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

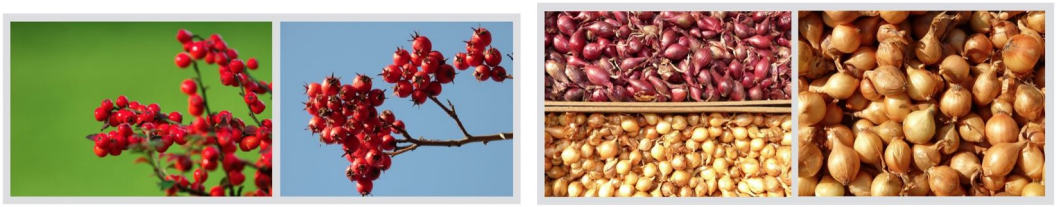
Supplementary Material

In this supplementary material, we provide extra information as follows:

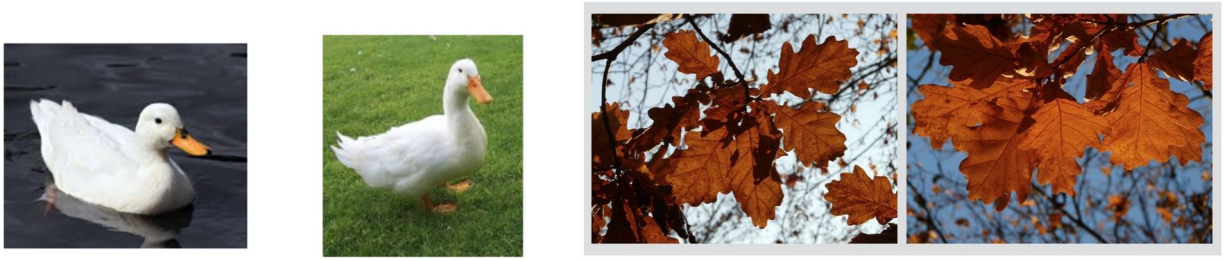
- Examples for color composition similarity rating scores (Section 1.
- Extra information on our dataset such as different strategies for crowd-source experiments and statistics analysis on the crowd-source results (Section 2).
- Image embedding visualizations based on the Mean Opinion Score (MOS) predicted by our color similarity metric as well as based on the L2 distance between the color similarity features derived from our network model (Section 3).
- Ranking examples for image retrieval by content similarity versus color similarity. Content features are extracted from the ‘fc7‘ layer of the VGG19 object classification network (Section 4).

1. Examples on Color Composition Similarity Ratings

(a) Score 1: the images in each pair are totally different



(b) Score 2: below 50% similar colors



(c) Score 3: about 50% similar colors

(d) Score 4: above 50% similar colors

(e) Score 5: very similar to identical

Figure 1: Examples for each similarity score (two pairs), shown in the introduction of the crowd-sourcing experiment.

2. Color Composition Similarity Dataset

2.1. Rating strategies

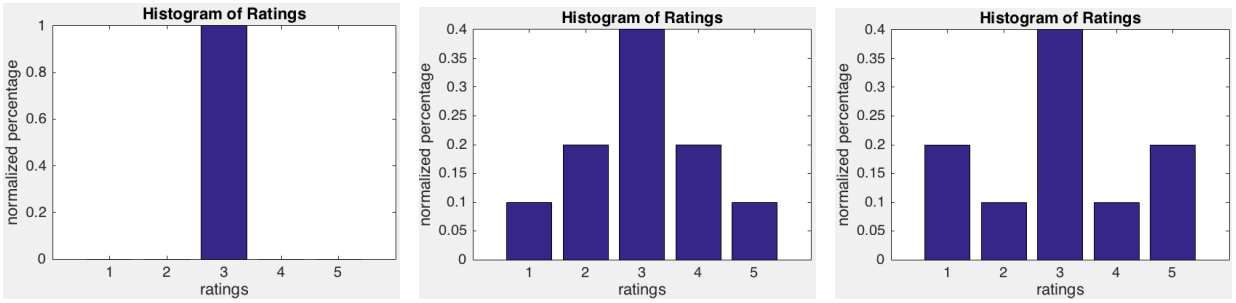
We did experiments for rating color similarity between pairs of images using different strategies: (a) pair comparison, (b) triplet comparison and (c) group rating. To study the effect of ratings of paired comparisons without and with context i.e. image group presentation, we conducted a small experiment. In without context comparison (Fig. 3(a)), we presented users with individual pairs of images (randomly selected) and asked them to rate the color composition similarity for the displayed pair. This is a normal pair comparison. In the with context comparison (Fig. 3(b)), we presented users with three images (triplet comparison): a reference image R, an image A and an image B. We asked users to rate the color composition similarity of image A to the reference R, and of image B to the reference R. Moreover, we also asked users to compare image A and B making sure the ratings are three-way consistent such that they follow the requirement in Eq. 1.

$$\begin{aligned} S(R,A) > S(R,B) &\Leftrightarrow Q(R,A) > Q(R,B) \\ S(R,A) = S(R,B) &\Leftrightarrow Q(R,A) = Q(R,B) \end{aligned} \tag{1}$$

where S is the perceptual similarity judgment, R is the reference image, A and B are evaluated images and Q is the rating.

If an evaluated image A is more similar to the reference image R than an image B is 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 R then their ratings should be the same. This strategy provides additional context for assigning ratings, and thus helps participants to adjust their individual ratings to become more consistent.

