Self-Supervised Color-Concept Association via Image Colorization

Ruizhen Hu, Ziqi Ye, Bin Chen, Oliver van Kaick, and Hui Huang

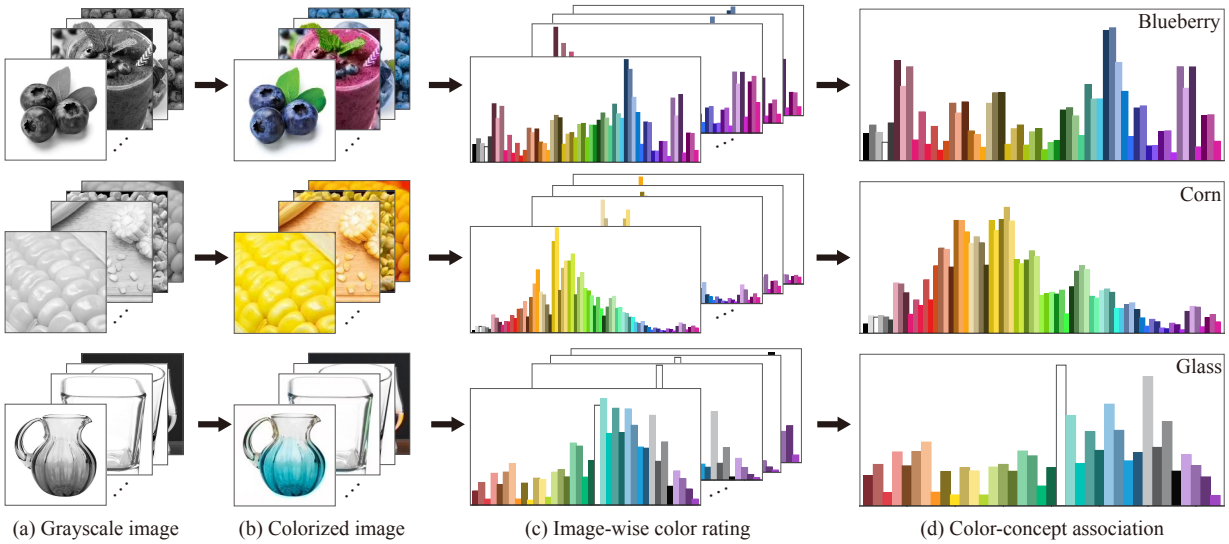


Fig. 1: We introduce a method for automatically extracting color-concept associations from natural images. We apply a colorization neural network to predict color distributions for input images. The distributions are transformed into ratings for a color library, which are then aggregated across multiple images of a concept, e.g., *blueberry*, *corn*, and *glass*, to provide the color-concept associations.

Abstract— The interpretation of colors in visualizations is facilitated when the assignments between colors and concepts in the visualizations match human's expectations, implying that the colors can be interpreted in a semantic manner. However, manually creating a dataset of suitable associations between colors and concepts for use in visualizations is costly, as such associations would have to be collected from humans for a large variety of concepts. To address the challenge of collecting this data, we introduce a method to extract color-concept associations automatically from a set of concept images. While the state-of-the-art method extracts associations from data with supervised learning, we developed a self-supervised method based on colorization that does not require the preparation of ground truth color-concept associations. Our key insight is that a set of images of a concept should be sufficient for learning color-concept associations, since humans also learn to associate colors to concepts mainly from past visual input. Thus, we propose to use an automatic colorization method to extract statistical models of the color-concept associations that appear in concept images. Specifically, we take a colorization model pre-trained on ImageNet and fine-tune it on the set of images associated with a given concept, to predict pixel-wise probability distributions in Lab color space for the images. Then, we convert the predicted probability distributions into color ratings for a given color library and aggregate them for all the images of a concept to obtain the final color-concept associations. We evaluate our method using four different evaluation metrics and via a user study. Experiments show that, although the state-of-the-art method based on supervised learning with user-provided ratings is more effective at capturing relative associations, our self-supervised method obtains overall better results according to metrics like Earth Mover's Distance (EMD) and Entropy Difference (ED), which are closer to human perception of color distributions.

Index Terms—Color-concept association, colorization, EMD

1 Introduction

People generally find it easier to interpret visualizations such as graphs and diagrams if the categories in the visualizations are represented with colors that match people's semantic expectations [15, 28, 30], since the ability to interpret visualizations also depends on the semantic discriminability of the colors in the visualization [19]. For example, reading a graph of nutritional information where strawberries are represented as red and mangos as yellow requires less conscious effort, as the two categories are associated to colors that match human perception of these fruits. In addition, Mukherjee et al. [19] showed with their semantic discriminability theory that in certain conditions colors can be associated to concepts usually not considered to be linked with colors. Thus, making color palettes semantically interpretable is useful for visualization and possible in many contexts.

Semantic palettes can be created from datasets of color-concept associations [15, 24, 30, 31]. However, manually creating such a dataset can be costly, as the dataset has to cover a large variety of concepts, leading to the collection of a large volume of data from human subjects. Thus, previous work has proposed automated methods for the extraction of associations from data. A common idea in many automated approaches is to collect color-concept associations from images asso-

- *R. Hu, Z. Ye, B. Chen, and H. Huang are with Shenzhen University, Visual Computing Research Center, China. E-mail:{ruizhen.hu, yeziqi.sd, codeb.box, hhzhiyan}@gmail.com.*
- *O. van Kaick is with Carleton University, School of Computer Science, Canada. E-mail: ovankaic@gmail.com.*
- *Hui Huang is the corresponding author of this paper.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

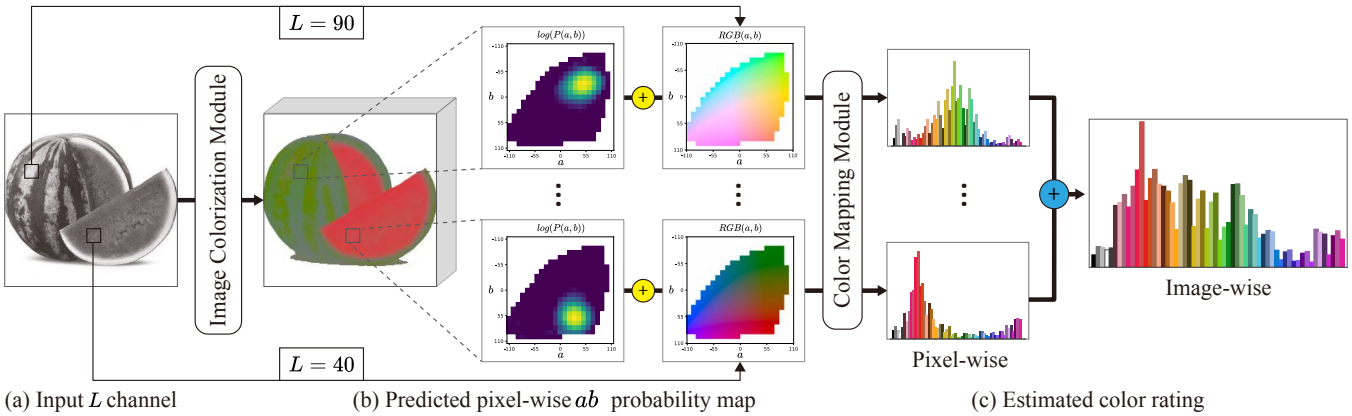


Fig. 2: Overview of our method for color-concept association. (a) Given the lightness channel (L) of an input image, we use an image colorization network to predict (b) an ab color probability distribution for each pixel. (c) Then, we convert the distributions into ratings for a color library with a color mapping module, and aggregate the ratings for all the pixels to obtain an image-wise color rating. The method further aggregates the ratings for all the images of a common concept to provide the final color-concept association.

ciated with keywords, where the data can be obtained through image retrieval from search engines according to given keywords [15, 24]. The image-keyword pairs are then further processed to provide the color-concept associations. For example, Lin et al. [15] analyze color distributions of the images and then assign colors to concepts according to an affinity score. Rathore et al. [24] learn a method to extract color distributions from images to match human color-concept associations, providing high-quality results. Although the core method is in principle unsupervised, in practice it is supervised as it relies on human ratings to guide the feature selection and weight optimization in the learning.

