

# Semantic-Sparse Colorization Network for Deep Exemplar-based Colorization

Yunpeng Bai<sup>1</sup>, Chao Dong<sup>2,3</sup>, Zenghao Chai<sup>1</sup>, Andong Wang<sup>1</sup>, Zhengzhuo Xu<sup>1</sup>, and Chun Yuan<sup>1,4</sup>

<sup>1</sup> Tsinghua Shenzhen International Graduate School, China

<sup>2</sup> Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

<sup>3</sup> Shanghai AI Laboratory, China

<sup>4</sup> Peng Cheng National Laboratory, China

{byp20, wad20, xzz20}@mails.tsinghua.edu.cn, chao.dong@siat.ac.cn,  
zenghaochai@gmail.com, yuanc@sz.tsinghua.edu.cn


**Abstract.** Exemplar-based colorization approaches rely on reference image to provide plausible colors for target gray-scale image. The key and difficulty of exemplar-based colorization is to establish an accurate correspondence between these two images. Previous approaches have attempted to construct such a correspondence but are faced with two obstacles. First, using luminance channel for the calculation of correspondence is inaccurate. Second, the dense correspondence they built introduces wrong matching results and increases the computation burden. To address these two problems, we propose Semantic-Sparse Colorization Network (SSCN) to transfer both the global image style and detailed semantic-related colors to the gray-scale image in a coarse-to-fine manner. Our network can perfectly balance the global and local colors while alleviating the ambiguous matching problem. Experiments show that our method outperforms existing methods in both quantitative and qualitative evaluation and achieves state-of-the-art performance.

**Keywords:** image colorization, sparse attention, exemplar-based colorization

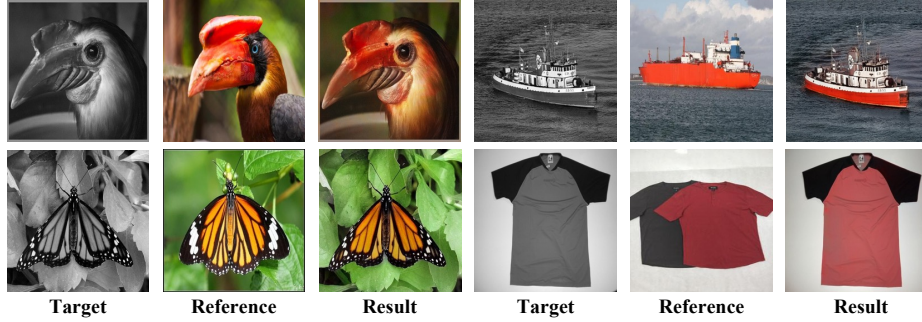
## 1 Introduction

Image colorization is a classic and appealing task that predicts the vivid colors from a gray-scale image. As there is no unique correct color for a given pixel, three classes of methods are proposed to constrain the output color space. The first one is called automatic colorization, such as [5, 40]. These methods generally rely on the powerful convolutional networks and learn a direct mapping from a large-scale image dataset. The second class introduces additional human intervention, such as user-guided scribbles [41, 28, 7] and text [25, 1]. They require users

---

 Corresponding author

to provide reliable color/text labels for more dedicated colorization. While the third class, denoted as exemplar-based method [23,11,33,35,19,2,6,10,36,21], is a trade-off between fully automatic and human intervention strategies. It adopts a reference image as guidance and generates a similar color-style image. These three kinds of methods have different applications and prior information, thus cannot be compared side-by-side. In this work, we study exemplar-based image colorization, due to its large flexibility and excellent performance.



**Fig. 1.** Overview colorization results of the proposed method. Our method can commendably build correspondence between the target and reference images and has the capability to generate a plausible colorization of gray-scale images.

The difficulty of exemplar-based image colorization is to build an accurate correspondence between the gray-scale image and the color reference. Some works regard colorization as a style transfer problem [35], and usually transfer the global color tones. As a result, they lack detailed color matching between semantically similar objects/parts. Other researchers [38,19,23,36,39] propose to construct a dense correspondence with a correlation matrix, whose elements characterize pairwise similarity between different image features. Although they have achieved considerable progress, they are still facing two obstacles. First, the correspondence is calculated using the luminance channel of the input image. However, as gray-scale images do not contain enough semantic information as color images (a common knowledge in image classification [11]), the correspondence based on the luminance channel [23,36,38,39] is inaccurate. Second, the dense correspondence itself will also bring in unavoidable drawbacks. It not only introduces wrong matching results for semantically unrelated objects, but also increases the computation burden.

To address the above mentioned problems, we propose a new coarse-to-fine colorization framework – Semantic-Sparse Colorization Network – to transfer both the global image style and the detailed semantic-related colors to the gray-scale image. Specifically, in the coarse colorization stage, we adopt an image transfer network to obtain a preliminary colorized result. The color information of the reference image is encoded as a vector, which is then migrated to

the gray-scale image by an AdaIN [12] operation. In the fine colorization stage, we will first calculate the semantic correspondence between the coarse result and the reference image. Specially, only the semantic-significant parts and some background regions are reserved for calculation, leading to a sparse correlation matrix. Then the attention mechanism will be used to re-weight the reference image and help generate the final color result. The proposed method can perfectly balance the global and local colors while alleviating the ambiguous matching problem caused by dense correspondence. Extensive experiments have shown the superiority of our network towards other state-of-the-art methods. To facilitate numerical evaluation, we also propose a unified evaluation pipeline for all exemplar-based colorization methods. Our code will be publicly available for research purpose.

Our main contributions are summarized as follows:

- We propose to build a more accurate correspondence between a coarse-colored result and the reference image. It not only minimizes the information gap between the gray-scale input and the color reference, but also achieves better performance on details.
- We propose a sparse attention mechanism to make the model focus on the semantically significant regions in the reference image. It could produce more detailed results with lower computation cost.
- We collect a new test dataset from ImageNet to solve the problem of fair comparison. We also design a new quantitative evaluation metric to evaluate exemplar-based colorization methods.

