defined the similarity metric between the reference image and target image as a hint information to encourage the network to propagate colors correctly. Analogously, our method combines the example-based method and the learn-based method. Given a target grayscale image and a reference image, we exploit the sparse representation approach to establish the mapping relations between them. Furthermore, a conditional broad-GAN is used to reconstruct the target grayscale image based on the match information. Motivated by the work [25], which utilized a reference image as additional information to guide the conditional GANs generate customized in-painting results, we propose a multi-task conditional broad-GAN to enforce the network produce semantically correct and visually delightful chromatic images.

To the best of our knowledge, our proposed technique is the first one to integrate the BLS and GANs in this manner to promote stability, and there is no such work devised in the literature before us.

III. Approach Overview

Our goal is to generate a plausible color image from a target grayscale image based on a color reference image, where some semantic content of the reference image may be partially related or completely unrelated to the target image. However, two major challenges must be solved to achieve this goal. First, it is difficult to establish a semantic map from reference image to target image, especially when the reference image is significantly unrelated to the target image in semantic content. Second, even if the map is established, it is still challenging to transfer suitable color and generate a plausible chromatic image.

To address the two challenges mentioned above, an end-to-end network architecture is proposed. This framework consists of two sub-networks that, respectively, address the semantic match and chromatic propagation between two input images. In this way, we decompose the complex colorization problem into two subproblems, each having a specific purpose. Our system uses the $CIE\ Lab$ color space, which has a luminance channel L and two chrominance channels a and b. Fig. 2 illustrates the system pipeline. The match sub-net takes a target grayscale image $\mathbf{X}_L \in \mathbb{R}^{H\times W\times 1}$ and a color reference image $\mathbf{Y}_{Lab} \in \mathbb{R}^{H\times W\times 3}$ as inputs, where H and W are the height and width of the input image. We observed that the patches with similar semantic features throughout the image and the pixels with close spatial distance in the local regions frequently have similar colors. In view of this, we formulate the similarity matching problem as a sparse representation problem. Global-local affinity energy is introduced to the match framework to constrain the sparse representation for global and local matching consistency. We utilize the pre-extracted features for semantic matching, and the matching result $Mat(\mathbf{X}_F, \mathbf{Y}_F)$(where $\mathbf{X}_F \in \mathbb{R}^{U\times N}$,$\mathbf{Y}_F \in \mathbb{R}^{U\times N}$ denote the features extracted from the given input target image and a reference image, N is the dimension of the feature vector, $U = H \times W$) will be used as one of the inputs of the colorization sub-net. The colorization sub-net, which consists of a generator G and a discriminator D, takes $\mathbf{X}_L \in \mathbb{R}^{H\times W\times 1}$, $\mathbf{Y}_{Lab} \in \mathbb{R}^{H\times W\times 3}$ and $Mat(\mathbf{X}_F, \mathbf{Y}_F)$ as inputs, and outputs a color image $\mathbf{Z}_{Lab} \in \mathbb{R}^{H\times W\times 3}$. In this stage, the extra reference information serves as a key 'hint' to guide appropriate color propagation. Furthermore, the BLS is utilized to assist in network training. Two loss functions are applied to this network. The chrominance loss function $\mathcal{L}_{chrom}$ measures the chrominance differences between the generated image $\mathbf{Z}_{Lab}$ and the reference image $\mathbf{Y}_{Lab}$. The perceptual loss function $\mathcal{L}_{perc}$ penalizes semantic content loss between $\mathbf{Z}_F$ and $\mathbf{T}_F$, where $\mathbf{Z}_F \in \mathbb{R}^{U\times N}$ and $\mathbf{T}_F \in \mathbb{R}^{U\times N}$



(a) (b)

Fig. 3: Demonstration of chrominance transfer results with (a) isolated sparse representation (Eq. (4)) and (b) our proposed global-structure and local-chromatic self-similarity constrained sparse representation (Eq. (1)).

denote the features extracted from the generated color image and the ground truth. Furthermore, users can get the desirable chromatic image by choosing different reference images.

IV. Match Sub-Network

Typically, in exemplar-based colorization methods, the network only considers the color distribution of the reference image and lacks semantic correspondence between the reference image and the target image [31]. Our work is inspired by the recent success of sparse representation in hyperspectral image super-resolution [28], which improves the performance and robustness dramatically. The complete pipeline of our match sub-network is illustrated in Fig. 2. Our main contribution is to propose a novel sparse representation method that ensures the matching consistency through global-local affinity energy of globally and locally self-similarity constraints. This novel method can be denoted as:

$$\mathbf{B}^* = arg\,\underset{\mathbf{B}}{min}\,\|\mathbf{X}_F - \mathbf{D}\mathbf{B}\|_F^2 + \alpha\eth(\mathbf{B}) + \mu\|\mathbf{B}\|_1 \tag{1}$$

where $\mathbf{X}_F$ is the feature vector containing the target images features that were extracted by the pre-trained VGG network, $\mathbf{B}$ is the coefficient matrix that reformulates the $\mathbf{X}_F$ by feature vectors from dictionary $\mathbf{D}$ that is composed of the feature vectors of the reference image, α, μ are weighting factors, $\|\mathbf{B}\|$ is the sparse constraint, and $\eth(\mathbf{B})$ is global-local affinity energy on the coefficient matrix. More details will be covered in the following two subsections. To get the result, both $\mathbf{X}_L$ and $\mathbf{Y}_{Lab}$ are fed into the match sub-network C_1, yielding the match information $Mat(\mathbf{X}_F, \mathbf{Y}_F)$:

$$Mat(\mathbf{X}_F, \mathbf{Y}_F) = C_1(\mathbf{X}_L, \mathbf{Y}_{Lab}, \mathbf{B}^*) \tag{2}$$

A. Formulation with Sparse Representation

In order to align the features between the target image and the reference image, we use the pre-trained VGG network [27] to extract high-level semantic features. Due to the large variety of semantic content in different images, learning a common dictionary for all target images tends to generate a dramatic matching error and essentially contradicts the exemplar-based methods. We instead learn the dictionary directly from the given reference image. The features, extracted from reference image by VGG network [27] comprise the dictionary $\mathbf{D} \in \mathbb{R}^{U\times N}$. Without considering the noise, the feature vectors $\mathbf{X}_F \in \mathbb{R}^{U\times N}$ containing the features extracted from the target image can be expressed as a linear transformation from $\mathbf{D}$ as:

$$\mathbf{X}_F = \mathbf{D}\mathbf{A} \tag{3}$$

where $\mathbf{A}$ is the coefficient matrix that reformulates the $\mathbf{X}_F$ by feature vectors from dictionary $\mathbf{D}$ without considering the noise. Therefore, given $\mathbf{D}$ and $\mathbf{X}_F$, the reconstruction coefficient matrix $\mathbf{B}$ can be estimated by minimizing the following error [32]:

$$\mathbf{B}^* = arg\,\underset{\mathbf{B}}{min}\,\|\mathbf{X}_F - \mathbf{D}\mathbf{B}\|_F^2 + \mu\|\mathbf{B}\|_1 \tag{4}$$

In general, we can reconstruct the features of the target image by Eq. (4), which establishes the map between the target image and the reference image. This matching information will be fed into the colorization sub-net serving as a ‘hint’. Therefore, users can obtain customized results by selecting varying reference images. Specifically, various reference images, which generate different dictionaries and sparse matrices, result in different sparse representations.

B. Global-Local Affinity Energy of Globally and Locally Self-Similarity Constraints

The sparse representation method discussed above achieves matching results between target features and reference features. In the task of exemplar-based image colorization, not only do we need to select the appropriate color to propagate, but the self-consistency of the target image is also important. However, Eq. (4) encodes the features of each pixel independently, which ignores the globally semantic structure and locally spatial structure consistency of target data. The isolated error matching by the traditional sparse representation methods results in unnatural chromatic images (see Fig. 3). Inspired by the work of [28] that suggests that a high-resolution hyperspectral image with a quality visual effect can be obtained by forcing the similarity of the sparse representations for pixels belonging to the same group and superpixel, we find that this method is also useful for image colorization. Based on this insight, two types of self-similarity constraints are introduced into Eq. (1):

- Global-semantic self-similarity constraint: Pixels, which belong to the patches with similar semantic structure, have a similarity of the sparse representations. These patches include both adjacent patches and non-adjacent patches.
- Local-structure self-similarity constraint: The sparse vectors for different pixels are similar in the local region. That means pixels with nearby spatial positions will have similar colors.

