# Color Composition Similarity and Its Application in Fine-grained Similarity

Mai Lan Ha
University of Siegen

Vlad Hosu
University of Konstanz

Volker Blanz
University of Siegen

{hamailan, blanz}@informatik.uni-siegen.de
vlad.hosu@uni-konstanz.de

## Abstract

*Assessing visual similarity in-the-wild, a core ability of the human visual system, is a challenging problem for computer vision methods because of its subjective nature and limited annotated datasets. We make a stride forward, showing that visual similarity can be better studied by isolating its components. We identify color composition similarity as an important aspect and study its interaction with category-level similarity. Color composition similarity considers the distribution of colors and their layout in images. We create predictive models accounting for the global similarity that is beyond pixel-based and patch-based, or histogram level information. Using an active learning approach, we build a large-scale color composition similarity dataset with subjective ratings via crowd-sourcing, the first of its kind. We train a Siamese network using the dataset to create a color similarity metric and descriptors which outperform existing color descriptors. We also provide a benchmark for global color descriptors for perceptual color similarity. Finally, we combine color similarity and category level features for fine-grained visual similarity. Our proposed model surpasses the state-of-the-art performance while using three orders of magnitude less training data. The results suggest that our proposal to study visual similarity by isolating its components, modeling and combining them is a promising paradigm for further development.*

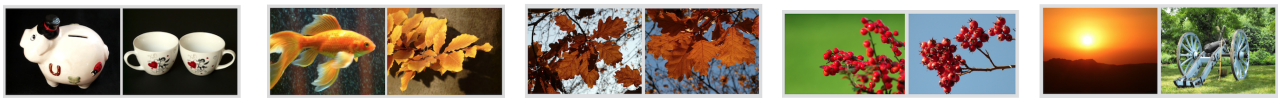
## 1. Introduction

Visual similarity is a long-standing research problem that has not been studied thoroughly. Its challenges come from the ambiguity in the problem definition as well as the subjective evaluation due to individual human perception. There are many factors that contribute to the overall visual similarity evaluation such as object categories, image composition, color layout, image style, etc.

The goal of our paper is to study the fundamental problem of visual similarity and propose novel ways to reduce the ambiguity in order to create better predictive models. We break down visual similarity into sub-problems (category and color similarity), finding a way to collect meaningful training data, and developing metrics and descriptors for color similarity. In contrast to existing approaches, we study visual color similarity "in-the-wild", which goes beyond pixel-based or patch-based approaches. For individual colors, we have a good standard to measure the perceptual similarity that is established via the CIE $\triangle E2000$ metric [20, 11]. However, multiple colors interact in complex ways in natural images. Existing metrics like CIE $\triangle E^*$ and handcrafted color descriptors [2, 12, 29, 5, 4, 21, 18, 15, 28] are not able to accurately predict color composition similarity.

While recent methods learn visual features for image search and visual similarity [6, 30, 23, 3, 31], they lack a dataset built directly from human judgments on color similarity for training and validation. These methods are trained on datasets that are labeled with object categories [7, 9]. Another set of approaches try to separate aspects of perception, by discovering and learning visual attributes for image search and retrieval [10, 32, 22, 8, 16, 24, 26]. The information they rely on involves textual description, attribute labeling and supervised learning on attribute labels. Attributes simplify the objectives of visual similarity by mapping a full range of perception into a discrete set of textual descriptions. Our approach is to ask participants to visually compare and rate images directly without having to use less accurate means of assessment such as textual descriptions.

As stated in [33], fine-grained similarity comparisons (including color) are critical for building perceptually accurate models. However, it is very difficult to measure the color similarity for images in-the-wild due to the high complexity of natural images and the subjectivity of perceptual judgments. Therefore, we introduce a new way to define visual color similarity, as color composition, and study it directly via human evaluations. The color composition assessment emphasizes hues and shades, their distributions and overall layout, independent of the semantic category. We create a dataset annotated with 5-point similarity ratings for color composition. This contrasts with other image similarity datasets which often rely on binary labels such



(a) score ≈ 5 (b) score ≈ 4 (c) score ≈ 3 (d) score ≈ 2 (e) score ≈ 1

Figure 1: Examples of ratings for color composition similarity on a scale of 1 (lowest similarity) to 5 (highest similarity).

as INRIA Holidays [14] or the triplets dataset [30]. One of the challenges for building a fine-grained similarity rating large-scale dataset is the cold start problem that arises from the very low probability in obtaining similar image pairs if we were to sample them randomly. We overcome this by using an active learning approach and iterating from binary to fine-grained ratings. We also account for many measures to ensure the quality of the dataset (Section 3). As a result, we contribute a large-scale (31,248 image pairs with at least 20 ratings each), high quality (ICC of 0.69, very high for crowd-sourcing) and novel dataset for color composition similarity. Our dataset is the first annotated dataset of its kind up to date, to the best of our knowledge [1].