## 2 Related Work

Because image colorization plays an essential role in image processing tasks such as old photo restoration and image editing, this subject has been studied for a long time [4,2,26,24,13]. Recently, many studies have used learning-based methods to solve this ill-posed problem. These approaches can be roughly grouped into three classes.

The first one is called automatic colorization, which directly maps gray-scale images to color images, such as [5] and [40]. They are the earliest methods to use convolutional networks to learn the mapping from a large-scale image dataset. MemoPainter [37] uses a memory network to “memorize” rare examples, which can avoid the interference of dominant color in the dataset and make the model perform well even without sufficient data. More recently, Transformer has also been applied to address this task [18]. Some works [3,31,32] take advantage of generative models to promote the diversity of results. For instance, [32] leverages the rich and diverse color priors encapsulated in a pretrained StyleGAN [15] to recover vivid colors. The variational autoencoder (VAE) architecture has also been used in [9]. However, the colorization process of these methods are lack of controllability.

The second class introduces additional human intervention, such as user-guided scribbles and text. They require users to provide reliable color/text labels

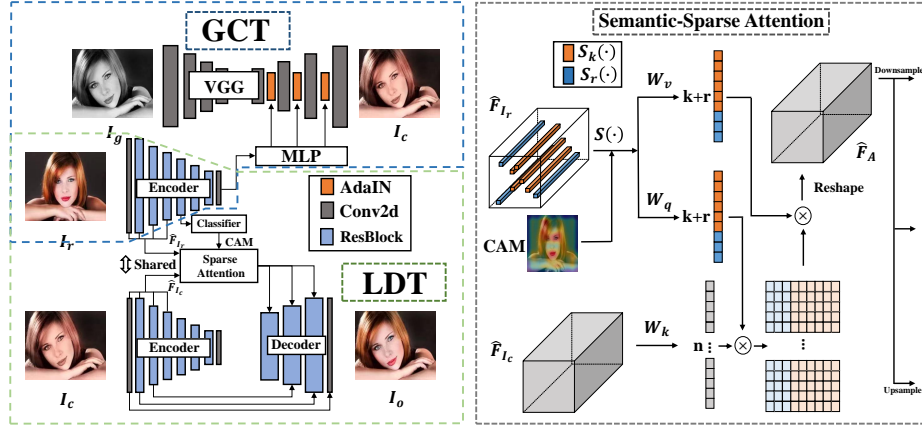
for more dedicated colorization. Traditional scribble-based colorization methods [20,34] usually propagate the local user hints to the whole image via an optimization approach, while learning-based methods [41,28,7] will combine color prior learned from large-scale image dataset with user’s intervention for colorization. Recently, some researchers [16] have found that leveraging user interactions would be a promising approach for reducing color-breeding artifacts. These methods require a certain amount of human effort, and the quality of results depends on the user’s skills. Text-based methods usually adopt image captions [25] or palettes converted from the text [1] as means of intervention. However, the color represented by text is challenging to transfer to the image accurately.

The third class, denoted as exemplar-based method, is a trade-off between fully automatic and human intervention strategies. Compared to the above two classes, it adopts sample reference images to provide rich colors without requiring the user to do too much manual work. The key and difficulty of exemplar-based colorization is to establish an accurate correspondence between these two images. DEPN [33] uses a pyramid structure to exploit multi-scale color information, but it only captures the global tones because no semantic correspondence is established. Some works [35] regard exemplar-based colorization as a style transfer problem, but cannot guarantee the correctness of semantics because they also lack a correspondence. Deep Image Analogy [22] was used in [11] to make the target and reference luminance channels aligned to get a coarse chrominance map for further refinement. [23] uses features extracted from the luminance channel of the target and reference images to obtain dense correspondence. However, inaccuracies caused by using luminance channels to calculate correspondence and wrong matching problems introduced by dense correspondence will lead to unsatisfactory results. A general attention based framework is proposed in [36] to fuse colors from the database when the correspondence is not established. However, this method sometimes will mistakenly use the colors from the database when the selected two images are highly semantically related, resulting in the final results looking different from the reference image.

### 3 Methods

#### 3.1 Overview of the Proposed Method

The task of exemplar-based colorization can be formulated as follows. Given a gray-scale image  $I_g$ , which only contains the luminance channel  $l$ , our goal is to predict the corresponding  $a$  and  $b$  color channels in the CIE Lab color space, according to the reference color image  $I_r$ . The main challenge is to build an appropriate correspondence between the gray-scale image and the color reference. In order to make full use of the color information in the reference image, we will utilize the reference image twice in a coarse-to-fine manner during the whole colorization process. The proposed framework, namely Semantic-Sparse Colorization Network (SSCN), consists of two auxiliary modules, which transfer global and local colors in the reference image, respectively.



**Fig. 2.** The illustration of the proposed two-stage image colorization framework. Our method uses a coarse-colored image to build more accurate correspondence, which is completely different from previous works. The right part shows our proposed sparse attention mechanism in detail. With the help of the semantic information provided by CAM, the model can accurately use the critical parts of the reference image and reduce the complex computation caused by the attention mechanism.