Based on the above two constraints, global-local affinity energy is introduced to reformulate the consistent matching. We use k-means [33] to cluster all similar patches and use superpixel segmentation method [34] to obtain superpixels. To realize global consistency, all similar patches form the global-semantic groups $\mathbb{G} = \{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_d\}$, which will have similar sparse representations. There are d groups in global-semantic groups, and $\mathbf{g}_d$ is the indexical vector composed

Algorithm 1 Image Matching with Self-similarity Constrained Sparse Representation

```
Input: Target Image $\mathbf{X}_L$, Reference Image $\mathbf{Y}_{Lab}$
Output: Matching Result:$Mat(\mathbf{X}_F, \mathbf{Y}_F)$
extract features from target image $\mathbf{X}_L$ and reference image $\mathbf{Y}_L$ by pre-trained VGG to form $\mathbf{X}_F$, $\mathbf{Y}_F$, and the dictionary $\mathbf{D} = \mathbf{Y}_F$;
build global-semantic groups $\mathbb{G} = \{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_d\}$ and local-structure superpixels $\mathbb{L} = \{\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_q\}$;
calculate the affinity matrices $\mathbf{W}^{\mathbb{G}}$ and $\mathbf{W}^{\mathbb{L}}$:
for each global-semantic weight $w^{\mathbb{G}}_{(n,i)}$ of global affinity matrix $\mathbf{W}^{\mathbb{G}}$; $n \leq N$ do
    compute the global-semantic weight with Eq. (6)
end for
for each local-structure weight $w^{\mathbb{L}}_{(n,j)}$ of local affinity matrix $\mathbf{W}^{\mathbb{L}}$; $n \leq N$ do
    compute the local-structure weight with Eq. (8)
end for
solve Eq. (1) with the dictionary $\mathbf{D}$ and the affinity matrices $\mathbf{W}^{\mathbb{G}}$ and $\mathbf{W}^{\mathbb{L}}$;
```

by superpixel indices of the d-th group. Likewise, similar sparse representations will be enforced for these pixels within the same superpixel. Local-structure self-similarity constraint is represented by superpixels $\mathbb{L} = \{\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_q\}$ (in total q superpixels), and $\mathbf{l}_q$ is also an indexical vector formed by pixel indices of q-th superpixels. Based on the above analysis, a weighted mean of the sparse matrix for all pixels in the same global group or in the same superpixel can be used to represent the sparse vector of a pixel. Therefore, the global-local affinity energy is defined as follows:

$$\mathbf{b}_n = \lambda \sum_{i \in \mathbf{g}_d} w^{\mathbb{G}}_{(n,i)} \mathbf{b}_i + (1 - \lambda) \sum_{j \in \mathbf{l}_q} w^{\mathbb{L}}_{(n,j)} \mathbf{b}_j \quad \text{with } n \in \mathbf{g}_d \wedge n \in \mathbf{l}_q \tag{5}$$

where $\mathbf{b}_n$ is the n-th sparse vector of sparse matrix $\mathbf{B}$, $w^{\mathbb{G}}_{(n,i)}$ is the global-semantic weight, which constrains the similarity between the i-th sparse vector $\mathbf{b}_i$ belonging to the same global-semantic group and sparse vector $\mathbf{b}_n$. Homoplastically, $w^{\mathbb{L}}_{(n,j)}$ is the local-structure weight, which constrains the similarity between j-th sparse vector $\mathbf{b}_j$ belonging to the same superpixels and sparse vector $\mathbf{b}_n$. And λ is a parameter used to maintain the balance between group-semantic and local-structure self-similarity constraints.

Specifically, $w^{\mathbb{G}}_{(n,i)}$ measures the semantic similarity between pixels within the same global-semantic group, and we calculate $w^{\mathbb{G}}_{(n,i)}$ as follows:

$$w^{\mathbb{G}}_{(n,i)} = \frac{1}{\hbar^{\mathbb{G}}_n} \exp\{-\frac{\|\mathbf{f}_n - \mathbf{f}_i\|^2}{\rho^2_{\mathbb{G}}}\} \tag{6}$$

where $\mathbf{f}_n$, $\mathbf{f}_i$ are, respectively, the feature vector of n-th pixel and i-th pixel, $\rho_{\mathbb{G}}$ is a scale parameter for the semantic measure, and $\hbar^{\mathbb{G}}_n$ is a normalization term guaranteeing that $\sum_i w^{\mathbb{G}}_{(n,i)} = 1$, and is defined as:

$$\hbar^{\mathbb{G}}_n = \sum_{i \in g_d} \exp\{-\frac{\|\mathbf{f}_n - \mathbf{f}_i\|^2}{\rho^2_{\mathbb{G}}}\} \tag{7}$$

Analogously, $w^{\mathbb{L}}_{(n,j)}$ measures the spatial distance between pixels within the same superpixel. It is defined in the exactly same format, as:

$$w^{\mathbb{L}}_{(n,j)} = \frac{1}{\hbar^{\mathbb{L}}_n} \exp\{-\frac{\|p_n - p_j\|^2}{\rho^2_{\mathbb{L}}}\} \tag{8}$$

where, p_n, p_j are, respectively, the spatial location of n-th pixel and i-th pixel, $\rho_{\mathbb{L}}$ is scale parameter for distance measure, and $\hbar^{\mathbb{L}}_n$ has analogous definition as Eq. (7).

Finally, the global-local affinity energy is formulated as:

$$\eth(\mathbf{B}) = \|\mathbf{B} - \lambda \mathbf{W}^{\mathbb{G}} \mathbf{B} - (1 - \lambda) \mathbf{W}^{\mathbb{L}} \mathbf{B}\|^2_F \tag{9}$$

where $\mathbf{W}^{\mathbb{G}} \in R^{N \times N}$ and $\mathbf{W}^{\mathbb{L}} \in \mathbb{R}^{N \times N}$ are global and local affinity matrices that can be computed by Eq. (6) and Eq. (8).

We summarize the overall matching algorithm between the target image and the reference image with global-semantic and local-structure self-similarity constrained sparse representation in Algorithm 1. Firstly, we extract the features of the target image and the reference image by pre-trained VGG [27], and the features from the reference image constitute the dictionary. Then, we create the global-semantic groups and local-structure superpixels to calculate the global and local affinity matrices. Finally, the matching can be obtained by solving Eq. (1).

V. Colorization Sub-Network

We divided the complex problem of colorization into two parts. The match sub-network solves the matching problem between the target image and the reference image. And the purpose of the colorization sub-network is to selectively propagate and predict colors correctly through the ‘hints’ of the match sub-network to complete the colorization of the target image. As shown in the right-side part of Fig. 2, the colorization sub-network C_2 takes target grayscale image $\mathbf{X}_L$, reference image $\mathbf{Y}_{Lab}$, and match information $Mat(\mathbf{X}_F, \mathbf{Y}_F)$ as inputs, and outputs color target image $\mathbf{Z}_{Lab}$:

$$\mathbf{Z}_{Lab} = C_2(\mathbf{X}_L, \mathbf{Y}_{Lab}, Mat(\mathbf{X}_F, \mathbf{Y}_F)) \tag{10}$$

Considering that GANs have produced high-quality colorization results, we propose using the reference image as the exemplar information of the conditional GAN to perform the colorization task. For our experiments, the model architecture consists of two networks. One is the generator G that is used to colorize the target grayscale image, and the other is the discriminator D that is used to judge whether the image is from the ground truth image or synthetic. To retain the semantic content of the target image, we utilize the BLS to extract the semantic features of the reference image and the target image, which helps the discriminator distinguish images at the semantic level. In the training process, for each target image x_i in the training set $X = \{x_1, x_2, ..., x_k\}$, there exists a corresponding reference image y_i in the reference set $Y = \{y_1, y_2, ..., y_k\}$, where k is the number of image pairs in the training set. We train our network with a large number of image pairs to achieve the desired results.