Using the dataset, we train a Siamese network to predict the distributions and mean opinion scores (Section 4). The network serves as a metric and a feature extractor for color composition similarity. We compare performances and create a benchmark for existing color descriptors and our trained color features in the field of color similarity for images in-the-wild ( Section 3.3). Trained global features using CNNs produce a very good performance (0.913 SROCC, Spearman correlation w.r.t the ground-truth). Even though L1 and L2 measuring on existing hand-crafted local descriptors with dense samplings yield lower performances, it is promising to train these descriptors to capture global features of color composition, leading to better performances (the best case is 0.862 SROCC with HueSIFT).

Furthermore, we validate our color features and metrics in a fine-grained similarity application (Section 6). Color had been previously modeled implicitly together with category. We propose a novel approach to combine category features via pair-wise correlations and color similarity as predicted from our models. These combined features are extracted from pairs of images leading to improvements in accuracy compared to learning on individual content or color features alone. Training an SVM using our proposed features on a small dataset yields better accuracy than the state of the art. Compared to a common baseline, the best existing method DeepRanking [30] trained on millions of images achieves a relative improvement of 3.5%. Our model, trained on less than 50K images in total, improves by a much higher margin of 12.5%. Despite using three orders of magnitude less training data, the absolute improvement of our method on different validation sets is still better than the state of the art, 86.2% for ours vs their 85.7%.

In summary, our contributions are: (i) A general framework for modeling sub-aspects of image similarity, that is designed to handle highly subjective measurements via crowd-sourcing and active learning. (ii) The first large-scale perceptual color composition similarity dataset in-the-wild with 5-point ratings. (iii) A global color similarity benchmark for color descriptors.(iv) A new type of brief but highly generalizing features for fine-grained similarity: the concatenation of the correlation of category features and color similarity extracted from pairs of images. Triplet ranking using SVM on these features surpasses the state of the art even when trained on a much smaller dataset.

## 2. Related Work

**Hand-crafted Features for Color Similarity:** From the famous SIFT descriptor [19] that describes local features for a set of interesting points in an image by histograms of gradient orientations, different extensions of SIFT are derived for color descriptors [2, 12, 29, 5, 4]. A common objective of these descriptors is to be robust against changes in lighting, scale, rotation, and so on. A complete evaluation of these SIFT variational color descriptors can be found in [28]. Another set of color descriptors, introduced in the MPEG-7 standard [21], relies on transformations to various color spaces. Color descriptors are often designed to be compact for fast indexing [18], or to maintain a level of photometric invariance [15]. However, all these descriptors operate locally on low-level image features and therefore lack the ability to capture global color information.

**Learned Features for Visual Similarity:** For complex natural images, it is challenging for hand-crafted descriptors to perform well. Deep Convolutional Neural Networks (DCNN) have been successfully applied to image similarity. One type of methods learns similarity metrics for pairs of images using pairwise similarity data [6, 33, 3, 34]. Another approach uses triplet data where a reference image is paired with a positive and a negative example [31, 30]. In either case, pairwise image similarity is labeled by category, attributes or binary classification. When labeling by category, colors are ignored. Binary classes often relate to generic visual similarity rather than specifically to colors. With attribute learning, the visual attributes are expressed in terms of textual descriptions [10, 32, 22, 8, 16, 24, 26] and therefore over-simplify the objectives of visual and color similarity. In this work, we aim to develop better metrics

[1] dataset download link: https://github.com/hamailan/Color-Composition-Similarity

for color similarity that can also provide color features that are beneficial for many Computer Vision applications.

**Datasets for Perceptual Similarity:** In order to train or validate perceptual similarity metrics and descriptors, we need to have datasets that are assessed by people as ground-truth. However, there is no fine-grained rating dataset for perceptual similarity for images in-the-wild. INRIA Holidays [14] is a dataset where similar images are grouped together and dissimilar images are assigned to different groups. Another type is a triplet dataset [30] that provides a coarse level of similarity. Recently published, the BAPPS dataset [34] contains natural images and their generated distortions for learning perceptual patch similarity. In this work, we fill in the missing gap by contributing a perceptual color similarity in-the-wild dataset for which the similarity is measured by participants' ratings on color composition using a fine-grained 5-point scale rating.

## 3. Process for Definition of Perceptual Color Composition Similarity

The complexity of color composition on natural images makes it extremely challenging to write down a set of rules or formulae to define the perceptual color composition similarity. If we need to give a verbal definition, the similarity criteria that we are aiming at are the layout of colors, color distribution, dominant colors, and the overall perceptual appearance of colors in the images. Instead of relying on such a description, our approach is to capture this definition directly via human judgments. Given a pair of reference and test images, participants rate the degree to which the pair is similar with respect to colors exclusively. We face two challenges. The first is selecting images for which ratings for color similarity make sense. Statistically, numbers of pairs that are different in color compositions are much higher than similar pairs. The second is to convey an unambiguous definition of color composition similarity to participants so that they can understand and provide useful and reliable ratings. Unlike other types of annotations, it is not easy to describe the degrees of similarity.

Our solution to the first problem is to build the dataset in two stages. In the first stage, we create a small binary dataset in which similarity is clearly defined: either very similar or completely dissimilar. We start with the least subjective data. It is possible to collect a small set of pairs or groups of similar images by using INRIA Holidays [14] dataset and Pixabay [1] images. From this starting binary dataset, we train a small binary classification network (*binary-net*) to identify similar/dissimilar labels for pairs of images. We then use the *binary-net* to sample more image pairs and have them annotated by participants as similar or dissimilar. The network performance is further improved with the extended set of annotated pairs. We name the improved binary network as *improved-binary-net*. For more details, see Section 3.1.