In this paper, we introduce a self-supervised method based on colorization for extracting color-concept associations from natural images (Fig. 1). Our key insight is that human color-concept associations are learned from past visual input [7, 18, 36]. Thus, natural images should also be sufficient input for a method that derives color-concept associations, i.e., there should be no need for explicitly collecting human ratings. In this spirit, our main idea is to extract associations from the model learned by a colorization neural network [37]. Colorization networks learn to assign colors to the pixels of a grayscale image, and thus these networks essentially learn how to associate colors to certain categories of objects in an implicit manner. Our method extracts such knowledge from a colorization network to explicitly discover color-concept associations with a self-supervised learning method.

Specifically, we train a colorization network to predict color distributions for the pixels of natural images (Fig. 2). Then, we convert these probabilities into ratings for a color library, which are finally transformed into color-concept distributions that aggregate the ratings of multiple images. Thus, the output of our method is a color probability distribution for each concept. Such color-concept distributions can then be further considered in color assignments for visualizations, e.g., through an optimization scheme that balances the frequency and distinctiveness of the colors assigned to different concepts [15].

The main advantage of our method over previous work is that, since colorization can be posed as a self-supervised learning task, the colorization network can be trained simply with a dataset of images of different concepts. The last step of our method that aggregates the ratings from multiple images requires a concept "label" for each image. However, this is also the common requirement of previous methods, and the assignment of these labels can be automated by retrieving images for a list of required concepts with a search engine, as in previous work [15, 24]. In addition, our method does not depend on feature selection based on human ratings, which could be biased towards the small number of classes for which the ratings were collected. Our method also extracts the foreground objects in the input images [22] to ensure that background regions do not negatively impact the learned color distributions.

We evaluate our method by comparing our results to the results of the state-of-the-art supervised method [24] with qualitative and quantitative analyses. Specifically, we evaluate the results quantitatively with the same measures used by previous work (Pearson correlation and total variation) but also use a perceptual measure (Earth Mover's Distance) to provide an evaluation that better reflects the perceptual differences between the results. We show that our self-supervised method provides results that are superior to the state-of-the-art method based on these perceptual measures while less effective according to non-perceptual measures due to the lack of supervision.

2 Related Work

We discuss the literature most related to our work, reviewing methods for automatic extraction of color-concept associations from data and the state-of-the-art methods for image colorization.

2.1 Color-concept association

Color-concept associations can be quantified through human judgement [12, 13, 20, 32]. However, it can be costly to create such data manually. Methods for automatic extraction of color-concept associations from data can be roughly divided into image-based [2, 15, 16] and natural language-based methods [9, 31].

Image-based methods extract associations from images paired with keywords. Lin et al. [15] extract color distributions from tagged images, and determine color-concept affinity scores from the histograms and keywords using an entropy-based method. The method provides an optimal assignment of one color to each concept. Similarly, Lindner et al. [16] use statistical methods to associate colors to concepts from annotated images.A few methods also compute color distributions for concepts. Lindner et al. [17] compute distributions in CIELAB space for over 9,000 color names in 10 different languages using images retrieved from a search engine. Bartram et al. [3] use clustering algorithms to extract palettes for images associated to affective concepts. More recently, Rathore et al. [24] perform feature selection based on human ratings to learn the best features for retrieving histograms from images, so that the histograms are likely associated to given concepts.

On the other hand, language-based methods extract color-concept associations from natural language descriptions. Setlur et al. [31] analyze co-occurrences of concepts and colors in n-grams extracted from a corpus of text, and then cluster images retrieved with the associated colors to create color palettes.

In summary, earlier work for color-concept extraction makes use of handcrafted automatic methods, which provide results that do not always correlate strongly with human selections [24]. Recent learning-based methods are more robust and validated on user ratings, but the

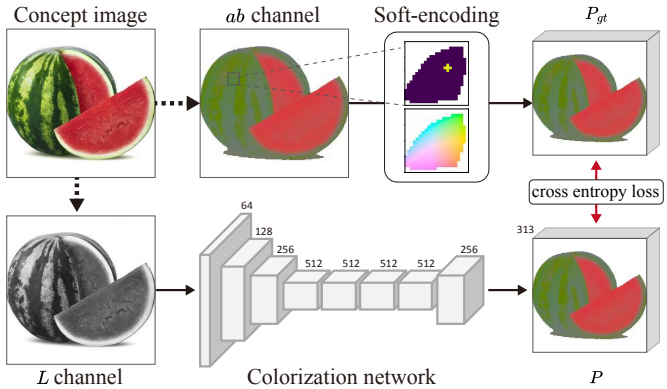


Fig. 3: Colorization network training. Top row: given an input image provided in perceptual *Lab* color space, we convert the image into a ground truth probability distribution P_{gt} in *ab* space. Bottom row: the network is trained to predict the distribution P from the lightness channel (L) via the loss that compares two distributions.

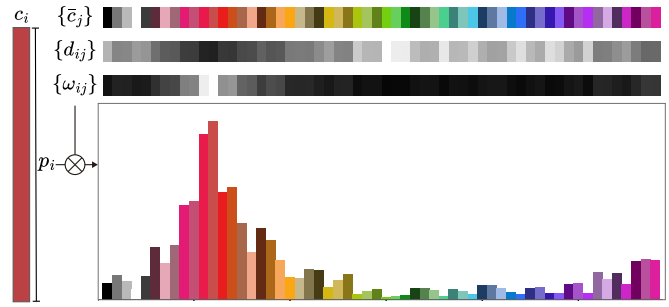


Fig. 4: Mapping from a color c_i to the color library $\{\bar{c}_j\}$, distributing the probability p_i of c_i to the colors in the library according to Equation 2.

learning is dependent on user input specific to the problem. In contrast, our work is a learning-based method that does not require data specifically prepared for the concept-color association task.

2.2 Image colorization

State-of-the-art colorization methods are based on neural networks learned on over a million images [4,5,8,10,33,34,37]. These networks can be trained in a self-supervised manner by taking color images, transforming them into grayscale, and then training the networks to predict the colored version of the image from the grayscale. Iizuka et al. [11] introduce a CNN-based colorization method that combines global priors with local features extracted from the images. Zhang et al. [37] provide a colorization method based on a convolutional neural network, which performs significantly better than past methods. This method is then used as the basis of a user-guided colorization which incorporates user hints to direct the coloring output [39]. Most of these previous methods predict single colors for each pixel in the input image.

Moreover, colorization has been used as a proxy task for visual understanding [14, 38], and in this paper, we utilize colorization to extract color-concept associations. One interesting aspect of the work of Zhang et al. [37] is that the output is a pixel-wise color distribution. Thus, we based our colorization network on this method since we also require distributions for computing color-concept associations.

3 Method

The key idea of our method is to use image colorization as a self-supervised task to extract color-concept associations automatically from natural images. A colorization network is able to assign correct colors for images with different concepts, which means that this network must implicitly learn color-concept associations. Thus, our goal is to extract the learned color-concept associations explicitly from such a network.

Fig. 2 shows an overview of our method. With the image colorization module pre-trained on ImageNet, for each color image associated with a given concept, we can predict a distribution of possible colors for each pixel by taking the lightness (L) channel as input and outputting the probability map over the quantized *ab* color space. The predicted pixel-wise color distribution is then converted to the color rating over a given color library, e.g., UW-58 colors [24] shown in this example, by weighted color mapping, and accumulated over all pixels of all images of the same concept, to provide the final color-concept association.