The overall pipeline of SSCN is illustrated in Figure 2. Specifically, taking the reference image  $I_r$  as input, our model will first encode it into features  $F_{I_r}$ . These features will be used in both global and local coloring modules. In the coarse colorization stage, the Global Color Transfer (GCT) module will use  $F_{I_r}$  to preliminarily color the gray-scale image  $I_g$ , and get a coarse-colored result  $I_c$ , which has similar global tones as  $I_r$ . Then the coarse output  $I_c$  will be further encoded into features  $F_{I_c}$  with the same encoder as  $F_{I_r}$ . In the fine colorization stage, the Local Details Transfer (LDT) module will use  $F_{I_r}$  and  $F_{I_c}$  to construct a correspondence that focuses on the semantically relevant regions of  $I_r$ . Note that these regions are sparsely selected according to their semantic levels. Based on the predicted mappings from LDT, the reference features  $F_{I_r}$  are reorganized and fused with  $F_{I_c}$  at different scales. Finally, the decoder takes the fused color features to produce the  $a$  and  $b$  channels of the input image  $I_g$ .

### 3.2 Global Color Transfer

We will first introduce the encoder of  $I_r$ , which is shared in both GCT and LDT modules. The encoder consists of six residual blocks. The last layer of  $F_{I_r}$  is passed through an MLP to form the style vector, which will be used in the GCT module for global style transfer. In GCT, the gray-scale image  $I_g$  will first be encoded into features  $\{x_1, x_2, \dots, x_n\}$ . Then, we perform coarse colorization in the feature space by changing feature statistics with AdaIN operation as:

$$AdaIN(x_i, y) = y_{s,i} \frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i}, \quad (1)$$

where  $\mu(x_i)$  and  $\sigma(x_i)$  represent the  $i^{th}$  feature map’s mean and variance, respectively.  $y_s$  and  $y_b$  are the affine parameters of the style vector, which is obtained from  $F_{I_r}$  via MLP transformation. Each feature map  $x_i$  is normalized separately and then scaled/biased using the corresponding coefficients from  $y(y_s, y_b)$ . After affine transformation, each feature channel will have the activation for certain color information. These features can be inverted to the Lab space by a convolutional decoder. We finally get the coarse colorized result  $I_c$  of the coarse colorization stage. In our implementation, the encoder uses sub-layers of the VGG19 [29], and the decoder is symmetric structure. AdaIN are added after CNN layers of the decoder.

### 3.3 Local Details Transfer

The target of the LDT module is to build a more detailed and accurate correspondence between the coarse-colored result  $I_c$  and the reference image  $I_r$ . To begin with, we encode  $I_c$  into the corresponding features  $F_{I_c}$ , with the same encoder as  $F_{I_r}$ . To find their correspondence, we extract features from the first four layers of  $F_{I_r}$  and  $F_{I_c}$ , and resize them to the same spatial size of  $1/4$  input image. Then these features are concatenated to form features  $\hat{F}_{I_r}$  and  $\hat{F}_{I_c}$ , corresponding to the latent states of coarse and reference image, respectively. Their spatial size is both  $d \times H/4 \times W/4$ , where  $d$  is the number of feature maps. To facilitate computation, they are further flattened in the last two directions, and form features of size  $d \times HW/16$ . In this way, we segment the input image into  $HW/16$  regions and represent each region with a  $d$  dimensional vector.

Based on the obtained features  $\hat{F}_{I_r}$  and  $\hat{F}_{I_c}$ , the LDT module will calculate a correlation matrix  $A$  via attention mechanism, whose element is computed by the scaled dot product [30] illustrated as Formula 2:

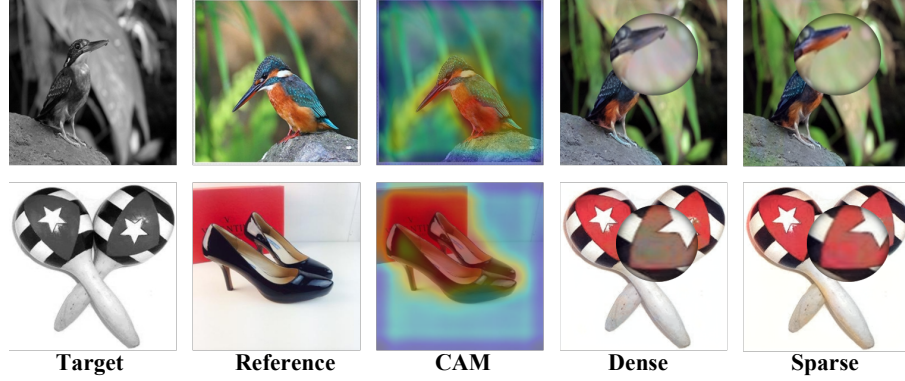
$$\alpha_{ij} = \underset{j}{softmax} \left( \frac{(W_q f_i^c) \cdot (W_k f_j^r)}{\sqrt{d}} \right). \quad (2)$$

Here,  $\alpha_{ij}$  represents the similarity between the  $i$ -th region of  $\hat{F}_{I_c}$  and the  $j$ -th region of  $\hat{F}_{I_r}$ .  $\hat{F}_{I_c}$  is used to retrieve relevant local details from  $\hat{F}_{I_r}$ . Then, we can re-weight the features  $\hat{F}_{I_r}$  to obtain the attended feature  $\hat{F}_a$  through a weighted sum operation as Formula 3:

$$f_i^a = \sum_j \alpha_{ij} W_v f_j^r, \quad (3)$$

where  $W_q$ ,  $W_k$  and  $W_v$  represent the linear transformation matrix into *query*, *key*, and *value* vectors, respectively. The attended features  $\hat{F}_a$  will be reshaped to the size of  $d \times H/4 \times W/4$  and further resized into a suitable shape, fused with the features  $F_{I_c}$  at different scales and fed into the U-Net [27] decoder for the final detailed result of the fine colorization stage.