In the second stage, we use the *improved-binary-net* to select images for the rating dataset. We ask participants to evaluate similarity on a finer-grained 5-point scale where 1 means a pair is completely different and 5 means the images are very similar or almost identical. It is important to choose evaluated images such that the ratings are present for all 5 options. We describe the detailed strategy in Section 3.2.

To solve the second problem, we ask participants to consider several cues that help them consistently compare pairs of images such as presence of dominant colors, distribution of colors, colors of foreground objects and the background, and the overall perceptual appearance of colors in the whole image. We quantify the rating from 1 to 5 as follows: 1 - the pair of images are totally different, 2 - below 50% similar colors, 3 - about 50% similar colors, 4 - above 50% similar colors and 5 - very similar to identical (e.g., Fig. 1). We present many rating examples, conduct an entrance test before participants can start working on the project and embed hidden test questions seamlessly into work items. The test pairs and their expected ratings serve as ground truth to assess whether participants' rating criteria are consistent with the requirements of the task. In the test cases, to allow room for subjectivity, for very similar pairs we set the candidates for correct ratings to $\{4, 5\}$. For pairs that are absolutely different, the correct rating candidates are set to $\{1, 2\}$. For non-extreme similarities, ratings of $\{2, 3, 4\}$ are allowed. Participants must pass the entrance test and maintain their accuracy above 70% throughout the study. Finally, the quality of our rating dataset is evaluated in Section 3.3.

### 3.1. Binary Dataset and Network

We combine the images from INRIA Holidays [14] and Pixabay [1] datasets to create our own dataset. We use an active learning approach to improve the binary network and expand the dataset (Fig. 2). The process starts with an equal number of 3,591 labels each for similar and dissimilar image pairs. This small set of labels are manually annotated by the authors (Fig. 2(a)) and are used to train the initial binary network named *binary-net* to classify similar or dissimilar images in term of color composition (Fig. 2(b)). Due to the limited amount of training data, we design a CNN architecture with few parameters. Instead of using a Siamese model, we stack pairs of RGB images into 6 channel inputs. We augment the images by horizontal flips, small rotations, and swap the two inputs. The output of the network is softmax scores for 2 classes: similar and dissimilar.

In the next step, we generate data for participants' evaluation on new pairs of images for binary classification using the initial binary network *binary-net* (Fig. 2(c)). We select 1,302 reference images that cover a wide variety of objects, textures, and scenes. We use *binary-net* to evaluate the binary similarity between each reference image against a set

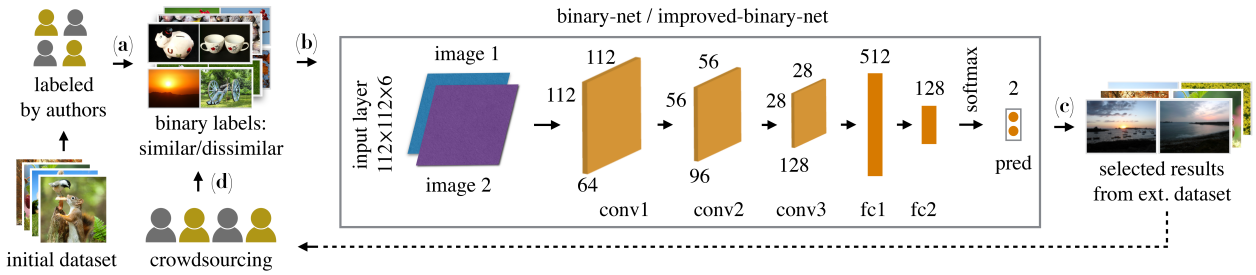



Figure 2: Active learning approach for building color composition similarity binary dataset: it starts with hand-picked similar image pairs (a), on which a classifier is trained (b) to select more similar image pairs (c), which in turn are annotated for similar or dissimilar by crowd-sourcing participants. The process is repeated by using accumulated user annotated data.

of 3,000 images from our large pool dataset. The results from the *binary-net* are then sorted from the most similar to the most different based on their similarity scores. For each reference image, only the first few dozen images are similar and the majority of images are different. Therefore, we select only 24 evaluated images per reference for participants to evaluate similar or dissimilar. These 24 images consist of 1 highly similar image from the initial set of 3,591 labels that are manually selected at the beginning, the first 20 images resulted from the *binary-net* and 3 dissimilar images that are taken randomly at the end of the *binary-net* result list. It yields 31,248 pairs of comparisons in total. Finally, the participants' evaluations are added to the binary dataset (Fig. 2(d)) and fed to re-train the initial *binary-net* to increase its accuracy (Fig. 2(b)). This re-trained network is called *improved-binary-net* and we use it to select images for fined ratings in Section 3.2.