More details about the image colorization module and the color mapping module are given in the following two subsections.

3.1 Image colorization

We perform the colorization task in the perceptual CIE Lab color space. The goal of our image colorization module is to take the L channel of an image as input and predict a distribution of possible colors instead of a single deterministic color for each pixel, so that we are able to quantify color-concept associations over a large range of colors. As the L channel is already given as input, we only need to predict the probability map over the *ab* color space, which then combined with the input L provides all the information needed to define a full color distribution. With this goal in mind, we adopt the image colorization neural network proposed by Zhang et al. [37], where the input and output both fit our setting. The network architecture and how the training data is constructed are illustrated in Fig. 3.

Given an input lightness channel $L \in R^{H\times W}$ of an image, the goal of the colorization module is to learn a mapping to a probability distribution over possible colors $P \in [0,1]^{H\times W\times Q}$, where H and W are the image dimensions, and Q is the number of quantized *ab* values. In all our experiments, we quantize the *ab* space into bins with grid size 10 and keep the $Q = 313$ values which are in-gamut, as in [37].

To define the training data for the network, we convert the ground truth color image I into a quantized *ab* probability distribution P_{gt} using a soft-encoding scheme. More specifically, for each pixel, we find the 5-nearest neighbors in the color space determined by the quantized *ab* space and its lightness, and then weight them proportionally to their distance from the ground truth using a Gaussian kernel with $\sigma = 5$.

The loss function is then defined as the multinomial cross entropy loss between the predicted probability distribution P and the ground truth probability distribution P_{gt} :

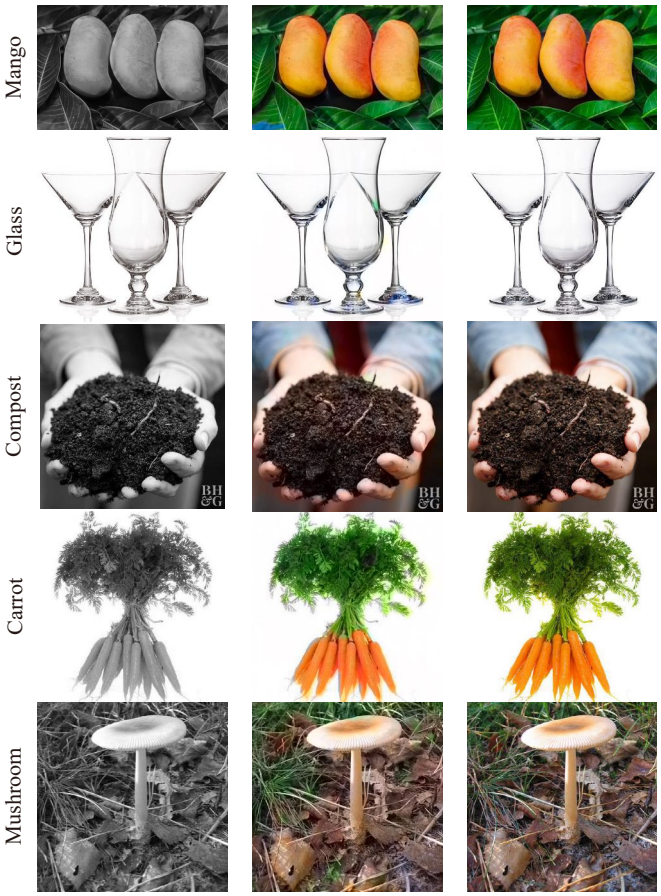
$$Loss(P, P_{gt}) = -\sum_{h,w} v(P(h,w,:)) \sum_{q} P(h,w,q) \log P_{gt}(h,w,q), \quad (1)$$

where $v(\cdot)$ is a weighting term that can be used to rebalance the loss based on color-class rarity as defined in [37], which gives less weight to desaturated values as the number of pixels in natural images at desaturated values are orders of magnitude higher than for saturated values due to the appearance of background elements such as clouds, pavement, dirt, and walls.

The network consists of 8 blocks, each of which contains 2 or 3 repeated convolutional and ReLU layers and a BatchNorm layer. The network is pre-trained on 1.3M images from the ImageNet training set [26], and fine-tuned on the given images with associated concepts to extract the color-concept associations. More details of the network and training can be found in the supplementary material.

3.2 Color mapping

The color mapping module then derives the color-concept associations from the probability distribution $P \in [0,1]^{H\times W\times Q}$ predicted by the colorization network. Specifically, for each pixel located at position (h,w), we use its lightness $L(h,w)$ to map the quantized *ab* space to the full color space with the transferred probability map $P^{h,w} = P(h,w,:) \in [0,1]^{Q}$. Then for each color c_i with probability $p_i = P_i^{h,w}$, $i = 1,\ldots,Q$, we distribute its probability p_i to the given color library $\{\bar{c}_j\}_{j=1,\ldots,N}$, e.g., UW-58 colors [24] , UW-71 colors [19] or BCP-37 colors [21], to



(a) Gray-scale image (b) Colorized image (c) Concept image

Fig. 5: Colorization results on unseen test images for different concepts. Note the overall plausibility of the colorization results.

obtain $\{\bar{p}_j^{h,w}\}_{j=1,\ldots,N}$:

$$\bar{p}_j^{h,w} = \sum_{i=1}^{Q} \omega_{ij} p_i, \tag{2}$$

where ω_{ij} are distribution weights based on the perceptual difference between two colors c_i and $\bar{c}_j$. Fig. 4 illustrates how p_i of color c_i is re-distributed to the given color library $\{\bar{c}_j\}_{j=1,\ldots,N}$.

To define ω, we first compute the Euclidean distances $\{d_{ij}\}_{j=1,\ldots,N}$ in Lab color space between c_i and all the colors in library $\{\bar{c}_j\}_{j=1,\ldots,N}$. Then, we compute their corresponding z-scores $\{z_{ij}\}_{j=1,\ldots,N}$, and apply softmax to the z-scores to get the final weights:

$$\omega_{ij} = \frac{e^{-z_{ij}}}{\sum_{j=1}^{N} e^{-z_{ij}}}. \tag{3}$$

To obtain the accumulated color-concept associations $\{\bar{p}_j\}_{j=1,\ldots,N}$ for the entire image, we simply compute the average probability distribution among all pixels:

$$\bar{p}_j = \sum_{(h,w)\in F} \bar{p}_j^{h,w} / |F|, \tag{4}$$

where F is the set of pixels inside the foreground region of the image. The foreground mask is obtained with the automatic foreground detection method of Qin et al. [22].

For a set of images associated with the same concept, we further compute the average probability distribution of all the images to get the accumulated color-concept association $\bar{p} \in [0,1]^N$. As natural images have great variability in lighting, we find that there can be regions in the images with white glow and dark shadows, which are not common colors associated to concepts and thus less useful for visualization. To reduce the impact of these colors on the computed distributions, we half the probability of the entry in $\bar{p}$ with both a and b channels equal to zero and then renormalize the distribution to obtain the final color-concept association $\hat{p}$.

(a) Concept image (b) Estimated image-wise color rating

Fig. 6: Example color ratings for different images with the same concept. Top three rows: *Apple*. Bottom three rows: *Metal*.

4 Results and Evaluation

In this section, we evaluate our method with qualitative and quantitative analyses on a variety of datasets.

4.1 Datasets

To evaluate our method, we collect four datasets from previous works with ground truth user ratings, which cover a variety of object categories, to fully explore the capability of our method.