In this work, we introduce two loss functions to solve the exemplar-based colorization:

$$\begin{aligned}(G^*, D^*) =& arg \min_{G} \max_{D} (\mathcal{L}_{cGAN}(G, D) \\ &+ \varepsilon \mathcal{L}_{chrom}(G) + \delta \mathcal{L}_{perc}(G))\end{aligned} \tag{11}$$

where $\mathcal{L}_{cGAN}$ denotes the adversarial loss function, $\mathcal{L}_{chrom}$ and $\mathcal{L}_{perc}$ are the chrominance and perceptual loss functions, respectively, ε and δ are relative weights. We use Eq. (11) to solve the min-max problem. We present the details of our proposed two loss functions in the following subsections.

A. Broad Learning System

Given a colorized image $\mathbf{Z}_F \in \mathbb{R}^{U \times N}$ and a ground truth image $\mathbf{T}_F \in \mathbb{R}^{U \times N}$, which is preprocessed by the pre-trained VGG [27] with U samples, each of N dimensions, $\hat{\mathbf{Z}}_F$ and $\hat{\mathbf{T}}_F$ are the output matrices that belong to $\mathbb{R}^{U \times C}$ (C is the dimension of broad features processed by the BLS), where $\hat{\mathbf{Z}}_F$ and $\hat{\mathbf{T}}_F$ denote the broad feature matrices of the generated image and the ground truth. The feature information of the output matrices is composed of mapped features and enhancement nodes. For U feature mappings, the image's mapping feature nodes can be obtained by the following equation:

$$\mathbf{E}_m = \phi(\mathbf{Z}\mathbf{W}_{em} + \boldsymbol{\beta}_{em}) \qquad m = 1, 2, ..., U \tag{12}$$

where $\mathbf{W}_{em}$ and $\boldsymbol{\beta}_{em}$ are generated with proper dimensions randomly corresponding to the weights and bias of the mth feature map, ϕ is the mapping function. All feature nodes are constructed as:

$$\mathbf{E}^m \equiv (\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_m) \tag{13}$$

The rth group of enhancement nodes is defined below:

$$\mathbf{H}_r = \psi(\mathbf{E}_r \mathbf{W}_{hr} + \boldsymbol{\beta}_{hr}) \qquad r = 1, 2, ..., U \tag{14}$$

where $\mathbf{W}_{hr}$ and $\boldsymbol{\beta}_{hr}$ are generated with proper dimensions randomly corresponding to the weights and bias of the rth group of enhancement nodes, ψ is the mapping function similar to ϕ. Analogously, all enhancement nodes are constructed as:

$$\mathbf{H}^r \equiv (\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_r) \tag{15}$$

Hence the colorized broad features can be denoted as:

$$\begin{aligned}\hat{\mathbf{Z}}_F &= [\mathbf{E}_1, \mathbf{E}_2, ..., \mathbf{E}_m | \mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_r]\mathbf{W}^r \\ &= [\mathbf{E}^m, \mathbf{H}^r]\mathbf{W}^r\end{aligned} \tag{16}$$

where $\mathbf{W}^r$ are the connecting weights for the broad network, here $\mathbf{W}^r = [\mathbf{E}^m, \mathbf{H}^r]^+ \mathbf{F}_Z$, which can be calculated by ridge regression approximation of $[\mathbf{E}^m, \mathbf{H}^r]^+$ [29]. The definition of $\hat{\mathbf{T}_F}$ is exactly in the same format. Fig. 4 shows the network of the broad learning system described above.

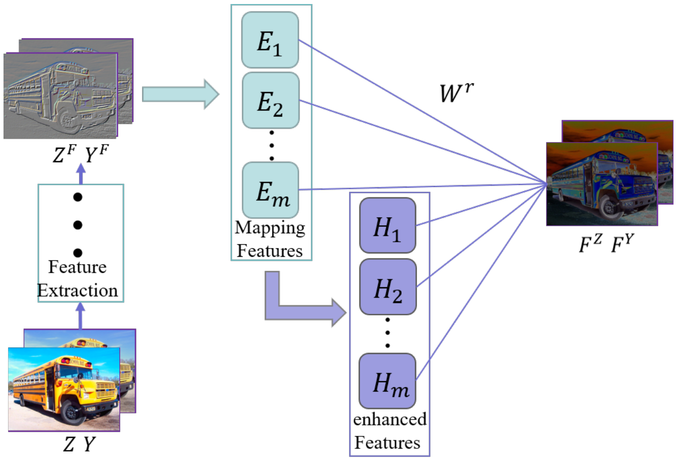


Fig. 4: The structure of our BLS. The input image (generated image Z and reference image Y) is pre-processed by the feature extraction and then convoluted to get the mapped features. Then, these mapped features are convoluted once again to obtain the enhancement features. Finally, all features are concatenated together to form the output F^Z F^Y of the flat network.

B. Feature-Aware Conditional Broad-GAN

The original GAN was considered unstable for two reasons: firstly, it trains two adversarial neural networks for contradictory purposes; secondly, the input image can be mapped to any potential image. For exemplar-based image colorization, we desire the output image's chrominance to be as similar as possible to the reference image. To reduce the space of possible mapping functions, we introduce feature-aware conditional broad-GAN that guides the generator G to predict colors by analyzing the semantic features of the reference image. G learns to generate a color image that fools the discriminator D. For our experiments, we used the network presented in [31] as our generator. To generate colorful and visually pleasant images, Xiao et al. [31] leveraged the inherent multi-scale pyramid structure, while our framework uses only a single-level network structure. The inputs to the generator G are the target grayscale image $\mathbf{X}_L$, the reference image $\mathbf{Y}_{Lab}$, and $Mat(\mathbf{X}_F, \mathbf{Y}_F)$.

Compared to the generator, the discriminator judges whether a target image is colorized or not. However, generally, GANs suffer from the model collapse problem [35], where the target image tends to map the color distribution but ignores the semantic features. More details are shown in Section VI-B. Therefore, the semantic features of the resulting color image are used as a criterion for the discriminator to judge the quality of colorization. Furthermore, we discover that the BLS can extract the semantic information of images efficiently. In this paper, the mapped features and the enhancement features are concatenated to construct the feature nodes and the enhancement nodes, respectively. Then, these two types of nodes are merged to output the resulting matrices. Finally, the output matrices will be fed into the traditional discriminator as the basis for judgment.

Algorithm 2 Image Colorization with Conditional Broad-GAN

```
Input: Grayscale Image $\mathbf{X}_L$, Reference Image $\mathbf{Y}_{Lab}$, and
    Matching Information $Mat(\mathbf{X}_F, \mathbf{Y}_F)$
Output: Colorized Target Image $\mathbf{Z}_{Lab}$
1: calculate the adversarial loss $\mathcal{L}_{cGAN}$ using Eq. (17);
2: compute the chrominance loss $\mathcal{L}_{chrom}$ using Eq. (18);
3: extract the feature maps $\hat{\mathbf{Z}}_F$ and $\hat{\mathbf{T}}_F$ of the colorized image
    and the ground truth image, respectively, through BLS;
4: using Eq. (19), calculate the perceptual loss $\mathcal{L}_{perc}$ with
    the help of features extracted through BLS;
5: use Eq. (11) to solve the min-max problem;
```

C. Loss Function

The objective of previous colorization works was to encourage the target color image $\mathbf{Z}_{Lab}$ to be as similar as possible to the ground truth image $\mathbf{T}_{Lab}$. However, it is inherently contradictory to exemplar-based colorization since the target color image should be perceived as being consistent with the reference image. Thus, it is not right to directly reduce the difference between $\mathbf{Z}_{Lab}$ and $\mathbf{T}_{Lab}$. Therefore, based on the above analysis, the objective of our cGAN is still to learn how to transform a grayscale image into a color image with the 'hint' of a reference image, and that learning objective is defined as:

$$\begin{aligned}\mathcal{L}_{cGAN} =&\mathbb{E}_{x_k, y_k \backsim p_{data}(x,y)}[logD(x_k, y_k)] \\ &+ \mathbb{E}_{y_k \backsim p_y, G(z_k) \backsim p_z}[log(1 - D(G(z_k, y_k)))]\end{aligned} \quad (17)$$

Nevertheless, two loss functions specific to the reference-based method can also be added to the network. There are two primary considerations in designing the loss function: first, our output image should be as close as possible to the reference image to get a customized result; second, we enforce the output color image to be as natural as possible, even when the reference image is extremely unrelated to the target. To achieve these goals, we need to ensure that appropriate reference colors are used for the input grayscale image while ensuring that the resulting color image retains the semantic content of the input target image. Thus, we adopt a multi-task learning framework that uses the same network and weights to train both the chrominance and perceptual branches. The two branches have their own input and loss functions for different purposes.