### 3.2. Rating Dataset

In the subsequent crowd-sourcing process we create a fine-grained rating dataset from the binary set. We ask participants to evaluate the similarity for pairs of images using a 5-point Likert-type scale, ranging from absolutely dissimilar (1) to very similar or identical (5). The rating data comprises 1,302 reference images. There are 24 evaluated images for each reference. It is important to choose the evaluated images such that their ratings span the entire 5-point scale. For very similar to identical (rating 5) pairs of images, we choose pairs from the 3,591 manual labeling data. For pairs of images for which the similarity ratings potentially range from 2 to 4, we select pairs from the top results of *improved-binary-net* sorted by similarity scores. Dissimilar pairs of images (rating 1) are accurately chosen from the bottom of the sorted *improved-binary-net* result list.

The important factors that control the quality of the rating dataset are the rating accuracy and consistency among work items of individual participants as well as the consistency among all participants for each work item. To reduce biases and promote the coherence of participants' ratings, for every reference image, we presented to participants a group of evaluated images at a time. We asked the participants to not only rate each pair of images individually but also compare among the group of evaluated images. If an evaluated image $A$ is more similar to the reference image $R$ than an evaluated image $B$ to $R$, then the rating for $A$ should be higher than for $B$ and vice versa. If both images $A$ and $B$ are equally similar to the reference image, then the ratings for both should be the same. This strategy provides an additional context for rating, thus helping participants to adjust their individual ratings to become more consistent.

### 3.3. Quality of the Rating Dataset

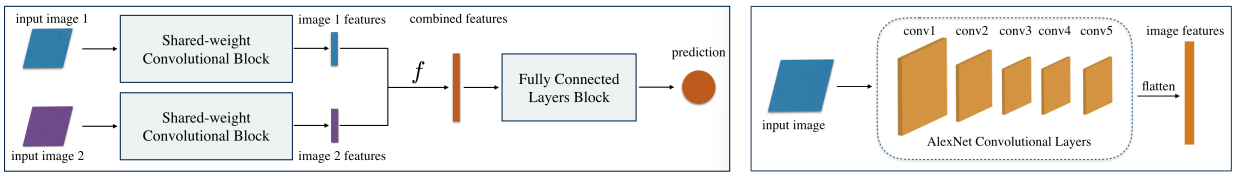
One important aspect of crowd-sourcing experiments is to have a sufficient number of participants working on each question. In highly subjective perceptual comparison tasks, we need a larger number of user judgments per item compared to less subjective tasks such as object labeling. Therefore, we conducted a preliminary experiment on a small part of the dataset (559 pairs) using 40 ratings per pair. We studied how well a smaller number of ratings can reproduce the mean of 40 ratings. We found that the mean opinion derived from 20 ratings suffices to obtain a 0.994 Pearson linear correlation with the mean for 40 ratings, with an MAE of 0.033 on a scale of [1,5]. Thus, we chose 20 ratings per pair.

To evaluate the quality and reliability of the dataset, we use the Intra-class Correlation Coefficient (ICC). The one-way ICC on our dataset is 0.69. This suggests a high agreement in the context of crowd-sourcing rating experiments, where values between 0.3 and 0.5 have been previously reported on several rating datasets [25, 13].

## 4. Computational Model of Perceptual Color Composition Similarity

With the rating dataset, we train Convolutional Neural Networks (CNN) to evaluate the perceptual color composition similarity. These networks can be used as similarity metrics and color feature extractors.

Different from binary networks, rating networks allow us to rank image similarity. We train two types of rating net-




(a) Siamese network for predicting ratings.

(b) Shared-weight Convolutional Block.

Figure 3: Siamese architecture (a) using Convolutional Neural Network (b) for training our color similarity metrics.

works: **COLSIM_RATE** to predict the participants rating distribution and **COLSIM_MOS** to predict the participants Mean Opinion Score (MOS). Both networks use the same architecture as in Fig. 3(a). The only difference is in the prediction layer, where we have a single output for MOS and five outputs for rating distributions. The overall architecture is a Siamese network that has two input images. Each input is fed into a shared-weight Convolutional Block that contains a series of convolutional layers to extract features. The features from the two input images are combined by a function $f$ defined in Eq. 1. Finally, a neural network with a few fully connected layers performs the predictions based on the combined features.

**Shared-weight Convolutional Block**: we use 5 convolutional layers that are similar to the Caffe implementation of AlexNet with Batch Normalization on the first 2 layers. The responses of the last convolutional layer are flattened to form a feature vector $v$ (Fig. 3(b)). We also train a compact network that contains only 3 convolutional layers on images of size 112×112 pixels. The compact network is smaller and faster, but there is a slight drop in performance (see Section 5).

**Image Features Combination**: to combine features of image 1 ($v_1$) and features of image 2 ($v_2$), we use 3 different metrics: absolute difference, squared difference and Hadamard product as follows:

$$f : (v_1, v_2) \rightarrow C(|v_1 - v_2|, |v_1 - v_2|^2, v_1 \cdot v_2) \quad (1)$$

where '$\cdot$' denotes the element-wise multiplication, and $C$ is the concatenation operator. The combined features resulting from Eq. 1 are used as the input to the Fully Connected Layer (FCL) block (Fig. 3(a)).