- **Recycling6** [30] consists of around 70 images for each of 6 types of recycling items, including *Compost, Glass, Metal, Paper, Plastic,* and *Trash*. The ground truth user ratings are defined on the Berkeley Color Project 37 (BCP-37) color library [21].
- **Fruit12** [24] consists of 50 images for each of 12 fruit concepts, including *Avocado, Blueberry, Cantaloupe, Grapefruit, Honeydew, Lemon, Lime, Mango, Orange, Raspberry, Strawberry,* and

Watermelon. The ground truth user ratings are defined on the University of Wisconsin 58 (UW-58) color library [29].

- **Fruit5** [19] consists of 50 images for each of another 5 fruit concepts, including *Apple, Banana, Cherry, Grape,* and *Peach*. The ground truth user ratings are defined on the UW-71 color library, an extension of the UW-58 colors. Note that, since only the ground truth ratings are available but without the corresponding example images, we collected corresponding images through Google Image search as in [24].

- **Vegetable5** [19] consists of 50 images for each of 5 vegetable concepts, including *Carrot, Celery, Corn, Eggplant,* and *Mushroom*. The ground truth user ratings are defined on the UW-71 color library and example images were retrieved from Google search similarly as in the previous item.

4.2 Qualitative results

Our method first colorizes each concept image to predict a pixel-wise *ab* probability distribution. Then, our method extracts a color rating from the distributions of each image and finally accumulates ratings for all the images with the same concept in the dataset to provide the final color-concept associations. We provide a qualitative analysis of these three steps of the method as follows. To illustrate the results of the colorization network, we test the learned model on unseen concept images and obtain the colorized results by taking the annealed-mean of the predicted distribution as in [37]. Fig. 5 shows several example colorization results for different grayscale concept images from the unseen test set. We see that the colorized images are perceptually quite close to the ground truth.

Furthermore, Fig. 6 illustrates the second step of the method, where several example estimated color ratings are extracted from different images of the same concept. Note how the color ratings capture the different color distributions in the images. For example, the distributions for apples have high probabilities for bright colors such as red, yellow, and green, while the distributions for metals have higher probabilities for grayscale colors.

Fig. 7 shows results of the third step of the method (in the middle column), where we see the final color distributions produced for different concepts based on multiple images. Although many colors have non-zero probabilities, the colors with the probabilities that stand out the most clearly correspond to colors commonly associated to the given concepts. Fig. 1 shows a few extra results with example input images.

4.3 Evaluation metric

To quantitatively evaluate our final results, that is, to compare an estimated color-concept association $\hat{p} = \{\hat{p}_i\}_{i=1}^N$ to the ground truth user rating $r = \{r_i\}_{i=1}^N$, defined on the color library $\{c_i\}_{i=1}^N$, where N is the number of colors in the library, we use several different metrics.

The first metric is the Pearson correlation coefficient (*Corr*) used in the work of Rathore et al. [24], which measures the linear correlation between two distributions. The second metric is the total variation (TV) used in the work of Mukherjee et al. [19], which is defined as half of the L1-distance between two distributions:

$$TV(\hat{p}, r) = \frac{1}{2}\sum_{i=1}^{N} |\hat{p}_i - r_i|. \tag{5}$$

The two metrics defined above consider two color distributions as two vectors that are compared in a manner that is oblivious to the association between the vector entries and specific colors. Thus, to take the perceptual difference of the colors into consideration, we propose to use the Earth Mover's Distance (EMD) [25] as an evaluation metric. EMD has been widely used to compare two probability distributions [6, 27, 35] and is known to be closer to user perception than other metrics.

Let us suppose that $\hat{p}$ and r are two mass distributions. Then, EMD measures the cost of transporting one heap of mass to another heap, and thus computes the minimal transport effort between $\hat{p}$ and r:

$$\begin{array}{lrcll} \min\limits_{f_{ij}} & \sum_{i=1}^N \sum_{j=1}^N f_{ij} d_{ij}, & & & \\ \text{s.t.} & f_{ij} & \geq & 0, & i,j \in \{1,2,\ldots,N\} \\ & \sum_{j=1}^N f_{ij} & \leq & \hat{p}_i, & i \in \{1,2,\ldots,N\} \\ & \sum_{i=1}^N f_{ij} & \leq & r_j, & j \in \{1,2,\ldots,N\} \\ & \sum_{i=1}^N \sum_{j=1}^N f_{ij} & = & 1, & \end{array} \tag{6}$$

where f_{ij} is the amount of mass transported from color c_i in $\hat{p}$ to color c_j in r, and d_{ij} is the Lab color distance between c_i and c_j. Note that as both $\hat{p}$ and r are normalized, the total amount of the flow is always equal to one as shown in the last constraint above. In addition, EMD is independent of the order in which the colors are stored in the distributions, since the color distances are taken into consideration for finding the optimal mass transport. Once the transportation problem is solved to obtain the optimal flow f_{ij}^*, the EMD is defined as the total cost of the flow:

$$EMD(\hat{p}, r) = \sum_{i=1}^{N}\sum_{j=1}^{N} f_{ij}^* d_{ij}. \tag{7}$$

Besides the three metrics discussed above that compute differences between distributions directly, as indicated in the work of Mukherjee et al. [19], *specificity* is one of the key properties that a color-concept association should possess, which refers to the 'peakiness' of the color-concept association distribution. Specificity is quantified using the entropy of the distribution, which captures how 'flat' vs. 'peaky' a distribution is, regardless of how many peaks there are. Thus, we also compute the entropy difference (ED) between two distributions as an auxiliary metric to quantify the similarity of specificity of two distributions:

$$ED(\hat{p}, r) = \left| \sum_{i=1}^{N} \hat{p}_i \log \hat{p}_i - \sum_{i=1}^{N} r_i \log r_i \right|. \tag{8}$$

4.4 Comparison to the state-of-the-art method

In this section, we compare our method to the state-of-the-art supervised method [24] on different datasets using all evaluation metrics as well as through a user study.

The results of the supervised method are obtained by cross-validation within each dataset. For example, for each fruit concept inside the Fruit12 dataset, we train the model of Rathore et al. [24] with the ground truth ratings on 11 fruit concepts and validate the model on the remaining concept by comparing the derived color rating to the ground truth color rating. Note that, in contrast, our method is self-supervised and can be trained directly with the images associated to each concept without any ground truth rating.

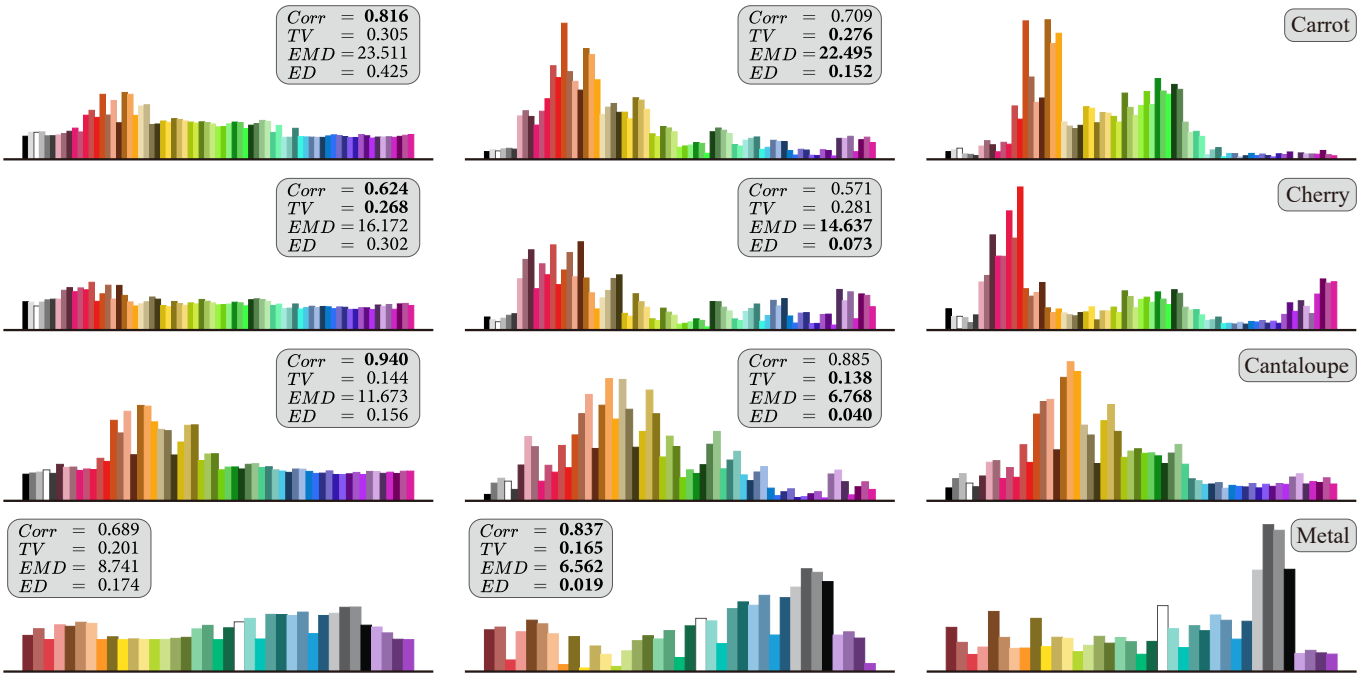
Comparison via evaluation metrics. Table 1 shows the comparison of the average evaluation scores among all the concepts for each dataset. We see that the values of different metrics are inconsistent with each other. The results obtained by the supervised method [24] have a higher linear correlation with the ground truth rating, while our results have lower EMD and ED values. The TV scores of the two methods are comparable.