In the chrominance branch, the objective of the network is to learn to selectively extract colors from the reference image and apply them to the target input image. In the preliminary stage of network training, we use the ground truth image as a reference image and feed it into the network. However, this method is not actually an exemplar-based approach. Therefore, we leverage the proposed sparse representation method to reconstruct a reference image $\mathbf{X}_{Lab}^{'}$ from the ground truth image. Hence, reconstructed ground truth image $\mathbf{X}_{Lab}^{'}$ during the training stage replaces the reference image $\mathbf{Y}_{Lab}$ in the training stage. Under the guidance of the reference image $\mathbf{X}_{Lab}^{'}$, the grayscale image $\mathbf{X}_L$ is colorized if the network chooses suitable colors. There, we use a simple loss function [22]:

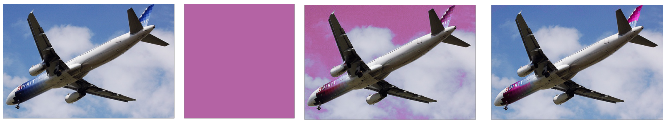

(a) Ground truth (b) Reference (c) Colorized result 1 (d) Colorized result 2

Fig. 5: Visualization of perceptual loss. Both colorized results have the same reference image, but the unnatural pink sky (c) is colorized by minimizing the chrominance loss only. A more plausible color (d) is colorized by minimizing both the chrominance loss and the perceptual loss.

$$\mathcal{L}_{chrom} = \mathbb{E}_{x_k, x_k^{'} \backsim p_{data}(x,x^{'}), z_k \backsim p_z}[\|G(x_k, z_k) - x_k^{'}\|_1] \quad (18)$$

where, we use $L1$ distance to enforce that the target image and the reference image are as similar as possible in terms of color distribution.

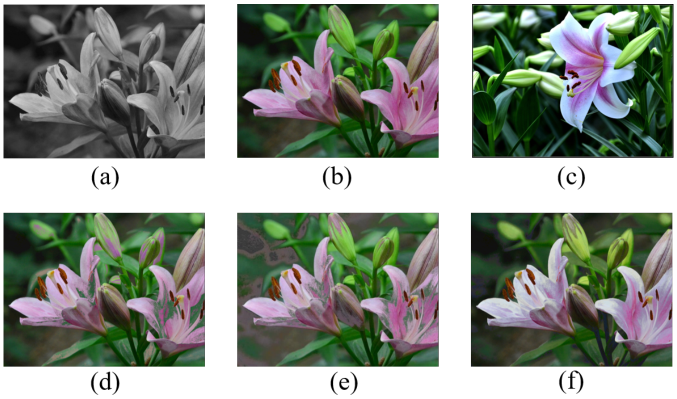
However, L_{chrom} only works when the target image and the reference image have significant similarities in semantic content. To encourage the network to get a perceptually reasonable color image even if there is no suitable reference image, we propose a perceptual branch. The perceptual loss minimizes the semantic differences of target image caused by unrelated reference image and improves the robustness of appearance differences, as shown in Fig. 5. In this branch, we adopt the feature maps in the BLS, which has been demonstrated to have good classification ability. Accordingly, we define the perceptual loss as:

$$\mathcal{L}_{perc} = \mathbb{E}_{p_t \backsim p_{data(p)}}[\|\hat{\mathbf{Z}}_F(p_t) - \hat{\mathbf{T}}_F(p_t)\|]_F^2 \quad (19)$$

where, $\hat{\mathbf{Z}_F}$ and $\hat{\mathbf{T}}_F$ refer to the feature maps of the colorized image and the ground truth image extracted from BLS, and p_t is the t-th pixel of the resulting color image or the ground truth image. The broad-discriminator is encouraged to distinguish whether the resulting image maintains the semantic similarity. Therefore, $\mathcal{L}_{perc}$ aims to minimize the semantic difference between the ground truth image and the colorized image, which is robust to the images with differences in appearance. We have also described the working of the colorization sub-net in Algorithm 2.

VI. Experimental Results

In this section, we show our experimental results on images with different semantic contents. Note that our network is trained using the Places Database [36] that contains 205 scene categories and approximately 2.5 million images and Pascal VOC data sets [37]. We take about 600,000 image pairs for our model training and 25,000 for model testing. The image categories contain popular categories, such as architectures, forest, people, vehicles, scenery, animals. To generate visually plausible color images for any reference image, the pairs are composed of images with different extents of similarity. Practically, 20% of reference images are the ground truth of target images, 50% of image pairs have significant similarity by our artificial selection, and the remaining 30% of image



(a) (b) (c)

(d) (e) (f)

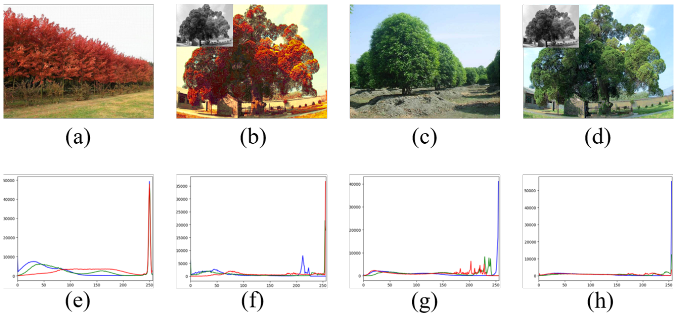
Fig. 6: The influence of parameters $\rho_{\mathbb{G}}$, $\rho_{\mathbb{L}}$ with different values. (a) target grayscale image, (b) ground truth, (c) reference image, (d)–(f) resulting color images corresponding to different parameters. (d) $\rho_{\mathbb{G}} = 1$, $\rho_{\mathbb{L}} = 10^{-8}$. (e) $\rho_{\mathbb{G}} = 10^{-8}$, $\rho_{\mathbb{L}} = 10$ (d) $\rho_{\mathbb{G}} = 1$, $\rho_{\mathbb{L}} = 10$

pairs are randomly selected within the same category. As shown in Fig. 1, we can get customized results easily by selecting different reference images. In addition, we evaluate our method qualitatively and quantitatively in the following aspects: first, we perform experiments to detail the role of the match sub-net; second, we set up a single-branch framework experiment and visualize the functions of the chrominance branch and perceptual branch. Then, we train our colorization model and compare it with several state-of-the-art methods; and lastly, we utilize a user study to validate our approach and quantitatively evaluate users' subjective preferences.

A. Role of Match Sub-Net

Our end-to-end network architecture includes two sub-nets: the match sub-net and the colorization sub-net. The match sub-net measures the semantic similarity between the reference image and the target image. And the colorization sub-net transfers colors from the reference image to the grayscale image. The match sub-net defines as a preprocessing step to provide the basis for the color transformation of the colorization sub-net. We analyze the effect of the match sub-net by fine-tuning the experiment framework.