**Fully Connected Layer (FCL) Block**: is comprised of two fully connected layers of size 512 and 128. We use dropout 0.5 for the first FCL and 0.2 for the second FCL. ReLU activation is used throughout the whole network.

**Rating Distribution prediction**: participants' ratings are distributed over the 5-point scale. Given a pair of images, we want to predict the participants' rating distribution. We use different metrics for computing the distribution losses, including Mean Absolute Error (MAE), Kullback-Leibler (KL) divergence and Huber loss. From numerical results, KL divergence consistently performs the best. Thus, we use KL divergence in all of our rating networks.

**Mean Opinion Score (MOS) prediction**: From the participants' rating distributions, we can compute MOS values that are useful for image ranking. The MOS is computed as $MOS = \sum_{i=1}^{n} i \cdot P(i)$ where $P$ is the normalized rating distribution and $n = 5$ for a 5-point rating scale. We also train networks that predict MOS (**COLSIM_MOS**) using Mean Squared Error (MSE) loss. Our experiments show that MOS derived from predicted rating distributions has lower errors compared to the results of networks that are trained directly on MOS data.

## 5. Color Descriptors Evaluations

In order to evaluate and compare the performances of different descriptors and networks on perceptual color similarity measurements, we split the dataset into an 80% training set (24,840 pairs) and a 20% test set (6,210 pairs). There are no common reference images in the two sets. All the algorithms are trained and validated on the training set and tested on the test set. The results reported in Table 1 are the SROCC on the test set, which measures the Spearman Rank Order Correlation between the predicted results and participants' ratings. We choose SROCC over other metrics such as MAE or MSE because it accounts for the changes in scale and non-linearity of the measurements coming from different descriptors and methods.

We divide color descriptor methods into three groups: histogram-based, SIFT-based and MPEG7. We use L1 and L2 for all descriptors in these three groups to measure the similarity between pairs of images in the test set, rank them, and compute the SROCC. We also train the descriptors using CNNs, and neural networks for SIFT and MPEG7 descriptors, respectively. To extract SIFT descriptors, we densely sample the images and compute SIFT color features at each sampled point. The resulting data is enough to train a CNN that has a similar architecture to our **COLSIM_RATE** network (in Section 4). MPEG7 descriptors, on the other hand, are very compact. Their sizes are 192 for Color Layout Descriptor (CLD), and 256 for Color Structure Descriptor (CSD) and Scalable Color Descriptor (SCD). Thus, we train a small neural network that has two fully connected layers with one prediction layer. The features produced by descriptors for pairs of images are com-

| Descriptor | Spearman correlation $\rho$ | | |
|---|---|---|---|
| | L1 | L2 | Trained MOS/Rating |
| nrghistogram | 0.503 | 0.546 | - |
| opponent histogram | 0.604 | 0.498 | - |
| hue histogram | 0.631 | 0.535 | - |
| lab histogram | -0.260 | -0.336 | - |
| rgsift | 0.259 | 0.277 | - / 0.754 |
| hsvsift | 0.327 | 0.277 | - / 0.757 |
| csift | 0.351 | 0.318 | - / 0.687 |
| opponentsift | 0.604 | 0.498 | - / 0.636 |
| huesift | 0.631 | 0.535 | - / 0.862 |

| Descriptor / Network | Spearman correlation $\rho$ | | |
|---|---|---|---|
| | L1 | L2 | Trained MOS/Ratings |
| CLD | 0.290 | 0.562 | - / 0.715 |
| CSD | 0.653 | 0.692 | - / 0.737 |
| SCD | 0.692 | 0.646 | - / 0.720 |
| VGG19 + L2 | - | - | 0.467 / - |
| VGG19 Transfer | - | - | 0.780 / 0.812 |
| VGG19 Fine-tune | - | - | 0.832 / 0.863 |
| Compact (ours) | - | - | **0.860 / 0.869** |
| COLSIM (ours) | - | - | **0.902 / 0.913** |

Table 1: Evaluation of color descriptors and learning methods on color composition similarity. Predictions are based on L1 and L2 norms, or trained on 'MOS' and distribution of 'Ratings'. The Spearman $\rho$ between the predictions and the MOS computed from user ratings is reported. Performance is highest when training on distributions of ratings.

bined using the function $f$ as described in Eq. 1.

The numerical results show that descriptors, even though designed for color similarity, do not correlate well with human evaluations. The maximum SROCC is 0.692 obtained with CSD and SCD descriptors. Training a CNN or Neural Network on the descriptors can improve the results up to a maximum of 0.860 SROCC in the case of **huesift**. Nevertheless, it takes an additional step to first compute descriptors before training them to get decent results.