Comparison via a user study. We further conduct a user study to compare the results based on human perception. For each concept in our dataset, we show the ground truth user rating together with the two estimated ratings obtained by our method and the supervised method [24] in random order, and ask the user to select which result they think is more similar to the ground truth user rating. The user can select either one of the two results or a "not sure" option.

We invited 30 participants to do the study for all the concepts in our dataset, which resulted in 840 answers in total. The vote percentages for the options "ours/the supervised method [24]/not sure" are 70.5%/16.9%/12.6%. We see that our method was selected much more frequently than the supervised method [24] as having better results, which shows that our method provides color-concept associations that are perceptually more similar to the ground truth user ratings.

Table 1: Comparison to the state-of-the-art supervised method [24] on all the datasets with the four different evaluation metrics introduced in Sect. 4.3. Our self-supervised method provides overall better results according to the EMD and ED metrics, but worse results according to the Correlation (Corr) metric. The results according to the total variation (TV) are comparable. For each metric, ↑: the higher, the better; ↓: the lower, the better.

Method	Recycling6				Fruit12				Fruit5				Vegetable5			
	Corr↑	TV↓	EMD↓	ED↓	Corr↑	TV↓	EMD↓	ED↓	Corr↑	TV↓	EMD↓	ED↓	Corr↑	TV↓	EMD↓	ED↓
Supervised [24]	0.682	0.169	8.242	0.120	**0.811**	**0.185**	12.356	0.185	**0.627**	0.308	21.710	0.383	**0.672**	**0.310**	22.985	0.384
Our method	**0.717**	**0.163**	**8.043**	**0.034**	0.697	0.210	**12.346**	**0.068**	0.497	**0.301**	**18.448**	**0.204**	0.510	0.323	**20.432**	**0.205**



(a) Supervised[24] (b) Our result (c) GT color-concept association

Fig. 7: Results obtained by the supervised method [24] (left) and our self-supervised method (middle) compared to the ground truth ratings (right). Each row shows one example concept from one dataset, including *Metal* in **Recycling6**, *Cantaloupe* in **Fruit12**, *Cherry* in **Fruit5**, and *Carrot* in **Vegetable5**. The scores of all four evaluation metrics are shown with each result and the best evaluation scores are shown in bold.

Table 2: Consistency between each metric and the user choices collected in the user study.

Metrics	Corr	TV	EMD	ED
Consistency	37.7%	45.6%	54.2%	70.5%

Analysis of the results. To obtain a better understanding of the results, we show all the metrics together with the user vote percentages for each concept in Table 3 and report the consistency between each metric and the user choices in Table 2. More specifically, for each metric, we consider that it is consistent with the user choice if the result regarded to be better using this metric is selected by the user. Then, we compute the percentages of the consistent results over all the 840 answers we collected as the final consistency score for each metric.

From the results shown in Table 2, we see that the EMD and ED metrics are more consistent with the user choices. Especially, the ED metric obtains the highest consistency score, which shows that humans focus more on the peakiness when comparing two distributions. Compared to Corr and TV, the EMD metric considers the perceptual distance between different colors and computes the minimal ground transport effort, which also leads to results more consistent with human perception.

Fig. 7 shows several visual examples of the predicted color-concept associations compared to the ground truth rating and the corresponding values of different evaluation metrics. For the results shown in the first three rows, the correlations of the results obtained by the supervised method are always higher than our method, although the distributions obtained by our method are perceptually more similar to the ground truth distributions shown on the right. The EMD and ED metrics better capture this difference. For the result shown on the bottom row, our method performs consistently better than the supervised method with respect to all the evaluation metrics, while some dominant colors do not stand out in the relatively uniform distribution obtained by the supervised method [24], which may be caused by the weighted combination of distributions obtained by different features in their method.

Note that, although our method obtains overall better results accordingly to the EMD and ED metrics, which consider the scale of association with a given concept and are more consistent with human perception, the supervised method [24] obtains overall better results according to the correlation metric. Correlation assesses the relative pattern of associations, regardless of the scale, which has been shown as the critical factor that influences semantic discriminability of the colors more than the absolute associations [19, 29, 30]. We believe that it would be worthy to explore ways to be able to handle both relative association and range well.

Table 3: Comparison to the state-of-the-art supervised method [24] on each concept in our dataset with the four different evaluation metrics as well as the user study result. For each concept, the results of the supervised method [24] are the top row and our results are the bottom row. For each metric, ↑: the higher, the better; ↓: the lower, the better.