**Semantic-Sparse Correspondence.** In the above description, we use a standard attention mechanism to calculate the dense correspondence between

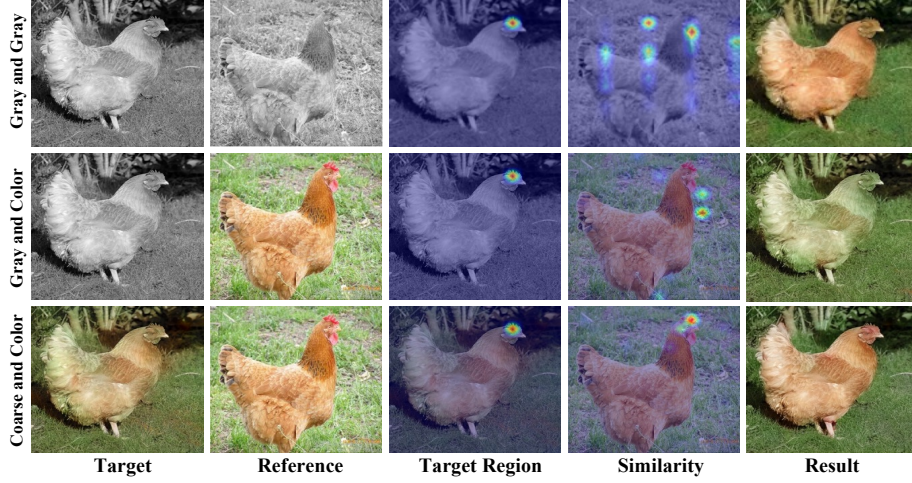


**Fig. 3.** Comparison results of dense and sparse correspondence strategies. The output results will be disturbed by the re-weighting process using dense correspondence. Sparse attention focusing on semantically important areas can solve this problem.

coarse and reference images. We further propose a semantically sparse correspondence for better results with less computation cost. To be specific, the reference features  $\hat{F}_{I_r}$  will go through a selection operation. First, the fifth layer of  $F_{I_r}$  will be fed into a classifier and get a class activation map (CAM) [42], which is used as the reference for selection. The CAM is flattened to  $C = \{c_1, c_2, \dots, c_{HW/16}\} \in \mathbb{R}^{HW/16}$ . The selection operation  $S(\cdot)$  contains the top- $k$  selection  $S_k(\cdot)$  and random selection  $S_r(\cdot)$  implemented upon  $C$ . The  $S_k(\cdot)$  selects the  $k$  largest elements of  $C$  and records their indexes  $\mathbf{T}_k$ . This encourages the attention mechanism to focus more on semantically significant areas and reduce the interference caused by insignificant parts. At the same time, the coloring of the background areas also needs reference. Thus  $S_r(\cdot)$  randomly selects  $r$  more indexes  $\mathbf{T}_r$ . Finally, we obtain  $S(C) = \mathbf{T}_k \cup \mathbf{T}_r$  and the semantic-sparse features  $\hat{F}_{I_r}[S(C)]$ . To calculate the new correspondence map, we can simply replace the features  $\hat{F}_{I_r}$  with  $\hat{F}_{I_r}[S(C)]$  in Formula 2,3. The other steps remain the same as above.

### 3.4 Discussion

**Dense Correspondence vs. Sparse Correspondence.** Dense correspondence will be easily affected by irrelevant regions, especially when the reference is completely different from the gray-scale image. Even if the target region has low similarity with most reference regions, the re-weighting process will still disturb the final result. In contrast, sparse correspondence can overcome this difficulty by focusing only on semantically important regions, which can reduce the interference of other regions. Moreover, the computational complexity goes from  $\mathcal{O}((HW)^2)$  to  $\mathcal{O}((k+r)HW)$ , while  $(k+r)$  is generally 8 to 16 times smaller than  $HW$ . The comparison results of these two strategies are shown in Figure 3.



**Fig. 4.** Comparison results of using three different data types to build the correspondence. The coarse-colored result we proposed to use can establish more accurate correspondence than the other two common types.

It can be observed that some details are more accurately colorized after reducing the interference.

**Coarse-colored vs. Gray-scale.** In this work, we propose to use a coarse-colored image to build the correspondence with the reference, which is completely different from previous works [23,36,38,39]. The coarse result is already consistent with the reference’s global color style, thus can produce more dedicated correspondence than directly using the gray-scale image. Moreover, the correspondence between color images is more accurate than that between gray-scale images (luminance channels). To verify this comment, we build a correlation matrix for three data types with the same operations. Figure 4 shows the comparison results of the similarity between one target region and all reference regions. It is clear that the chicken comb is correctly matched between two color images, even with different colors.

### 3.5 Objective Functions

**Smooth-L1 Loss.** To avoid simply using the average scheme for solving the ambiguity colorization problem, a widely used loss function Smooth-L1 loss is adopted in image colorization tasks. This loss is added to the results of both two stages in our architecture as  $L_{stage1}$  and  $L_{stage2}$ . The following Formula 4 can calculate the Smooth-L1 loss between  $T_{ab}$  and  $\hat{T}_{ab}$ :

$$L_{stage1,2}(T_{ab}, \hat{T}_{ab}) = \begin{cases} \frac{1}{2}(T_{ab} - \hat{T}_{ab})^2 & \text{for } |T_{ab} - \hat{T}_{ab}| \leq \delta \\ \delta |T_{ab} - \hat{T}_{ab}| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases} \quad (4)$$

**Classification Loss.** There is a classification loss  $L_{cls}$  in the classifier to get a CAM as a reference for  $S(\cdot)$ . This loss can also improve the encoder’s ability of extracting color features. When  $F_{I_r}$  is fed into the classifier, its label vector is predicted.  $L_{cls}$  is defined as the cross-entropy between the classification vector  $\hat{y}$  and its ground truth one-hot label.