A straightforward approach is to fine-tune a network or train one from scratch on our dataset. We do transfer learning from pre-trained features and then fine-tuning using the VGG19 network [27]. As VGG19 is trained for object categorization, it cannot perform well out of the box on color similarity. The SROCC result for L2 distance on fc7 features of the VGG19 is 0.467. It shows that content and color are not highly correlated. The SROCC result of VGG19 transfer learning is 0.812 and improves to 0.863 with fine-tuning. Even though the results are satisfying, we observe that the features in VGG19 favor classification and hence still affect the performance of color similarity measurement. Thus, we train a rating network **COLSIM_RATE** described in Section 4 from scratch. The SROCC of **COLSIM_RATE** is 0.913, the best of all methods. We also train a Compact network that contains 3 convolutional layers, 2 fully connected layers and 1 prediction layer on images of size 112×112 pixels. Even though the performance is lower at 0.869 SROCC, the network has fewer parameters while having comparable performance to the fine-tuned VGG19. The MOS prediction network **COLSIM_MOS** has an SROCC of 0.902, which is slightly lower than **COLSIM_RATE**.

Regarding errors, we plot the cumulative distribution function (CDF) of the MAE between the participants' distribution of ratings and our **COLSIM_RATE** network's predictions in Fig. 4(a). The MAE is below 0.1 for 70% of the test data and only increases substantially in the last 5%.

## 6. Fine-grained Image Similarity

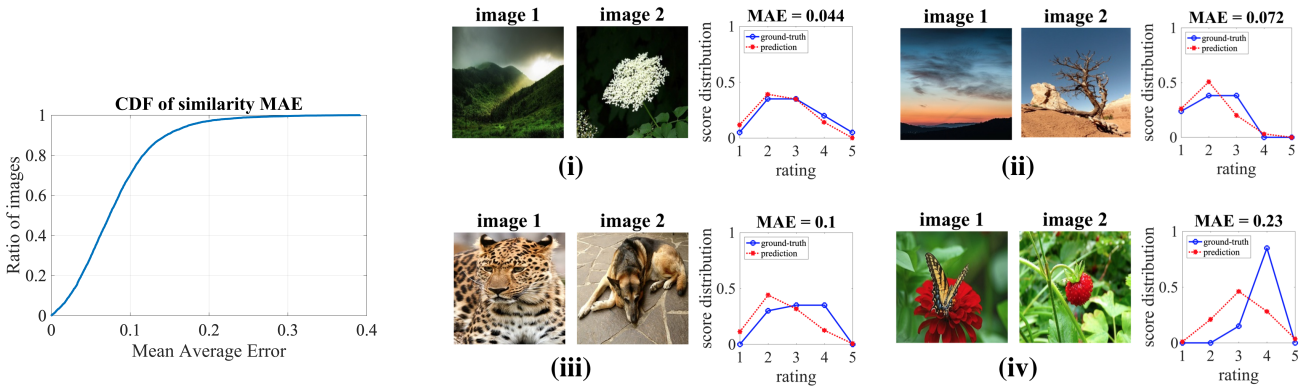
Fine-grained image similarity measures not only the content difference among image classes but also the visual difference within a class. Image retrieval by class or categorical features does not consider colors as a part of the ranking procedure. For instance, when searching for an image of a black poodle, retrieval prioritizes semantic information and returns poodles with various colors. This is not always desirable. We show that by using our visual color similarity metric, the relevance of the ranking results is improved.

### 6.1. Related work

Existing methods relate visual similarity to fine-grained classification or visual attribute similarity. These two main approaches are only beginning to tackle the complex nature of perceptual comparisons as part of visual search. Visual similarity is contextual because of the subjective judgments and its use-case. For instance, a query for an image depicting a leopard pup at the zoo could be intended to retrieve images of leopards (pure class), young leopards (object attribute and class), or yellow animals (color and class).

The first type of methods learn features for general visual similarity [6, 30, 23, 3, 31], starting from category labels, textual descriptions, or triplet data. The second type of approaches separate aspects of visual similarity, by learning from human-nameable visual attributes or discovering new ones for image retrieval [10, 32, 22, 8, 16, 24, 26, 31]. Attribute learning complements category-level recognition by learning the degree to which one or more attributes are present in an image. Attributes are very specific and combining them is challenging [26] due to their interactions.

We propose to separate visual similarity into multiple factors that can be individually studied. In this work, we focus on the color composition factor. This is not a per-image attribute as we cannot quantify the amount of color composition in an image, nor can we say that an image has more




(a) CDF of similarity ratings MAE.

(b) Examples of different error levels (MAE). Blue graph: ground-truth, red graph: prediction

Figure 4: The MAE between participants' rating distributions and the **COLSIM_RATE** network's predictions on the test set. For most images the MAE is small, e.g., (i) and (ii) whereas only 3% have an MAE $> 0.2$, e.g., (iv).

or less color composition than another. However, it allows us to better specify the context in visual search. We use the correlation between pairs of content features and color similarity to improve fine-grained visual similarity prediction.

### 6.2. Features and training model

Our hypothesis for improving fine-grained similarity is that the combination of category and color features helps to better predict the similarity of image pairs compared to the individual features alone. The similarity in the categorical feature space is computed as the correlation between two feature vectors of pairs of images. The color similarity features are extracted from our color composition similarity metric or L2 distance for existing hand-crafted color descriptors. The detailed formulations for the content correlation and color similarity are explained below. Our hypothesis is verified by numerical results in Table 2.

Wang *et al*. [30] have introduced a fine-grained similarity database which contains 5,033 ranked triplets. A triplet comprises a query $Q$, and two compared images $A$ and $B$. If the visual similarity $sim(Q, A) > sim(Q, B)$ which means A is more similar to Q than B, then the correct ordering of the triplet is $(Q, A, B)$.