For feature extraction, we have used the pre-trained VGG model [27] that was trained on the RGB color space. All the other details, like normalization and parameter settings, can be found in the original paper [27]. The parameter λ balances the weights between global-semantic and local-structure self-similarity constraints. Here $\lambda = 0.5$ is used, which means global constraint and local constraint are equally important. Besides, $\rho_{\mathbb{G}}$ determines the feature similarity globally, and $\rho_{\mathbb{L}}$ determines the spatial consistency locally. To show the effect of parameter adjustment on the results, we compared the distinctions when $\rho_{\mathbb{G}} = 1$, $\rho_{\mathbb{L}} = 10^{-8}$, $\rho_{\mathbb{G}} = 1$, $\rho_{\mathbb{L}} = 10$, and $\rho_{\mathbb{G}} = 10^{-8}$, $\rho_{\mathbb{L}} = 10$, respectively. Fig. 6 shows the colorized results with different parameters. In extreme cases, the local constraint is almost ignored in Eq. (9) when we fixed $\rho_{\mathbb{L}}$ close to 0 in Eq. (8). Fig. 6(d) shows that many local patches were miscolored when we neglected the local-chromatic constraint. For example, many isolated patches of pink flowers are mismatched to the green leaves. At the other extreme, when $\rho_{\mathbb{G}}$ is close to 0 in Eq. (6), the global term is ignored, and the self-similarity constraint is determined by the local term in Eq. (9). Fig. 6(e) shows that many superpixels with similar semantics do not match similar colors, resulting in a global inconsistency. Lastly, as shown in Fig. 6(f), the experiments demonstrate that these suitable parameters exhibit better performance. For the rest of the experiments, the parameters are set as global-semantic similarity measure $\rho_{\mathbb{G}} = 1$, local-structure similarity measure $\rho_{\mathbb{L}} = 100$, weighting factor $\lambda = 0.5$, balance factors $\alpha = 0.0001$, and $\beta = 0.025$.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 7: Visualization of color selection in the chrominance branch. (a) and (c) different reference images, (b) and (d) colorized images corresponding to different reference images, (e)–(h) corresponding color histograms.

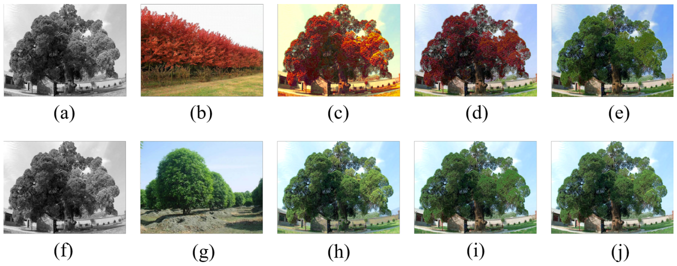
B. Chrominance Branch and Perceptual Branch

Our colorization sub-net is created by the exemplar-based conditional broad-GAN that learns to select suitable color samples, propagate colors, and predict colors. To accomplish this effectively, we have designed a multi-task learning framework, including chrominance branch and perceptual branch. Each branch serves a distinct purpose by minimizing a diverse loss objective. To learn the influence of chrominance branch on the colorization sub-net, we only train the chrominance branch of the C_2 by setting $\varepsilon = 100$, $\delta = 0$ in Eq. (11). As shown in Fig. 7, we visualize it through an example to learn its role intuitively. The colorized results with diverse reference images are shown in Fig. 7(b) and Fig. 7(d). In order to estimate the color transformation performance between the reference image and the target image intuitively, we plot the color histogram to visualize the results in Fig. 7(e)–Fig. 7(h). As we can see, the histograms of the resulting images are very close to the reference images. This illustrates that our network can catch the color information from the reference image and apply it to the whole target image.

To understand the effect of perceptual branch, we only train the perceptual branch of the C_2 by setting $\varepsilon = 0$, $\delta = 0.5$ in Eq. (11). As shown in Fig. 8, we use the same image set to compare the roles of chrominance loss and perceptual loss. Fig. 8(e) and Fig. 8(j) demonstrate that the perceptual branch predicts the colors by semantic features of the target image from the large-scale data, e.g., the leaves

TABLE I: Evaluation of each component of our method.

Component	-Global	-Local	-Chrom	-Perc	**+All**
MEAN PSNR	27.78	26.95	26.43	28.13	28.87
Drop Rate	-3.78%	-6.65%	-8.45%	-2.56%	-


(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Fig. 8: Comparison of results from the training with different branch configurations. (a) and (f) target grayscale image, (b) and (g) different reference images, (c)–(e) and (h)–(j) resulting color images corresponding to different loss functions. (c) and (h) chrominance loss only ($\varepsilon = 100$, $\delta = 0$), (e) and (j) perceptual loss only ($\varepsilon = 0$, $\delta = 0.5$), (d) and (i) both loss functions ($\varepsilon = 100$, $\delta = 0.3$).

are colorized with green no matter what color the leaves are in the reference image. Therefore, the perceptual loss controls the image colorization by semantic features, which improves the mismatch caused by chrominance loss only. Considering the complementarity of the two branches, we minimize both loss functions simultaneously by leveraging a multi-task learning network. Hence, we train both branches with the objective in Eq. (11). We estimate the effect of varying weights for the colorization results in Fig. 8(d) and Fig. 8(i). This indicates that our network transfers colors when there exist well-matched patches between the target image and the reference image but predicts color from the large-scale database when the reference image and the target image are highly unrelated. In this paper, we set $\varepsilon = 100$, $\delta = 0.3$ by default after performing many experiments.

C. Evaluation of Model Components

We analyzed the effectiveness of different components by comparing the peak-signal-to-noise ratio (PSNR) and structural similarity index (SSIM) performances [39], when one of the model components is discarded. The components are as follows: global-semantic self-similarity constraint, local-structure self-similarity constraint, chrominance branch, and perceptual branch. Table I and Fig. 9 show the impact on the colorization performance of our approach when lacking one of the model structures. As we can see, each component of our proposed method plays an individual positive role in colorization performance. As shown in Table I, the first row indicates the average PSNR (dB) performance of colorized images on the whole test set when one of the network elements is not used. The second row presents the drop rate of reconstruction performance compared with the results obtained when all components are integrated together. The last column shows the performance of the complete architecture. From these results, it is evident that we achieve the best performance when combining all components together. As shown in Fig. 9, the components can be sorted by importance as: chrominance branch, local-structure self-similarity constraint, global-semantic self-similarity constraint, and perceptual branch.

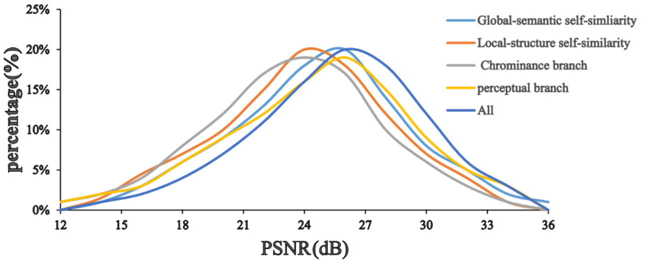
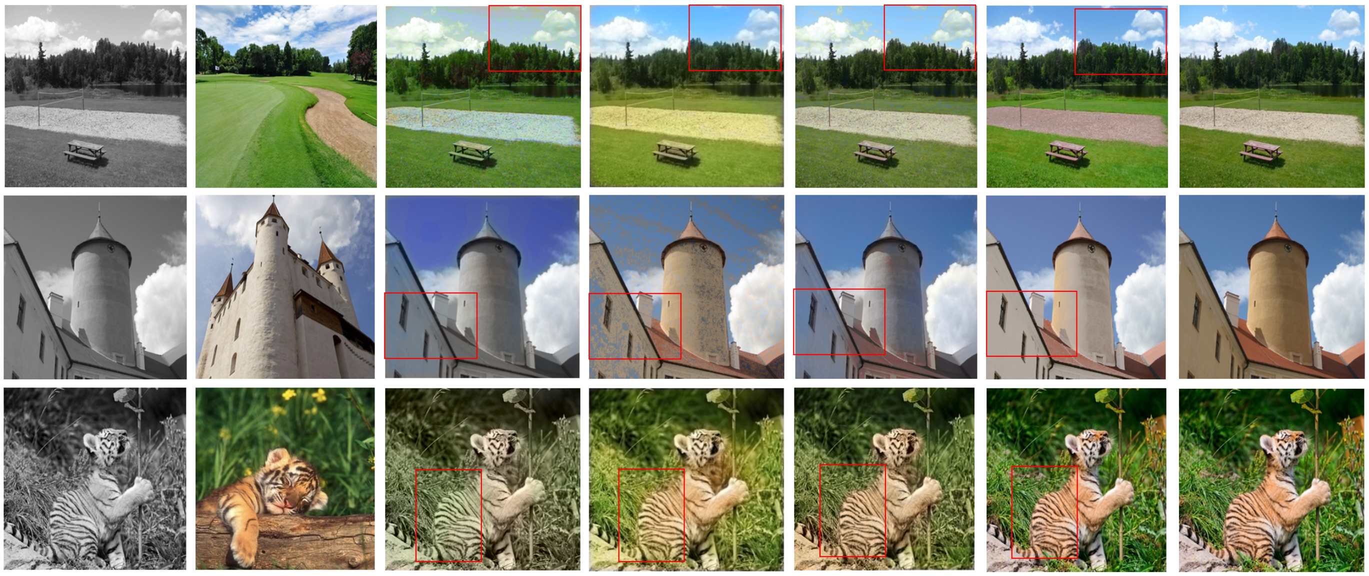