**Color Histogram Loss.** To transfer the color distribution of the reference image to the target image accurately, we also add a histogram loss to the final output as Formula 5. Similar to the previous work [40], we treat the problem as multinomial classification. We quantify  $\hat{T}_{ab}$  output space into bins with  $gridsize = 10$  and keep the in-gamut  $Q = 313$ . The mapping to predicted color distribution  $\hat{Z} \in [0, 1]^{H \times W \times Q}$  is also learned with the decoder. The  $L_{his}$  is defined as a cross-entropy loss for every pixel to measure the distance between predicted distribution  $\hat{Z}$  and ground truth  $Z$ , and sum over all pixels.

$$L_{his}(\hat{Z}, Z) = - \sum_{h,w} \sum_q Z_{h,w,q} \log(\hat{Z}_{h,w,q}) . \quad (5)$$

**TV Regularization.** To encourage spatial smoothness in the output result  $\hat{T}_{ab}$ , we follow previous work [14] and apply the total variation regularization  $L_{TV}(\hat{T}_{ab})$  to the output of the fine colorization stage.

In summary, the overall loss function for the entire network is defined as:

$$L_{total} = \lambda_{stage1} L_{stage1} + \lambda_{stage2} L_{stage2} + \lambda_{TV} L_{TV} + \lambda_{cls} L_{cls} + \lambda_{his} L_{his} , \quad (6)$$

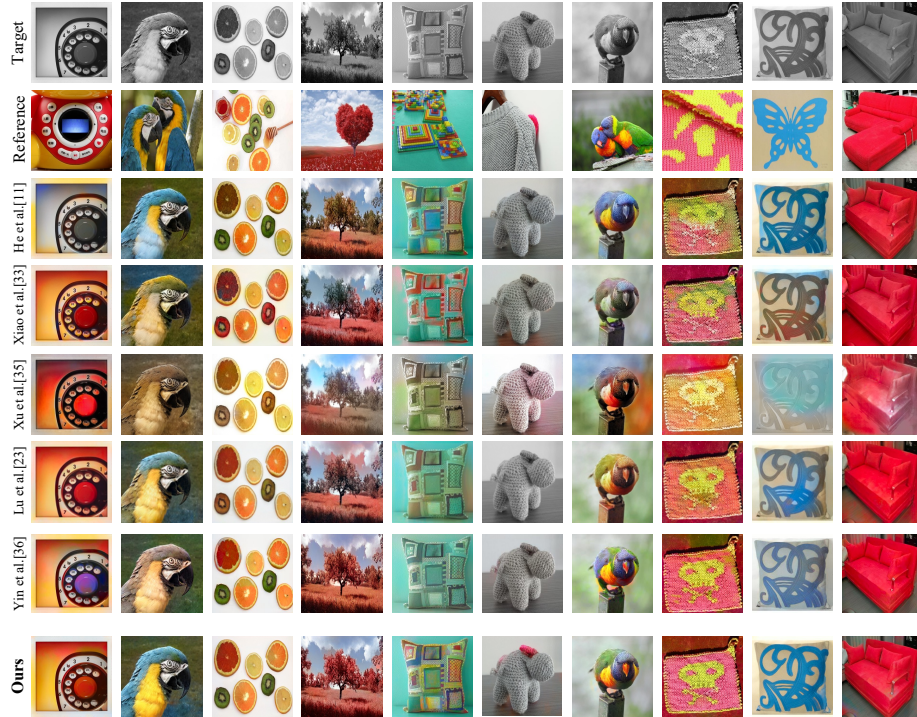
where  $\lambda_{stage1}$ ,  $\lambda_{stage2}$ ,  $\lambda_{TV}$ ,  $\lambda_{cls}$  and  $\lambda_{his}$  are hyperparameters to constrain different loss terms.

## 4 Experiments

### 4.1 Implementation Details

We use ImageNet’s [8] total training set to train the entire network with 5 epochs and set mini-batch size as 8. During training, the input image will be resized to  $256 \times 256$ . We use Adam [17] for optimization with  $\beta_1 = 0.9, \beta_2 = 0.999$ . The learning rate is set to 0.0001. We set the coefficients for each loss function as follows:  $\lambda_{stage1} = 100$ ,  $\lambda_{stage2} = 100$ ,  $\lambda_{cls} = 0.1$ ,  $\lambda_{TV} = 10$ , and  $\lambda_{his} = 1$ . For the  $S(\cdot)$ , both  $k$  and  $r$  are set to 256.

For the exemplar-based colorization method, it is impossible to find enough source-reference pairs to train the network. We adopt a scheme similar to [19]. The reference is generated from the original image by geometric distortion, which can provide complete color information for the target image. The geometric distortion is realized by thin plate splines (TPS) transformation. The distortion is randomly applied to each image. In the training process, we apply violent transformation to some images to simulate semantically unrelated reference images.



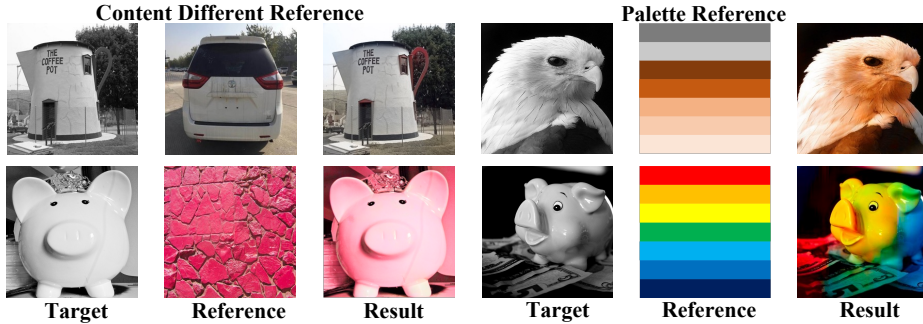
**Fig. 5.** Qualitative comparison of colorizing results with previous methods. The target image, reference image, and each method’s colorized images are displayed from top to bottom. The proposed method outperforms other models and achieves state-of-the-art performance.