Dataset	Concept	Metrics					Dataset	Concept	Metrics				
		Corr↑	TV↓	EMD↓	ED↓	Users↑			Corr↑	TV↓	EMD↓	ED↓	Users↑
Recycling6	Trash	0.224	**0.221**	**10.558**	0.116	10.0%	Fruit12	Orange	**0.935**	0.154	9.867	0.191	16.7%
		0.487	0.230	12.836	**0.012**	**73.3%**			0.918	**0.143**	**8.622**	**0.024**	**76.7%**
	Glass	0.809	0.185	10.109	0.148	26.7%		Raspberry	0.448	**0.265**	19.683	0.179	3.3%
		0.829	**0.141**	**6.837**	**0.001**	**66.7%**			**0.453**	0.268	**18.568**	**0.041**	**80.0%**
	Compost	**0.755**	**0.156**	**7.896**	0.110	6.7%		Strawberry	**0.662**	**0.242**	**11.110**	0.174	13.3%
		0.743	0.161	8.909	**0.024**	**90.0%**			0.589	0.266	15.016	**0.054**	**80.0%**
	Paper	0.833	0.131	6.564	0.100	13.3%		Watermelon	**0.648**	**0.241**	**10.636**	0.190	**43.3%**
		0.917	**0.124**	**5.704**	**0.097**	**86.7%**			0.398	0.293	15.344	**0.049**	40.0%
	Plastic	**0.782**	**0.120**	**5.583**	0.073	26.7%	Fruit5	Apple	**0.690**	**0.283**	**21.838**	0.371	10.0%
		0.492	0.155	7.411	**0.050**	**50.0%**			0.418	0.314	21.947	**0.220**	**70.0%**
	Metal	0.689	0.201	8.741	0.174	6.7%		Banana	**0.875**	0.354	31.448	0.556	3.3%
		0.837	**0.165**	**6.562**	**0.019**	**90.0%**			0.725	**0.302**	**21.284**	**0.374**	**80.0%**
Fruit12	Avocado	**0.760**	**0.227**	15.369	0.258	23.3%		Cherry	**0.624**	**0.268**	16.172	0.302	6.6%
		0.591	0.269	**14.739**	**0.096**	**56.7%**			0.571	0.281	**14.637**	**0.073**	**83.3%**
	Blueberry	**0.844**	**0.214**	**16.878**	0.261	**66.7%**		Grape	**0.122**	**0.332**	**22.470**	0.289	3.3%
		0.351	0.307	20.173	**0.117**	30%			-0.088	0.416	25.805	**0.198**	**46.7%**
	Cantaloupe	**0.940**	0.144	11.673	0.156	13.3%		Peach	0.827	0.304	16.623	0.397	10.0%
		0.885	**0.138**	**6.768**	**0.040**	**86.7%**			**0.856**	**0.191**	**8.566**	**0.154**	**90.0%**
	Grapefruit	0.838	**0.150**	**7.063**	0.113	10.0%	Vegetable5	Carrot	**0.816**	0.305	23.511	0.425	0.0%
		0.849	0.152	11.801	**0.080**	**86.7%**			0.709	**0.276**	**22.495**	**0.152**	**83.3%**
	Honeydew	**0.906**	**0.112**	**8.161**	0.095	30.0%		Celery	**0.766**	**0.402**	34.760	0.614	33.3%
		0.802	0.168	8.977	**0.083**	**60.0%**			0.376	0.450	**31.335**	**0.492**	**23.3%**
	Lemon	**0.919**	**0.133**	12.309	0.170	3.3%		Corn	**0.806**	0.273	24.132	0.343	10.0%
		0.911	0.140	**7.697**	**0.071**	**93.3%**			0.745	**0.242**	**15.786**	**0.165**	**73.3%**
	Lime	**0.913**	**0.184**	14.934	0.259	**60.0%**		Eggplant	**0.494**	**0.295**	**18.965**	0.307	13.3%
		0.705	0.250	**13.068**	**0.121**	33.3%			0.179	0.388	22.474	**0.150**	**63.3%**
	Mango	**0.916**	0.151	10.590	0.176	6.7%		Mushroom	0.479	0.277	13.555	0.232	3.3%
		0.916	**0.126**	**7.380**	**0.037**	**86.7%**			**0.541**	**0.261**	**10.072**	**0.066**	**93.3%**

4.5 Ablation studies

To justify several key design choices of our method, we conduct ablation studies with the following settings:

- No Fine-tuning: we use the model pre-trained on the ImageNet dataset directly, without fine-tuning it on the concept images;
- No Pre-training: we train the colorization network directly on the concept images, without using the network pre-trained on the large-scale ImageNet dataset;
- No Seg-mask: we make use of all the pixels in the concept images without masking out the ones in the background during the color mapping;
- No Post-processing: we take the mapped distribution directly without halving the probability of colors with both *ab* channels equal to zero to reduce the effect of lighting in real images.

Table 4 shows the results of the ablation studies on all four datasets. We see that our method with the full pipeline ("Our method" in the table) provides consistently better results.

For the *No Fine-tuning* setting, the network learns information purely from images with various concepts in ImageNet, which makes the distribution extracted not specifically related to the concepts in our datasets and thus leads to the worst results. For the *No Pre-training* setting, the network learns associations between colors and concepts from the specific concept images in our datasets, which leads to much better results than the *No Fine-tuning* setting. This shows that our method can obtain reasonable results with a small dataset. However, since each dataset is quite small with only 50 images per concept, this makes the network sometimes biased by the samples given in the datasets and thus does not learn a more general color-concept association. The performance becomes consistently better when the colorization network is pre-trained on ImageNet and fine-tuned on the concept images. Fig. 8 shows an example concept image and the colorization and color ratings obtained under different training settings.

Table 4: Ablation studies to justify the key components of our method. Details of the different settings can be found in Sect. 4.5. For each metric, ↑: the higher, the better; ↓: the lower, the better.

Method	Recycling6				Fruit12				Fruit5				Vegetable5			
	Corr↑	TV↓	EMD↓	ED↓	Corr↑	TV↓	EMD↓	ED↓	Corr↑	TV↓	EMD↓	ED↓	Corr↑	TV↓	EMD↓	ED↓
No Fine-tuning	0.590	0.180	8.626	0.050	0.460	0.279	17.172	0.059	0.347	0.353	22.144	0.214	0.366	0.358	22.549	0.236
No Pre-training	0.717	0.174	8.753	0.043	0.660	0.223	12.905	0.078	0.429	0.326	20.624	0.214	0.492	0.331	20.845	0.209
No Seg-mask	0.687	0.173	8.264	0.048	0.546	0.249	15.120	0.073	0.313	0.347	22.526	0.256	0.333	0.360	23.233	0.262
No Post-processing	**0.757**	0.166	8.524	0.067	0.642	0.227	13.253	**0.059**	0.409	0.329	20.188	0.231	**0.517**	0.326	21.309	0.225
Our method	0.717	**0.163**	**8.043**	**0.034**	**0.697**	**0.210**	**12.346**	0.068	**0.497**	**0.301**	**18.448**	**0.204**	0.510	**0.323**	**20.432**	**0.205**

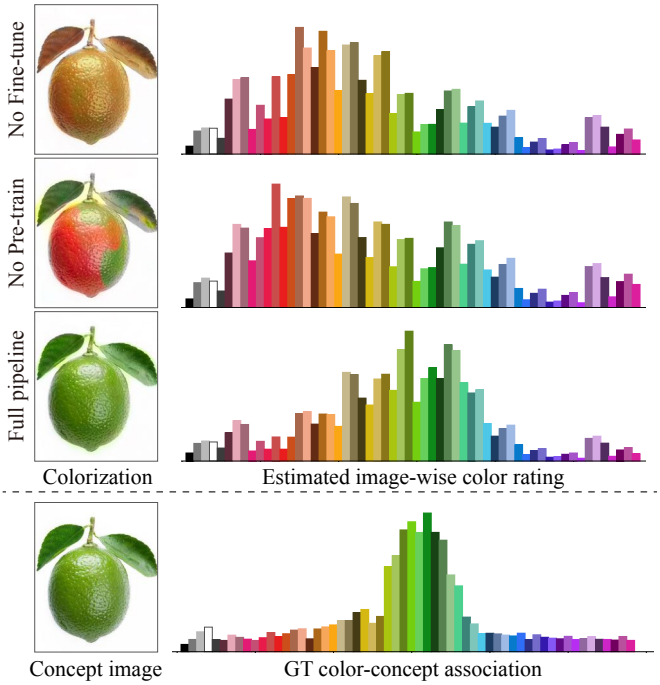


Fig. 8: Comparison of colorization and corresponding color ratings obtained under different training settings.

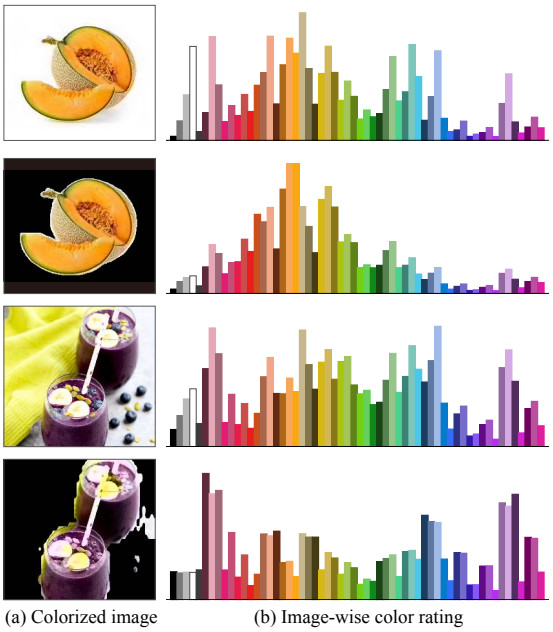


Fig. 9: Example color ratings extracted from concept images with and without foreground segmentation masks.