Using this dataset, we study different similarity measures on category and color features individually and in combination. We use the L2 distance to measure the visual similarity between pairs of images. For content features, we evaluate L2 on the fc8 layer of AlexNet and the Global Average Pooling (GAP) layer of ResNet50. For color features, we evaluate L2 for all color SIFT descriptors, MPEG7 descriptors and COLSIM features extracted from our model. The L2 distance on individual types of features does not yield good results (Table 2). Therefore, we train a binary classifier (SVM, RBF kernel) on the triplet data using combinations of features. In general, the input features to the SVM are a pair of similarities $(sim(Q, A), sim(Q, B))$ for a correct triplet $(Q, A, B)$. Wrongly ranked triplets are created from the correct ones, by reversing the relationships $(sim(Q, B), sim(Q, A))$.

The features that are used when training the SVM are: the direct color similarity produced by the **COLSIM** network $S_{COLSIM}(X, Y)$, and the Pearson Linear Correlation Coefficient (PLCC) between *GAP* content features [2] $F^{GAP}$ extracted from a pre-trained **ResNet50** network: $S_{GAP}(X, Y) = PLCC(F^{GAP}(X), F^{GAP}(Y))$ where $PLCC(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i-\overline{x}}{\sigma_x})(\frac{y_i-\overline{y}}{\sigma_y})$ where $n$ is the number of dimensions of the features $x, y$. Therefore, the input features for "SVM ResNet GAP correlation" is $[S_{GAP}(Q, A), S_{GAP}(Q, B)]$ which contains only content features for a triplet $(Q, A, B)$. The input features for "SVM ResNet + **COLSIM**" are the combination of content similarity and color similarity, and defined as $[S_{COLSIM}(Q, A), S_{COLSIM}(Q, B), S_{GAP}(Q, A), S_{GAP}(Q, B)]$ for a triplet $(Q, A, B)$.

### 6.3. Results analysis and discussion

The DCNN methods in [30] have been evaluated on a validation dataset of 14,000 triplets. However, the authors [30] make available only a subset of 5,033 triplets. We evaluate our models on this subset by using 20 repetitions of random 80%/20% train/validation splits. The optimal hyper-parameters for each split are estimated by 5-fold cross-validation. Our proposed model using the combined category and color feature similarities performs best. We do not have access to the other methods to directly compare their performances on the "Subset 5k" database. Thus, we use a shared baseline model for comparison: the L2 distance between fc8 features from AlexNet [17], named "ConvNet AlexNet fc8". This common baseline performs much better on the 14K dataset than on the subset 5K (Table. 2). Therefore, we expect equivalent methods will perform better when tested on the 14K compared to the 5K subset.

[2] the terms category and content features are used interchangeably

| Model | Validation 14k (not available) | Subset 5k ($\subset$ 14k) |
|---|---|---|
| L2 on ConvNet AlexNet fc8 | 82.8% (baseline) | 73.7% (baseline) |
| Single-scale Ranking | 84.6% | - |
| OASIS on Single-scale Ranking | 82.5% | - |
| Single-Scale & Visual Feature | 84.1% | - |
| DeepRanking | **85.7% (+3.5%)** | - |
| L2 on *sift descriptors | - | 62.9% - 65.4% |
| L2 on MPEG descriptors | - | 62.3% - 65.1% |
| L2 on COLSIM features | - | 69.1% |
| L2 on ResNet GAP | - | 79.1% |
| SVM on COLSIM correlation | - | 73.7% |
| SVM on ResNet GAP correlation | - | 84.3% |
| SVM on ResNet + **COLSIM** | - | **86.2% (+12.5%)** |

| Color Descriptor | Combined features | Color features |
|---|---|---|
| csift | 84.5% | 50.7% |
| rgsift | 84.6% | 61.3% |
| oppsift | 84.8% | 64.5% |
| hsvsift | 85.1% | 62.7% |
| huesift | 85.3% | 65.4% |
| CSD | 85.5% | 68.9% |
| SCD | 85.5% | 62.8% |
| **COLSIM** (ours) | **86.2%** | **73.7%** |

SVM results on the 5k subset when training with (combined) and without content features. Except COLSIM, we use L2 for the rest of color descriptors to compute SVM features.

Table 2: Evaluation on the DeepRanking triplet dataset. Results for the 'Validation 14k' column are reproduced from [30].

**State-of-the-art performance:** the accuracy of our method is 86.2% compared to 85.7% for the best DeepRanking [30] approach. However, our method shows a substantially higher improvement of 12.5% relative to the shared baseline, compared to the improvement of 3.5% for DeepRanking. As the performance of the baseline method on 'Subset 5K' (73.7%) is much lower than on 'Validation 14K' (82.8%), the relative % improvement suggests a much better overall performance for our method.

**Feature combination vs individual features:** even though the SVM training on ResNet GAP correlation and COLORSIM scores achieves the best results, we also test the model on different hand-crafted descriptors. The results, on the right of Table 2, show that: (i) COLSIM outperforms hand-crafted descriptors; (ii) the combination of content feature correlation and color similarity yields better accuracy compared to using L2 on descriptors or ResNet GAP alone (on the left of Table 2).