Fig. 9: The PSNR distribution with/without each model component. The proposed method achieves the best performance when integrating all components together.

D. Comparison with Existing Colorization Methods

To evaluate the performance of our method visually, we compare the resulting images of our algorithm with existing image colorization methods [8], [11], [16]–[18], [38]. The methods [8], [11], [38] are exemplar-based methods that propagate chrominance information from the reference image to the target image, and methods [16]–[18] are learning-based techniques that predict the color distribution of grayscale image by analyzing a large number of images from image database. To ensure fair comparison, we extract the colorization results from their papers or run their attainable code.

We compare color images generated from our proposed method and other exemplar-based methods [8], [11], [38]. Several results are shown in Fig. 10. Gupta et al. [8] leveraged the features that were extracted from the reference image and the target at the level of superpixels to guide the colorization process, and added an image space voting mechanism to enforce the spatial coherence simultaneously. However, there is still a great number of mismatches between the target image and the reference image. As seen in the 2nd column of Fig. 10, some part of the tiger's skin is mismatched to the green grass. Both [11] and [38] see the color reconstruction as a problem that selects the most appropriate color from the reference image. To obtain the best color candidates and enforce spatial constraint, a variational framework is proposed in [11]. The method [11] ignores the relation between luminance and chrominance that leads to a de-saturation effect of the colorized image. Pierre et al. [38] introduced a novel regularization term that is capable of considering both of the channels of luminance and chrominances. However, for both algorithms [11] and [38], global structure constraint and local chrominance consistency are disregarded in the process of selecting suitable colors (e.g., the 3rd and 4th column of Fig. 10). In comparison, our method can not only choose the most appropriate color candidates according to the semantic guidance but also guarantees global and local consistency by adding a constraint term in the sparse representation. As shown



(a) Gray image (b) Reference image (c) Gupta et al. (d) Bugeau et al. (e) Pierre et al. (f) Proposed (g) Ground truth

Fig. 10: Comparison of our colorization results with exemplar-based methods [8], [11], [38]. Visually, our results look more natural than that of Pierre et al. [38] and are closer to the ground truth. In addition, numerous isolated error matching patches of Gupta et al. [8] and Bugeau et al. [11] result in really poor colorized images. In comparison, the colorization results of the proposed method have a higher semantic consistency. Besides, the de-saturation effect of the colorized image of the method [11] still leads to artifacts of color inconsistency.

in Fig. 10, our method obtains a superior visual quality than the other techniques. For example, in the last low of Fig. 10, walls in the first three methods' results are mismatched with the blue sky, but this is not the case with our approach.

Then, we compare the performance with learning-based methods [16]–[18]. As shown in Fig. 11, the colorization results of techniques [16]–[18] are automatically generated based on a large-scale database, but ours are generated based on a reference image. Generally, plausible colorized results are generated by these methods. Based on Convolutional Neural Networks, Iizuka et al. [16] proposed a hidden layer that is able to merge global priors and local features. However, as shown in the 3rd column of Fig. 11, large areas with complex textures are not colorized correctly. For example, some of the flowers in the 3rd row are still gray. Zhang et al. [17] took grayscale image colorization as a classification task and enriched the color gradients by class-rebalancing in the training process. Nonetheless, as we can see in the 4th column of Fig. 11, a yellow patch appears in the center of the image, which affects the image colorization quality. Larsson et al. [18] utilized both low-level and semantic features to generate realistic images but failed to control the local consistency. Furthermore, such learning-based methods only generate a single result and cannot achieve customized results. Comparatively, our approach merges the advantages of both exemplar-based and learning-based approaches. As seen in the 6th column of Fig. 11, our network colorizes the target image under the guidance of the reference image but predicts the colors based on the semantic features when the reference image lacks corresponding patches. However, there are still some limitations because of the global-semantic and local-structure self-similarity constraints. As shown in the last row of Fig. 11, the information of flowers' red color was lost during the image reconstruction process, but the colorized image is still semantically meaningful.

TABLE II: Quantitative evaluations on different resolutions with the ground truth as the reference image.

Method	PSNR (dB)			SSIM		
	256^2	512^2	1024^2	256^2	512^2	1024^2
Gupta et al. [8]	24.95	24.78	24.32	0.72	0.70	0.68
Bugeau et al. [11]	26.32	26.11	25.87	0.81	0.79	0.73
Pierre et al. [38]	27.59	27.23	26.65	0.86	0.83	0.80
Iizuka et al. [16]	27.58	27.11	26.59	0.80	0.79	0.76
Zhang et al. [17]	28.33	27.98	27.34	0.87	0.85	0.81
Larsson et al. [18]	28.59	28.34	28.21	0.85	0.84	0.81
Our proposed	**28.98**	**28.66**	**28.47**	**0.89**	**0.85**	**0.83**

We conduct another experiment to further evaluate the effectiveness of our colorization architecture. In this experiment, we compare the PSNR and SSIM performances of the generated results with both exemplar-based and learning-based methods. Note that, the techniques [8], [11], [38] and ours take the ground truth as the reference image. Besides, all methods have the same input grayscale image. Therefore, we evaluate the PSNR and SSIM of colorized results directly by comparing them with their ground truth images. Table II presents the PSNR and SSIM results. The reason why method [38] achieves higher PSNR and SSIM than the other two exemplar-based methods is that the approach of [38] couples luminance and chrominances channels. Different from exemplar-based method [8], [11], [38], the learning-based

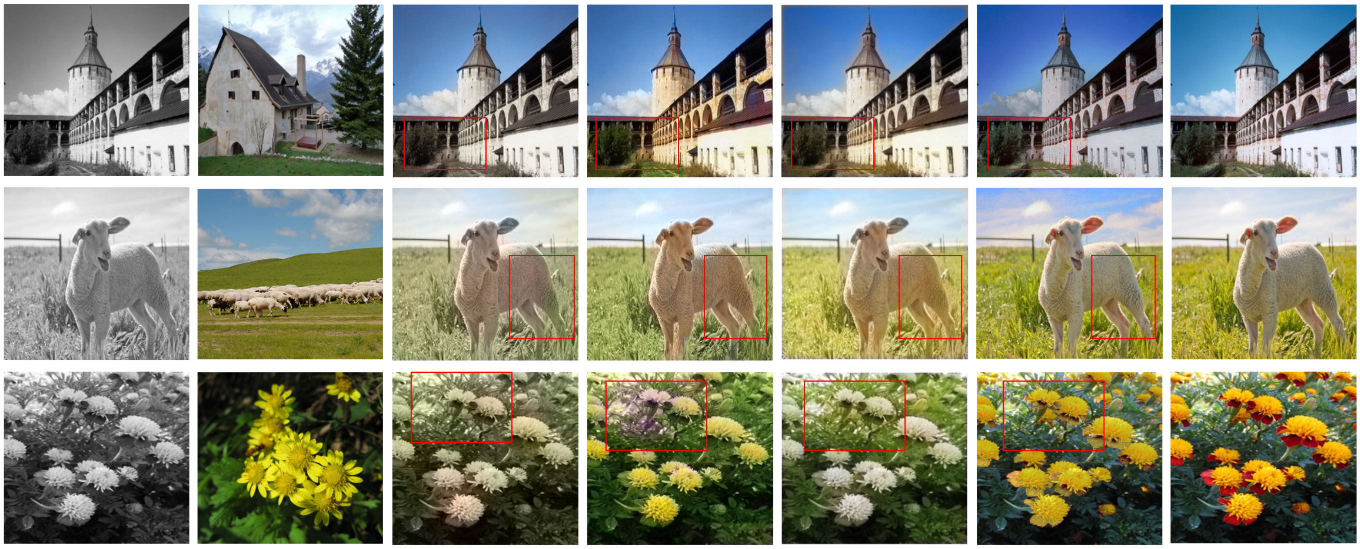

(a) Gray image (b) Reference image (c) Iizuka et al. (d) Zhang et al. (e) Larsson et al. (f) Proposed (g) Ground truth