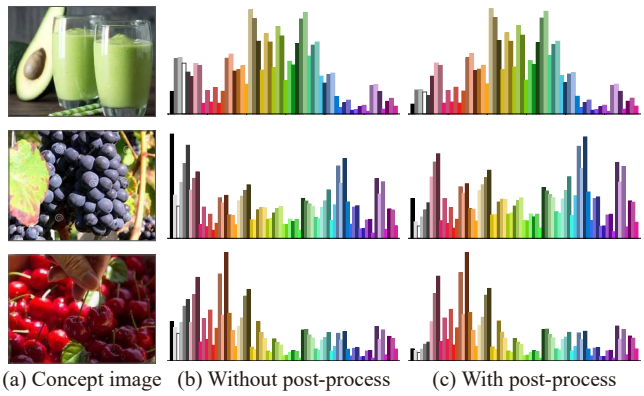


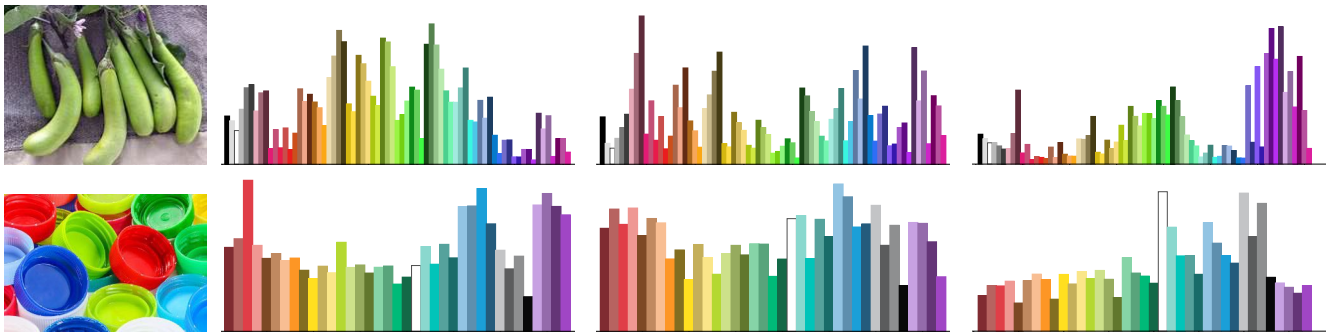
Fig. 10: Example concept images with glow and shadows and their corresponding color ratings with and without post-processing.

For the *No Seg-mask* setting, we find that the background can add some random noise or bias to the color distribution. Fig. 9 shows a comparison of several example results with and without using the foreground segmentation mask automatically obtained by the method of Qin et al. [22]. We can see that for the *cantaloupe* shown in the first row, the rating of the white color is extremely high due to the white background, and becomes more reasonable when the background is masked out. For the *blueberry* shown in the third row with a more noisy background, the color rating has clearer peaks after masking out the background, even though the segmentation is not perfect.

For the *No Post-processing* setting, we see that the correlations of the Recycle6 and Vegetable5 datasets are better than for our full method, while all the other metrics are better in the full method. We find that the main reason is that there are concepts in these two datasets where white or grey colors dominate the images, e.g., *plastic* and *mushroom*. Thus, decreasing the probability of colors with both *ab* channels equal to zero may lead to worse results for these concepts. However, when averaging the performance on the whole dataset among all the concepts, post-processing generally provides better results. Fig. 10 shows several representative concept images with significant glow and shadows, leading to the high probability of white or black colors, which is alleviated through post-processing.

4.6 Data visualization application

To demonstrate how our method can be used for data visualization applications, we follow the method of [15] to automatically select semantically-resonant colors based on the color-concept associations obtained by our method. We take the set of company brands used



(a) Concept image (b) Estimated image-wise color rating (c) Our color-concept association (d) GT color-concept association

Fig. 11: Failure cases due to the variety of images in the dataset. Top row: our method can capture the dominant green color in the given eggplant image, but as there are only very few images with green eggplants in the datasets, and most of eggplants have dark colors such as purple, the final color-concept rating that our method extracted has low probabilities for green colors compared to the ground truth user rating. Bottom row: most of the images given in the dataset for the Plastic concept are colorful, which leads to a near-uniform distribution of the final color distribution.

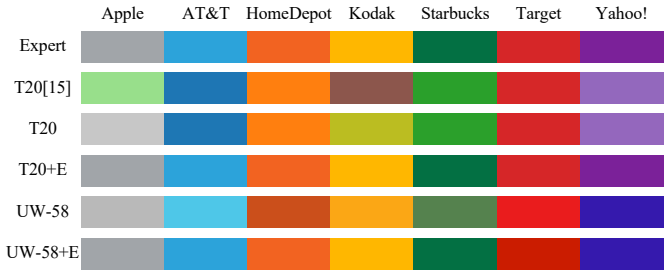


Fig. 12: Semantically-resonant colors selected for different brands using different methods when given different color palettes.

in [15], including *Apple, AT&T, HomeDepot, Kodak, Starbucks, Target*, and *Yahoo!*, as the running example, to show the color selection process. More results can be found in the supplementary materials.

We start with collecting images for each concept via search engines as we did for the Fruit5 and Vegetable5 datasets. Then, the network pre-trained on ImageNet is fine-tuned to obtain the color-concept association for each concept as described in Sect. 3.1. The color selection method in [15] is finally applied to select the final color assignment given any pre-defined color palette.

Fig. 12 shows the semantically-resonant colors selected for different brands using different methods when given different color palettes. The expert-chosen colors are shown in the first row and the result of [15] selected from the Tableau20 palette is shown in the second row for comparison. The following four rows show the colors selected based on the color-concept associations obtained by our method but from different color palettes, where "+E" indicates that the set of expert-chosen colors is included in the color palettes.

We see that, due to the limited color choices provided in the Tableau20 palette, most of the colors selected by our method are slightly different from the expert-chosen ones. However, once we add those expert-chosen colors into the palette for selection, all the chosen colors match exactly with expert choices. When given UW-58 as the color palette, the chosen colors are closer to expert choices than those from the Tableau20 palette, which also shows that our method is able to select more appropriate colors for different concepts from a larger color library. Again, when further provided with expert-chosen colors, most of the selected colors are improved other than the one for *Yahoo!*. When checking the selection process in detail, we find that although the expert-chosen purple color does have a higher association in our method, the color entropy is slightly larger than the final color that our method chooses, which leads to a slightly lower affinity score. It would be interesting to explore other color selection methods based on our color-concept associations instead of using the method in [15].

5 Conclusion

We introduced a method for extracting color-concept associations from datasets of natural images via colorization. We showed that our method leads to improved results over the state-of-the-art supervised method according to evaluation metrics more consistent with human perception, while requiring minimal data preparation, since colorization networks can be trained in a self-supervised manner. Thus, the method can be easily applied to extract color associations for concepts beyond the ones tested in our work, since the main data preparation task is to gather images of the new concepts.

Limitations and future work. Since our method is unsupervised, our results are entirely determined by the data provided as input to the method. As an example of this limitation, a lack of variety in the images of the datasets can lead to failure cases, as shown in Fig. 11. Thus, it is an interesting direction for future work to explore ways to automatically select the set of representative images for each concept for color rating extraction. Moreover, the simple post-processing applied by our method can reduce the negative effect of glow and shadows caused by certain lighting conditions to some extent, but it may lead to worse results for some specific categories. Thus, it would be interesting to explore more sophisticated ways to resolve this problem.