**Feature correlation vs L2 distance:** using content or color descriptors alone, we find that training an SVM on the PLCC of the features results in a better accuracy than L2 distance on the respective features. For instance, the accuracy for SVM on ResNet GAP correlation is 84.3% compared to 79.1% for L2 on ResNet GAP features.

**Features vs end-to-end training:** while DeepRanking [30] used 14 million google search images during training, and a large set of triplets ($\approx$ 50k), our method relies on a much smaller set of 5,033 triplets and our own database of 30k image pairs. The improved performance of our approach, using combined category and color features, shows that embedding domain knowledge in our model achieves both excellent performance and efficient training. Training on the proposed low-dimensional pairwise features is much faster than an alternative end-to-end triplet network.

## 7. Conclusion

We hypothesize that visual similarity can be better studied by isolating its multitude of aspects and modeling them individually. This approach requires the means to isolate, model, and combine multiple aspects. We isolate the aspect of color composition similarity, define an efficient data collection and annotation strategy including an active learning approach for this subjective measurement task. This process leads to the first large-scale dataset for measuring color composition similarity for images in-the-wild. Our dataset has enabled us to train accurate DCNN models for perceptual color similarity and benchmark the performance of existing color descriptors. The numerical results show that few existing descriptors are informative for global color similarity, except for deep features that are trained on our dataset. We create an improved model for visual ranking similarity, by introducing a novel way to combine non-homogeneous representations such as color similarity and category features. These multi-aspect, low-dimensional features have proven to be extremely effective in training visual ranking models, surpassing the existing state of the art 'DeepRank that was trained on substantially more data. Overall, the results prove that our proposed approach better predicts visual similarity. We expect that future works will improve visual similarity models by isolating and studying other aspects such as texture, style, etc.


## Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the research training group GRK-1564 "Imaging New Modalities", and as part of Project-ID 251654672 - TRR 161 (Project A05).

## References


[1] Pixabay dataset. https://pixabay.com/. Accessed: 2019-03-22. 3

[2] A. E. Abdel-Hakim and A. A. Farag. Csift: A sift descriptor with color invariant characteristics. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1978–1983, 2006. 1, 2

[3] S. Bell and K. Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics*, 34(4):98:1–98:10, July 2015. 1, 2, 6

[4] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008. 1, 2

[5] G. J. Burghouts and J.-M. Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48–62, Jan. 2009. 1, 2

[6] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *The Journal of Machine Learning Research*, 11:1109–1135, Mar. 2010. 1, 2, 6

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1

[8] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and fisher vectors for efficient image retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 745–752, June 2011. 1, 2, 6

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1

[10] V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, Dec. 2007. 1, 2, 6

[11] E. D. G. Sharma, W. Wu. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research and Application*, 30(1), Feb. 2005. 1

[12] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, Dec. 2001. 1, 2

[13] V. Hosu, H. Lin, and D. Saupe. Expertise screening in crowdsourcing image quality. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2018. 4

[14] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision*, pages 304–317, 2008. 2, 3

[15] R. Khan, J. van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat. Discriminative color descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2866–2873, June 2013. 1, 2

[16] A. Kovashka and K. Grauman. Attribute adaptation for personalized image search. In *IEEE International Conference on Computer Vision*, pages 3432–3439, Dec. 2013. 1, 2, 6

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012. 7

[18] V. Ljubovic and H. Supic. A compact color descriptor for image retrieval. In *International Conference on Information, Communication and Automation Technologies*, pages 1–5, Oct. 2013. 1, 2

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov. 2004. 2

[20] M. R. Luo, G. Cui, and B. Rigg. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, 26(5), 2001. 1

[21] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715, June 2001. 1, 2

[22] D. Parikh and K. Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, pages 503–510, 2011. 1, 2, 6

[23] N. Pourian and B. S. Manjunath. Pixnet: A localized feature representation for classification and visual search. *IEEE Transactions on Multimedia*, 17(5):616–625, May 2015. 1, 6

[24] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3310–3317, June 2013. 1, 2, 6

[25] E. Siahaan, A. Hanjalic, and J. Redi. A reliable methodology to collect ground truth data of image aesthetic appeal. *IEEE Transactions on Multimedia*, 18(7):1338–1350, July 2016. 4

[26] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 801–808, June 2011. 1, 2, 6

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[28] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, Sept. 2010. 1, 2

[29] J. van de Weijer, T. Gevers, and A. D. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):150–156, Jan. 2006. 1, 2

[30] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014. 1, 2, 3, 6, 7, 8

[31] X. Wang, K. M. Kitani, and M. Hebert. Contextual visual similarity. *arXiv:1612.02534*, 2017. 1, 2, 6

[32] X. Yang, T. Zhang, C. Xu, S. Yan, M. S. Hossain, and A. Ghoneim. Deep relative attributes. *IEEE Transactions on Multimedia*, 18(9):1832–1842, Sept. 2016. 1, 2, 6

[33] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199, Sept. 2014. 1, 2

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 2, 3