Fig. 11: Comparison of our colorization results with the output of previous learning-based methods [16]–[18]. In general, the existing approaches being compared generate reasonable colorized images. However, there are still some limitations exhibited by them. For instance, some parts of grass and sky are colorized in gray by Iizuka et al. [16] in the 3rd column, an incorrect yellow patch appears at steamship by Zhang et al. [17] in the 4th column, and a region of snow mountain is colorized in blue by Larsson et al. [18]. In comparison, our method can generate plausible images by semantic guidance.

TABLE III: The comparison of HIS scores of our proposed technique with other state-of-the-art algorithms.

Gupta et al. [8]	Bugeau et al. [11]	Pierre et al. [38]	Iizuka et al. [16]	Zhang et al. [17]	Larsson et al. [18]	**Our proposed**
0.42	0.46	0.48	0.57	0.63	0.66	**0.68**

TABLE IV: Inference time (in seconds) of our approach in comparison to the state-of-the-art methods.

Gupta et al. [8]	Bugeau et al. [11]	Pierre et al. [38]	Iizuka et al. [16]	Zhang et al. [17]	Larsson et al. [18]	**Our proposed**
–	–	1.09	0.41	**0.12**	0.57	0.35

methods [16]–[18] get performance improvement due to the utilization of a learning system. However, as we can see, our approach obtains the highest performance as compared to these algorithms. In addition, the PSNR and SSIM of our method have little changes on different resolutions. That is because our algorithm can find out the correspondent semantic patches between the reference image and the target grayscale image easily. To further test the robustness of our proposed method, we performed the evaluation using the histogram intersection similarity (HIS) metric [40], which assesses the color histograms of the reference and resultant images in terms of cumulative resemblance. The HIS score comparison is presented in Table III. These results also highlight the superiority of the proposed approach as our technique has obtained the best HIS score compared to the other state-of-the-art algorithms.

The only area in which our proposed technique lags a bit behind one of the existing algorithms [17] is the inference time. Table IV shows the inference times of our designed approach and the state-of-the-art methods. Although the proposed technique takes more time than [17], the overall performance is not too bad considering the superiority of the proposed algorithm in other evaluation results.

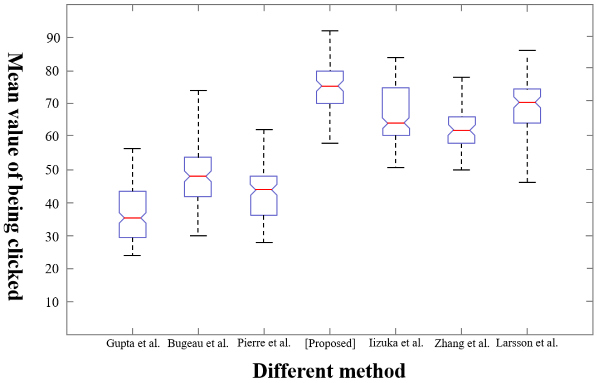
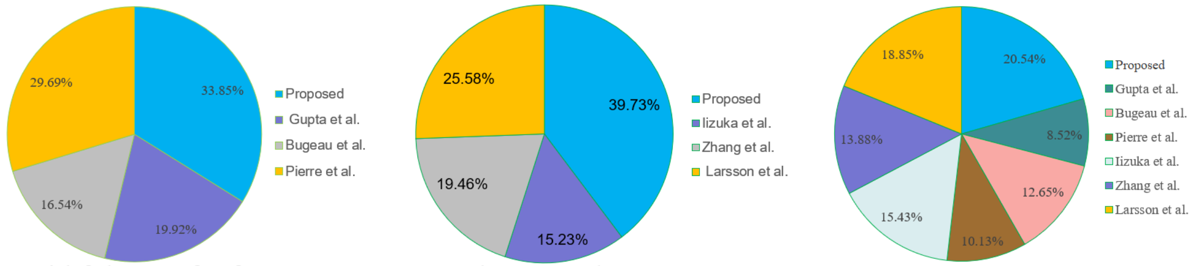


Fig. 12: Boxplots of user preferences for our proposed approach with different methods, including Gupta et al. [8], Bugeau et al. [11], Pierre et al. [38], Iizuka et al. [16], Zhang et al. [17], Larsson et al. [18], showing the mean (red line), quartiles, and extremes (black lines) of the distributions.

E. User Study

In addition to visual evaluation, a user study was conducted to compare the naturalness of the generated images with the existing methods. Although the Peak Signal to Noise Ratio (PSNR) can be used to evaluate the quality of colorization through human perceptual differences, it may not be entirely suitable for all colorized results. For example, the target



(a) Faithfulness of reference (b) Naturalness (c) Fidelity

Fig. 13: The results of three user studies from our approach in comparison with other state-of- the-art methods, including Gupta et al. [8], Bugeau et al. [11], Pierre et al. [38], Iizuka et al. [16], Zhang et al. [17], Larsson et al. [18]. Each section of the pie indicates the percentage of participants that voted for this method. (a) Compared with exemplar-based methods, (b) compared with learning-based methods, and (c) compared with both exemplar-based methods and learning-based methods. The results indicate that the output color images generated by our technique look more natural and authentic than the existing state-of-the-art approaches. Our algorithm can transform more faithful colors from the reference image.

grayscale image colorized by the exemplar-based method could be completely different from the ground truth, but the results can be natural and in line with human visual perception. Therefore, a user study is designed to evaluate our technique against other six approaches [8], [11], [16]–[18], [38].

In order to guarantee the fairness, we showed the results to users and asked them to select one of them that looked more natural. One hundred (100) users aged between 14 and 65 were invited to help complete our user study. We divided the $6\times 7 = 42$ images (6 different grayscale images were colorized by 7 methods, [8], [11], [16]–[18], [38] and ours) into 6 groups and presented them to the users, and asked them to choose the most plausible image in each group. The method accumulated one point if the image generated by it was selected. Hence, each technique could score a maximum of 100 points and a minimum of 0 points in each group. The statistical results of user preference for each approach of the user study are shown in Fig. 12. The red line in the chart indicates the average score of each method. As shown in Fig. 12, the learning-based methods are obviously superior to the exemplar-based algorithms. In addition, our method has the highest average score. This demonstrates that the proposed approach is capable of generalizing well and generating plausible color images.

In addition, the techniques [16]–[18] input only one grayscale image into their networks and do not use any hint. However, the works [8], [11], [38] and our method have the reference image and the target grayscale image as inputs. We qualitatively evaluate the performance of colorization results with naturalness and reference faithfulness, respectively. Then, we compare the fidelity of colorized image of all techniques. For each group of results in Fig. 11, users were asked a single choice question: '*which image do you think is more natural?*' Fig. 13(b) shows the result of the user study. From these results, it can be observed that our algorithm can generate a more natural image than the other learning-based approaches [16]–[18]. In the same way, we implement the second and third user study with the questions of '*which image is more faithful to the reference image?*' and '*which image is more authentic to the ground truth image?*'. Fig. 13(a) and Fig. 13(c) show the results. As shown in Fig. 13(a), our approach is able to match the semantic information between the reference image and the target image and propagate the colors correctly. Fig. 13(c) shows that our technique is able to restore the image details better than state-of-the-art methods.