Furthermore, although our results are perceptually better, the relative patterns of association are more effectively captured by the state-of-the-art supervised method [24]. It would be interesting to explore ways of handling both relative association and range well, for example, by providing light supervision to adjust the color rating obtained by our method to better capture the relative pattern. Last but not least, as the colorization network is pre-trained on ImageNet, which consists of images of real objects, when fine-tuning it using images searched for abstract concepts without specific semantically-resonant visual representations, the final color-concept association might be unreasonable. We believe that this is a common limitation for unsupervised methods relying on images only and it would also be interesting to investigate ways of extending our method to make it work well for abstract concepts. One possible solution could be the automatic selection or even generation of representative images for each abstract concept based on the recent CLIP model [23] that connects text and images.

Acknowledgments

We thank the reviewers for their valuable comments. This work was supported in parts by NSFC (61872250, U21B2023, 62161146005), GD Natural Science Foundation (2021B1515020085), GD Talent Plan (2019JC05X328), Shenzhen Science and Technology Program (RCYX20210609103121030, RCJC20200714114435012), Natural Sciences and Engineering Research Council of Canada, and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ).

REFERENCES

[1] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014.

[2] L. Bartram, A. Patra, and M. Stone. Affective color in visualization. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 1364–1374, 2017.

[3] L. Bartram, A. Patra, and M. Stone. Affective Color in Visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pp. 1364–1374. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3026041

[4] Y. Cao, Z. Zhou, W. Zhang, and Y. Yu. Unsupervised diverse colorization via generative adversarial networks. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 151–166. Springer, 2017.

[5] A. Deshpande, J. Lu, M.-C. Yeh, M. Jin Chong, and D. Forsyth. Learning diverse image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6837–6845, 2017.

[6] S. Dutta, C.-M. Chen, G. Heinlein, H.-W. Shen, and J.-P. Chen. In situ distribution guided analysis and visualization of transonic jet engine simulations. *IEEE transactions on visualization and computer graphics*, 23(1):811–820, 2016.

[7] A. J. Elliot, M. A. Maier, A. C. Moller, R. Friedman, and J. Meinhardt. Color and psychological functioning: the effect of red on performance attainment. *Journal of experimental psychology: General*, 136(1):154, 2007.

[8] C. Furusawa, K. Hiroshiba, K. Ogaki, and Y. Odagiri. Comicolorization: semi-automatic manga colorization. In *SIGGRAPH Asia 2017 Technical Briefs*, pp. 1–4. 2017.

[9] C. Havasi, R. Speer, and J. Holmgren. Automated Color Selection Using Semantic Knowledge. In *2010 AAAI Fall Symposium Series*, Nov. 2010.

[10] M. He, D. Chen, J. Liao, P. V. Sander, and L. Yuan. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)*, 37(4):1–16, 2018.

[11] 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*, 35(4):110:1–110:11, 2016. doi: 10.1145/2897824.2925974

[12] D. Jonauskaite, A. M. Abdel-Khalek, A. Abu-Akel, A. S. Al-Rasheed, J.-P. Antonietti, Á. G. Ásgeirsson, K. A. Atitsogbe, M. Barma, D. Barratt, V. Bogushevskaya, et al. The sun is no fun without rain: Physical environments affect how we feel about yellow across 55 countries. *Journal of Environmental Psychology*, 66:101350, 2019.

[13] D. Jonauskaite, J. Wicker, C. Mohr, N. Dael, J. Havelka, M. Papadatou-Pastou, M. Zhang, and D. Oberfeld. A machine learning approach to quantify the specificity of colour–emotion associations and their cultural differences. *Royal Society open science*, 6(9):190741, 2019.

[14] G. Larsson, M. Maire, and G. Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6874–6883, 2017.

[15] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting Semantically-Resonant Colors for Data Visualization. *Computer Graphics Forum*, 32(3pt4):401–410, 2013. doi: 10.1111/cgf.12127

[16] A. Lindner, N. Bonnier, and S. Süsstrunk. What is the Color of Chocolate? – Extracting Color Values of Semantic Expressions. *Conference on Colour in Graphics, Imaging, and Vision*, 2012(1):355–361, Jan. 2012.

[17] A. Lindner, B. Z. Li, N. Bonnier, and S. Süsstrunk. A Large-Scale Multi-Lingual Color Thesaurus. *Color and Imaging Conference*, 2012(1):30–35, Jan. 2012.

[18] L. W. MacDonald, C. P. Biggam, and G. V. Paramei. *Progress in colour studies: cognition, language and beyond*. John Benjamins Publishing Company, 2018.

[19] K. Mukherjee, B. Yin, B. E. Sherman, L. Lessard, and K. B. Schloss. Context Matters: A Theory of Semantic Discriminability for Perceptual Encoding Systems. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):697–706, 2022. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2021.3114780

[20] L.-C. Ou, M. R. Luo, A. Woodcock, and A. Wright. A study of colour emotion and colour preference. part i: Colour emotions for single colours. *Color Research & Application*, 29(3):232–240, 2004.

[21] S. E. Palmer and K. B. Schloss. An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, 107(19):8877–8882, 2010.

[22] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 106:107404, Oct. 2020. doi: 10.1016/j.patcog.2020.107404

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

[24] R. Rathore, Z. Leggon, L. Lessard, and K. B. Schloss. Estimating Color-Concept Associations from Image Statistics. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1226–1235, 2019. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2019.2934536

[25] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, Nov. 2000. doi: 10.1023/A:1026543900054

[26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[27] B. E. Ruttenberg and A. K. Singh. Indexing the earth mover's distance using normal distributions. *Proceedings of the VLDB Endowment*, 5(3), 2011.

[28] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang. Mapping color to meaning in colormap data visualizations. *IEEE Trans. Vis. Comput. Graph.*, 25(1):810–819, 2019.

[29] K. B. Schloss, Z. Leggon, and L. Lessard. Semantic discriminability for visual communication. *Ieee transactions on visualization and computer graphics*, 27(2):1022–1031, 2020.

[30] K. B. Schloss, L. Lessard, C. S. Walmsley, and K. Foley. Color inference in visual communication: the meaning of colors in recycling. *Cognitive Research: Principles and Implications*, 3(1):5, Feb. 2018.

[31] V. Setlur and M. C. Stone. A Linguistic Approach to Categorical Color Assignment for Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):698–707, 2015. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2015.2467471

[32] D. S. Y. Tham, P. T. Sowden, A. Grandison, A. Franklin, A. K. W. Lee, M. Ng, J. Park, W. Pang, and J. Zhao. A systematic investigation of conceptual color associations. *Journal of Experimental Psychology: General*, 149(7):1311, 2020.

[33] P. Vitoria, L. Raad, and C. Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2445–2454, 2020.

[34] S. Wan, Y. Xia, L. Qi, Y.-H. Yang, and M. Atiquzzaman. Automated colorization of a grayscale image with seed points propagation. *IEEE Transactions on Multimedia*, 22(7):1756–1768, 2020.

[35] Y. Wang, F. Han, L. Zhu, O. Deussen, and B. Chen. Line graph or scatter plot? automatic selection of methods for visualizing trends in time series. *IEEE transactions on visualization and computer graphics*, 24(2):1141–1154, 2017.

[36] C. Witzel, H. Valkova, T. Hansen, and K. R. Gegenfurtner. Object knowledge modulates colour appearance. *i-Perception*, 2(1):13–49, 2011.

[37] R. Zhang, P. Isola, and A. A. Efros. Colorful Image Colorization. In B. Leibe, J. Matas, N. Sebe, and M. Welling, eds., *Computer Vision – ECCV 2016*, pp. 649–666. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-46487-9_40

[38] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1058–1067, 2017.

[39] 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*, 36(4):119:1–119:11, 2017. doi: 10.1145/3072959.3073703