## 4.2 Comparison with Previous Methods

**Visual Comparison.** We compare the results of our method with previous exemplar-based colorization approaches [33,35,11,23,36]. We run all 6 models on 230 pairs of images collected from ImageNet validation set and show several representative results. All comparison results are obtained by public available codes. We show the qualitative comparison in Figure 5. See our supplementary materials for more results.

The 4th column of Figure 5 shows the results of colorizing objects with unusual or artistic colors. Compared with method [11] constrained by the perceptual loss, the proposed method can appropriately colorize the target image according to the user’s requirement. Since [23] tends to make the color histograms of the two images consistent, resulting in the wrong spatial distribution of colors.

In the 6th column, when there are large regions with less semantics in the image, our method can pay more attention to the semantically relevant areas, e.g., the the pink area, while other methods fail to colorize the object or simply get a smooth result. In the 2nd column, the parrots in two input images are



**Fig. 6.** Colorization results of using content different references and palettes. Visually satisfactory results can also be obtained using these two types of references.

highly semantically related, while [36] uses the colors in the database, resulting in an unsatisfactory final result.

When the reference image is semantically unrelated to the target image (shown as 1st column in Figure 5), due to the dependence on prior color knowledge, [11] will ignore the colors from the reference image. Histogram-based methods [33] can get plausible results by transferring global tones, whereas our method can yield better results. For some images with many details, [23] cannot properly colorize these details due to the inappropriate correspondence constructed with two gray-scale images, while the proposed method allows the target image to be colored correctly, e.g., 5th and 7th columns in Figure 5.

These experimental results show that the proposed method can transfer color information for different image pairs accurately and effectively. We also show some results of using content different references and palettes in Figure 6. Even when the semantics of the reference image are irrelevant or have no semantics, our method can also get satisfactory results.

**Self-Augmentation PSNR/SSIM.** Unlike automatic colorization, in exemplar-based colorization setting, when given a target-reference pair, there is no ground truth that has both the target’s shape and the reference’s color. The histogram intersection similarity (HIS) used in previous work [23,36] is not a suitable index. Mismatches may also occur in the spatial color distribution of the result with high histogram similarity with the reference image. In order to make a quantitative evaluation of the colorization results, similar to the training process, we use the augmentation of a color image as the reference to colorize its luminance channel, so that the original color image can be used as ground truth. With ground truth available for comparison, some existing evaluation metrics, such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM), can be used for evaluation.

We select 5000 images from the validation set of ImageNet to do three different data augmentation, including TPS, random rotation (RR), and random



**Fig. 7.** Comparison results of using random cropping reference image to colorize the target image. The results obtained by other methods are not satisfactory even when using such a suitable reference.

**Table 1.** Quantitative comparisons of self-augmentation PSNR/SSIM. A higher value indicates a better preference, while the proposed method outperforms other models.

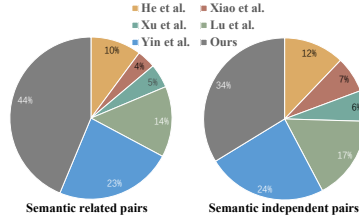
| Methods               | TPS                | RR                 | RC                 | Mean               |
|-----------------------|--------------------|--------------------|--------------------|--------------------|
| He et al.(2018)[11]   | 28.51/0.902        | 28.67/0.903        | 27.57/0.898        | 28.25/0.901        |
| Xiao et al.(2020)[33] | 25.17/0.912        | 25.30/0.913        | 24.98/0.910        | 25.15/0.911        |
| Xu et al.(2020)[35]   | 22.46/0.873        | 21.65/0.846        | 21.55/0.862        | 21.88/0.860        |
| Lu et al.(2020)[23]   | 27.93/0.913        | 29.80/0.931        | 27.12/0.907        | 28.28/0.917        |
| Yin et al.(2021)[36]  | 31.87/0.948        | 34.24/0.952        | 29.85/0.939        | 31.98/0.946        |
| <b>Ours</b>           | <b>36.32/0.969</b> | <b>35.49/0.966</b> | <b>32.39/0.958</b> | <b>34.73/0.964</b> |

cropping (RC) as references to get different results. The quantitative comparisons of three different augmentation are reported in Table 1. Figure 7 shows an example of using a RC reference and comparing the results with other methods. We will release this test dataset for future comparison.

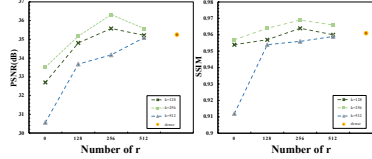
**User Evaluation.** We conduct user evaluation to verify the proposed method’s effectiveness subjectively. In this part, we randomly select 50 groups from the above results. Semantically dependent pairs and semantically unrelated pairs are distributed in half. Eventually, all  $6 \times 50$  color images are distributed anonymously and randomly to 30 college participants.

For fairness, the images with the same reference are shown simultaneously in a random order. All participants were asked to observe the images for no more than 5 seconds and choose the image that better matches the reference.

As shown in Figure 8, we show the percentage of votes for each method in the form of pie chart. It shows that images of our method are mostly preferred.



**Fig. 8.** The users’ preferences for six different methods. Under two different image pairs, our results have been the most selected by users.

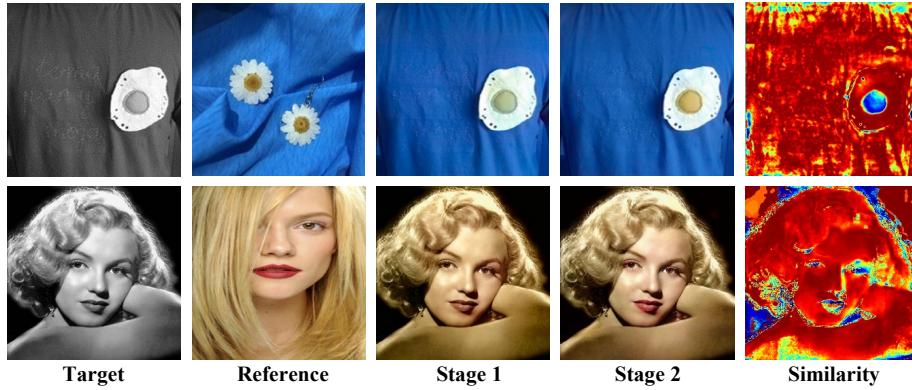


**Fig. 9.** The picture shows how the numerical changes of  $k$  and  $r$  affect the final result. Larger or smaller  $k$  and  $r$  will reduce the quality.