VII. Conclusion and Future Work

This paper presents a novel example-based image colorization architecture with three significant technical advances: i) to enforce the matching consistency, we design a self-similarity constrained sparse representation that considers global-semantic and local-structure simultaneously. ii) to encourage customization of results, an auxiliary matching information is added to the adversarial network. iii) to stabilize the GAN training, we introduce an exemplar-based conditional broad-GAN for feature-aware image colorization. Many experiments of a separate sub-net and the whole architecture using visual evaluation and user study have shown that the proposed method dramatically outperforms state-of-the-art algorithms. However, our approach relies on global and local consistency matching, which ignores the color bleeding near edges in some cases. In the future, we would like to improve the colorization algorithm by taking edge preservation into account and then extend it to video colorization.

References

[1] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 689–694, 2004.

[2] Y.-C. Huang, Y.-S. Tung, J.-C. Chen, S.-W. Wang, and J.-L. Wu, "An adaptive edge detection based colorization algorithm and its applications," in *ACM Multimedia*, 2005, pp. 351–354.

[3] Y. Qu, T.-T. Wong, and P.-A. Heng, "Manga colorization," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 1214–1220, 2006.

[4] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1120–1129, 2006.

[5] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Eurographics Conference on Rendering Techniques*, 2007, pp. 309–320.

[6] G. Charpiat, M. Hofmann, and B. Schölkopf, "Automatic image colorization via multimodal predictions," in *European Conference on Computer Vision*, 2008, pp. 126–139.

[7] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images," *ACM Transactions on Graphics*, vol. 30, no. 6, pp. 1–8, 2011.

[8] R. K. Gupta, A. Y.-S. Chia, D. Rajan, E. S. Ng, and H. Zhiyong, "Image colorization using similar images," in *ACM Multimedia*, 2012, pp. 369–378.

[9] X. Liu, L. Wan, Y. Qu, T.-T. Wong, S. Lin, C.-S. Leung, and P.-A. Heng, "Intrinsic colorization," *ACM Transactions on Graphics*, vol. 27, no. 5, pp. 152:1–152:9, 2008.

[10] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 277–280, 2002.

[11] A. Bugeau, V.-T. Ta, and N. Papadakis, "Variational exemplar-based image colorization," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 298–307, 2014.

[12] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan, "Deep exemplar-based colorization," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 47:1–47:16, 2018.

[13] B. Li, F. Zhao, Z. Su, X. Liang, Y.-K. Lai, and P. L. Rosin, "Example-based image colorization using locality consistent sparse representation," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5188–5202, 2017.

[14] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *IEEE International Conference on Computer Vision*, 2015, pp. 415–423.

[15] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in *IEEE International Conference on Computer Vision*, 2015, pp. 567–575.

[16] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 110:1–110:11, 2016.

[17] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision*, 2016, pp. 649–666.

[18] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *European Conference on Computer Vision*, 2016, pp. 577–593.

[19] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 119:1–119:11, 2017.

[20] Z. Yang, L. Liu, and Q. Huang, "Learning generative neural networks for 3D colorization," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 2580–2587.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014, pp. 2672–2680.

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5967–5976.

[23] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, pp. 1–7, 2014.

[24] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "CartoonGAN: Generative adversarial networks for photo cartoonization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9465–9474.

[25] B. Dolhansky and C. C. Ferrer, "Eye in-painting with exemplar generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7902–7911.

[26] X. Chen, C. Xu, X. Yang, L. Song, and D. Tao, "Gated-GAN: Adversarial gated networks for multi-collection style transfer," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 546–560, 2019.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015, pp. 1–14.

[28] X.-H. Han, B. Shi, and Y. Zheng, "Self-similarity constrained sparse representation for hyperspectral image super-resolution," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5625–5637, 2018.

[29] C. L. P. Chen and Z. Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 1, pp. 10–24, 2018.

[30] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6836–6845.

[31] C. Xiao, C. Han, Z. Zhang, J. Qin, T.-T. Wong, G. Han, and S. He, "Example-based colourization via dense encoding pyramids," *Computer Graphics Forum*, pp. 1–14, 2019.

[32] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[33] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *ACM-SIAM Symposium on Discrete Algorithms*, 2007, pp. 1027–1035.

[34] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "TurboPixels: Fast superpixels using geometric flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2290–2297, 2009.

[35] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *International Conference on Learning Representations*, 2017, pp. 1–17.

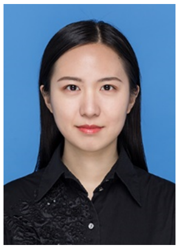
[36] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Neural Information Processing Systems*, 2014, p. 487495.

[37] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://host.robots.ox.ac.uk/pascal/VOC/voc2012, 2012.

[38] F. Pierre, J.-F. Aujol, A. Bugeau, N. Papadakis, and V.-T. Ta, "Luminance-chrominance model for image colorization," *SIAM Journal on Imaging Sciences*, vol. 8, no. 1, pp. 536–563, 2015.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5967–5976.



Haoxuan Li received the M.A. degree in art from the Zhengzhou University of Light Industry, Zhengzhou, China, in 2017. She is currently a research intern with the Visual Media and Data Management Laboratory, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. She is also a Deputy Director of graduate management with the College of Mechanical and Electrical Engineering, Zhengzhou University of Light Industry, Zhengzhou, China. Her current research interests include image colorization, broad learning system, deep learning, and computer vision.

Bin Sheng (Member, IEEE) received the B.A. degree in English and the B.Eng. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2004, and the M.Sc. degree in software engineering from the University of Macau, Taipa, Macau, in 2007, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Shatin, Hong Kong, in 2011. He is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology. His current research interests include virtual reality and computer graphics.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Shatin, Hong Kong, in 2013. He is currently a Research Assistant Professor with The Hong Kong Polytechnic University, Kowloon, Hong Kong. He has one image/video processing national invention patent and has excellent research project reported worldwide by *ACM TechNews*. His current research interests include image/video stylization, artistic rendering and synthesis, and creative media.

Riaz Ali received the B.Eng. degree in software engineering from the Mehran University of Engineering & Technology, Jamshoro, Pakistan, in 2010, and the M.Eng. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2020. He is currently a Lecturer with the Department of Computer Science, Sukkur IBA University, Sukkur, Sindh, Pakistan. His current research interests include image processing, broad learning system, and computer vision.

C. L. Philip Chen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988. He is currently a Chair Professor and the Dean of the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Being a Program Evaluator of the Accreditation Board of Engineering and Technology Education (A-BET) in the U.S., for computer engineering, electrical engineering, and software engineering programs, he successfully architects the University of Macau's Engineering and Computer Science programs receiving accreditations from Washington/Seoul Accord through Hong Kong Institute of Engineers (HKIE), of which is considered as his utmost contribution in engineering/computer science education for Macau as the former Dean of the Faculty of Science and Technology. He is a Fellow of IEEE, AAAS, IAPR, CAA, and HKIE; a member of Academia Europaea (AE), European Academy of Sciences and Arts (EASA), and International Academy of Systems and Cybernetics Science (IASCYS). He received IEEE Norbert Wiener Award in 2018 for his contribution in systems and cybernetics, and machine learnings. He is also a highly cited researcher by Clarivate Analytics in 2018 and 2019. His current research interests include systems, cybernetics, and computational intelligence. Dr. Chen was a recipient of the 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University (in 1988), after he graduated from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA in 1985. He was the IEEE Systems, Man, and Cybernetics Society President from 2012 to 2013, the Editor-in-Chief of the IEEE Transactions on Systems, Man, and Cybernetics: Systems (2014-2019), and currently, he is the Editor-in-Chief of the IEEE Transactions on Cybernetics, and an Associate Editor of the IEEE Transactions on Fuzzy Systems. He was the Chair of TC 9.1 Economic and Business Systems of International Federation of Automatic Control from 2015 to 2017, and currently is a Vice President of Chinese Association of Automation (CAA).