## 5 Ablation Studies

**Ablation study of  $S_k(\cdot)$  and  $S_r(\cdot)$ .** The use of sparse correspondence will lead to the question: how to select an appropriate number of regions in the process? Then we further study the effect of  $k$  and  $r$ , and use TPS reference to evaluate results quantitatively as described above. When the resolution of the reference image is  $256 \times 256$ , there are 4096 features available for selection. We increase  $k$  and  $r$  gradually from 128 and 0, respectively. The comparison results are shown in Figure 9. Without random selection, the value of PSNR/SSIM will be much lower because some areas of the background are incorrectly colored. Increasing  $r$  gradually can improve the results, but increasing  $r$  further will cause the result deteriorate again. In addition, it can be seen from the comparison of the three broken lines that a larger or smaller  $k$  will reduce the quality.

**Ablation study of two-stage architecture.** To illustrate the importance of the two-stage structure in our model, we conduct ablation study on  $k = 256, r = 256$  version. First, we evaluate the first stage results with PSNR and SSIM values of 30.02 and 0.937. There is a huge gap between them and the final results, thus illustrating the importance of LDT. To further validate the importance of preliminary coloring, we remove GCT from the whole architecture for comparison. Instead, we use another network with a similar structure to the encoder of  $I_r$  but with one channel input to extract the features of gray-scale image and calculate the correspondence in the same way. Due to the lack of information in the gray-scale image, the PSNR and SSIM values will decrease by 3.30 and 0.014. We also analyze the relationship between the results of the two stages in the encoder feature space. The similarity of features is shown in the form of heat map in Figure 10. We can see that the differences between the two are mainly concentrated in some semantic details, which are completed in the second stage.



**Fig. 10.** Ablation study on the relationship between the results of the two stages. The bluer the part in the heat map, the less similar the features. The main differences are concentrated in some semantic details.

**Ablation study of Loss Functions.** In order to verify that the classifier does not only provide a CAM but also help the encoder extract color features, we ablate the classification loss on the dense version. After this loss is removed, the corresponding PSNR and SSIM values are 33.73 and 0.952, while the PSNR and SSIM values of the dense version are 35.25 and 0.961. In addition, we also ablate color histogram loss of the best version ( $k = 256, r = 256$ ) to analyze its effect. The PSNR and SSIM values will decrease by 1.28 and 0.009. Removing either of these losses will reduce the model’s performance, especially in the  $L_{cls}$ .

## 6 Conclusions

This paper proposes a colorization framework named Semantic-Sparse Colorization Network (SSCN) to colorize the target image in a coarse-to-fine manner. Specifically, an image transfer network is adopted in the coarse colorization stage to obtain a preliminary colorized result. In the fine colorization stage, semantically related areas of the reference image will be selected to color the details of the target image. Thus, SSCN can adequately transfer a reference image’s global color and local details onto a gray-scale image. It provides a way to obtain different levels of color information from the reference image hierarchically and accurately. Extensive experiments show that the proposed method outperforms previous state-of-the-art approaches by a large margin.

**Acknowledgment.** This work was supported by SZSTC Grant No.JCYJ201908 09172201639 and WDZC20200820200655001, Shenzhen Key Laboratory ZDSYS20210623092001004.

## References

1. Bahng, H., Yoo, S., Cho, W., Park, D.K., Wu, Z., Ma, X., Choo, J.: Coloring with words: Guiding image colorization through text-based palette generation. In: ECCV 2018. pp. 443–459. Springer (2018) [1](#), [4](#)
2. Bugeau, A., Ta, V., Papadakis, N.: Variational exemplar-based image colorization. *IEEE Trans. Image Process.* **23**(1), 298–307 (2014) [2](#), [3](#)
3. Cao, Y., Zhou, Z., Zhang, W., Yu, Y.: Unsupervised diverse colorization via generative adversarial networks. In: ECML PKDD 2017. pp. 151–166. Springer (2017) [3](#)
4. Charpiat, G., Hofmann, M., Schölkopf, B.: Automatic image colorization via multimodal predictions. In: ECCV 2008. pp. 126–139. Springer (2008) [3](#)
5. Cheng, Z., Yang, Q., Sheng, B.: Deep colorization. In: ICCV 2015. pp. 415–423. IEEE Computer Society (2015) [1](#), [3](#)
6. Chia, A.Y.S., Zhuo, S., Gupta, R.K., Tai, Y., Cho, S., Tan, P., Lin, S.: Semantic colorization with internet images. *ACM Trans. Graph.* **30**(6), 156 (2011) [2](#)
7. Ci, Y., Ma, X., Wang, Z., Li, H., Luo, Z.: User-guided deep anime line art colorization with conditional adversarial networks. In: MM 2018, pp. 1536–1544. ACM (2018) [1](#), [4](#)
8. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: CVPR 2009. pp. 248–255. IEEE Computer Society (2009) [9](#)
9. Deshpande, A., Lu, J., Yeh, M., Chong, M.J., Forsyth, D.A.: Learning diverse image colorization. In: CVPR 2017. pp. 2877–2885. IEEE Computer Society (2017) [3](#)
10. Gupta, R.K., Chia, A.Y.S., Rajan, D., Ng, E.S., Huang, Z.: Image colorization using similar images. In: MM 2012. pp. 369–378. ACM (2012) [2](#)
11. He, M., Chen, D., Liao, J., Sander, P.V., Yuan, L.: Deep exemplar-based colorization. *ACM Trans. Graph.* **37**(4), 47:1–47:16 (2018) [2](#), [4](#), [10](#), [11](#), [12](#)
12. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017. pp. 1510–1519. IEEE Computer Society (2017) [3](#)
13. Huang, Y., Tung, Y., Chen, J., Wang, S., Wu, J.: An adaptive edge detection based colorization algorithm and its applications. In: MM 2005. pp. 351–354. ACM (2005) [3](#)
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV 2016. pp. 694–711. Springer (2016) [9](#)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019. pp. 4401–4410. Computer Vision Foundation / IEEE (2019) [3](#)
16. Kim, E., Lee, S., Park, J., Choi, S., Seo, C., Choo, J.: Deep edge-aware interactive colorization against color-bleeding effects. *CoRR* (2021) [4](#)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR 2015 (2015) [9](#)
18. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer. *CoRR* **abs/2102.04432** (2021) [3](#)
19. Lee, J., Kim, E., Lee, Y., Kim, D., Chang, J., Choo, J.: Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In: CVPR 2020. pp. 5800–5809. IEEE Computer Society (2020) [2](#), [9](#)

20. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. *ACM Trans. Graph.* **23**(3), 689–694 (2004) [4](#)
21. Li, H., Sheng, B., Li, P., Ali, R., Chen, C.L.P.: Globally and locally semantic colorization via exemplar-based broad-gan. *IEEE Trans. Image Process.* pp. 8526–8539 (2021) [2](#)
22. Liao, J., Yao, Y., Yuan, L., Hua, G., Kang, S.B.: Visual attribute transfer through deep image analogy. *ACM Trans. Graph.* **36**(4), 120:1–120:15 (2017) [4](#)
23. Lu, P., Yu, J., Peng, X., Zhao, Z., Wang, X.: Gray2colornet: Transfer more colors from reference image. In: *MM 2020*. pp. 3210–3218. ACM (2020) [2](#), [4](#), [8](#), [10](#), [11](#), [12](#)
24. Luan, Q., Wen, F., Cohen-Or, D., Liang, L., Xu, Y., Shum, H.: Natural image colorization. In: *Proceedings of the Eurographics Symposium on Rendering Techniques 2007*. pp. 309–320. Eurographics Association (2007) [3](#)
25. Manjunatha, V., Iyyer, M., Boyd-Graber, J.L., Davis, L.S.: Learning to color from language. In: *NAACL-HLT 2018*. pp. 764–769. Association for Computational Linguistics (2018) [1](#), [4](#)
26. Qu, Y., Wong, T., Heng, P.: Manga colorization. *ACM Trans. Graph.* **25**(3), 1214–1220 (2006) [3](#)
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI 2015. Lecture Notes in Computer Science*, vol. 9351, pp. 234–241. Springer (2015) [6](#)
28. Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Scribbler: Controlling deep image synthesis with sketch and color. In: *CVPR 2017*. pp. 6836–6845. IEEE Computer Society (2017) [1](#), [4](#)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *ICLR 2015* (2015) [6](#)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*. pp. 5998–6008 (2017) [6](#)
31. Vitoria, P., Raad, L., Ballester, C.: Chromagan: Adversarial picture colorization with semantic class distribution. In: *WACV 2020*. pp. 2434–2443. IEEE (2020) [3](#)
32. Wu, Y., Wang, X., Li, Y., Zhang, H., Zhao, X., Shan, Y.: Towards vivid and diverse image colorization with generative color prior. *CoRR* (2021) [3](#)
33. Xiao, C., Han, C., Zhang, Z., Qin, J., Wong, T., Han, G., He, S.: Example-based colourization via dense encoding pyramids. *Comput. Graph. Forum* **39**(1), 20–33 (2020) [2](#), [4](#), [10](#), [11](#), [12](#)
34. Xu, K., Li, Y., Ju, T., Hu, S., Liu, T.: Efficient affinity-based edit propagation using K-D tree. *ACM Trans. Graph.* **28**(5), 118 (2009) [4](#)
35. Xu, Z., Wang, T., Fang, F., Sheng, Y., Zhang, G.: Stylization-based architecture for fast deep exemplar colorization. In: *CVPR 2020*. pp. 9360–9369. IEEE (2020) [2](#), [4](#), [10](#), [12](#)
36. Yin, W., Lu, P., Zhao, Z., Peng, X.: Yes, ”attention is all you need”, for exemplar based colorization. In: *MM ’21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*. pp. 2243–2251. ACM (2021) [2](#), [4](#), [8](#), [10](#), [11](#), [12](#)
37. Yoo, S., Bahng, H., Chung, S., Lee, J., Chang, J., Choo, J.: Coloring with limited data: Few-shot colorization via memory augmented networks. In: *CVPR 2019*. pp. 11283–11292. Computer Vision Foundation / IEEE (2019) [3](#)
38. Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., Chen, D.: Deep exemplar-based video colorization. In: *CVPR 2019*. pp. 8052–8061. Computer Vision Foundation / IEEE (2019) [2](#), [8](#)

- 39. Zhang, J., Xu, C., Li, J., Han, Y., Wang, Y., Tai, Y., Liu, Y.: Scsnet: An efficient paradigm for learning simultaneously image colorization and super-resolution. CoRR (2022) [2](#), [8](#)
- 40. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV 2016. pp. 649–666. Springer (2016) [1](#), [3](#), [9](#)
- 41. Zhang, R., Zhu, J., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Trans. Graph. **36**(4), 119:1–119:11 (2017) [1](#), [4](#)
- 42. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR 2016. pp. 2921–2929. IEEE Computer Society (2016) [7](#)