We analyze the user agreement for both triplet comparison and pair comparison. Given a reference image R and an evaluated image A, the rating $\tilde{Q}(R,A)$ is a rating distribution from 1 to 5 for the color composition similarity between the reference image R and the evaluated image A from all the users. Let $H(\tilde{Q})$ be the normalized histogram of $\tilde{Q}$ (bins sum to 1), and $H^k(\tilde{Q})$ be the normalized histogram of $\tilde{Q}$ at bin k, $k \in [1..5]$, and $p(H)$ is the mean of $H(\tilde{Q})$. $p(H)$ is also known as Mean Opinion Score (MOS). Ideally, all users agree on one rating such that $p(H) \in [1..5]$ and $H^{p(H)}(\tilde{Q}) = 1$. However, users seldom agree on their ratings. An illustration of users' rating distribution is in Fig. 2.



(a) MOS=3, $\Phi(H) = 1$ (b) MOS=3, $\Phi(H) = 0.8$ (c) MOS=3, $\Phi(H) = 0.6$

Figure 2: Example of rating distributions. In all three examples, the Mean Opinion Score are the same (MOS = 3). $\Phi(H)$ is computed as $\sum_a^b H(\tilde{Q})$ where $a = 2$ and $b = 4$ which are one scale below and one scale above MOS value. In (a): 100% users agree on the rating 3 and therefore, $\Phi(H) = 1$. In (b): $\Phi(H) = 0.8$, users' ratings spread out such that 80% of users vote around the MOS. The worst case is (c) where $\Phi(H) = 0.6$, only 60% of users rate around MOS.

To measure the users' agreement $\tilde{Q}(R,A)$, we sum up the normalized histogram $H(\tilde{Q})$ around the MOS $p(H)$. The result is $\Phi(H) = \sum_a^b H(\tilde{Q})$ where $p(H) \in [a,b]$, and a and b are the floor and ceiling values of the MOS. When the MOS is an integer, a and b are chosen to be one rating below and one above the MOS. We then compute the mean and standard deviation of $\Phi(H)$ for all pairs of rating images.

$\Phi(H)$ measures agreement, ranging from a lowest value of 0 to the highest value of 1. In Table 1 we show that both the paired and the triplet comparison have a very high average agreement. However, the triplet comparison also has a small Std($\Phi(H)$) meaning that the user agreement is roughly the same high value for all pairs. An experimental methodology that gives consistent results in general e.g. a high minimum agreement for any pair, is clearly better. This confirms that the three-way comparative judgments reduce the overall subjectivity of user ratings.

	$\mu(\Phi(H))$	$\sigma(\Phi(H))$
Pairwise rating	0.983	0.204
Triplet rating	0.952	0.083

Table 1: Evaluation of users' agreement for pairwise and triplet ratings

Based on the numerical results, we conclude that pairwise rating with relative comparisons between a group of images increases the users' agreement. Thus, in the rating experiment, we extend the triplet rating to group rating where each reference image has 24 evaluated images. The 24 evaluated images are divided into 3 groups of 8 images. Users are then asked to compare each evaluated image to the reference image with, making sure to consider the consistency of relationships within the group (Fig. 3(c)).

Our experiments result in a high rating consistency for our dataset: the ICC reaches 0.69. This is very high for crowdsourcing studies, relative to results obtained in other similar rating experiments [1, 2].

How similar are the colors of the two images?

1. Very different	2. Substantially different	3. Fairly similar	4. Substantially similar	5. Very similar
○	○	○	○	○

(a) User's rating on pair comparison.

Image A | Reference Image | Image B

Compare the color between Reference and A: (required)
- ○ 1. Not similar at all
- ○ 2. Less than 50% of the colors are similar
- ○ 3. About 50% of the colors are similar
- ○ 4. More than 50% of the colors are similar
- ○ 5. Very similar

Compare the color between Reference and B: (required)
- ○ 1. Not similar at all
- ○ 2. Less than 50% of the colors are similar
- ○ 3. About 50% of the colors are similar
- ○ 4. More than 50% of the colors are similar
- ○ 5. Very similar

(b) User's rating on triplet comparison.

Reference image

How similar is the color composition of the following images to the reference above?

image A

Image A vs Reference:
- ○ 1. Not similar at all
- ○ 2. Below 50% similar colors
- ○ 3. About 50% similar colors
- ○ 4. Over 50% similar colors
- ○ 5. Very similar

image B

Image B vs Reference:
- ○ 1. Not similar at all
- ○ 2. Below 50% similar colors
- ○ 3. About 50% similar colors
- ○ 4. Over 50% similar colors
- ○ 5. Very similar

image C

Image C vs Reference:
- ○ 1. Not similar at all
- ○ 2. Below 50% similar colors
- ○ 3. About 50% similar colors
- ○ 4. Over 50% similar colors
- ○ 5. Very similar

image D

Image D vs Reference:
- ○ 1. Not similar at all
- ○ 2. Below 50% similar colors
- ○ 3. About 50% similar colors
- ○ 4. Over 50% similar colors
- ○ 5. Very similar

(c) User's rating on group comparison.

Figure 3: Different rating strategies: pairwise, triplet and group.

2.2. Number of rating per pair

We first conducted a preliminary experiment on a small part of the dataset (559 pairs) for which we collected 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 scores (MOS) derived from 20 ratings is sufficient to obtain a 0.994 Pearson linear correlation coefficient (PLCC) with the MOS for 40 ratings, with an MAE of 0.033 on a scale of $[1, 5]$. Thus, we chose 20 ratings per pair.

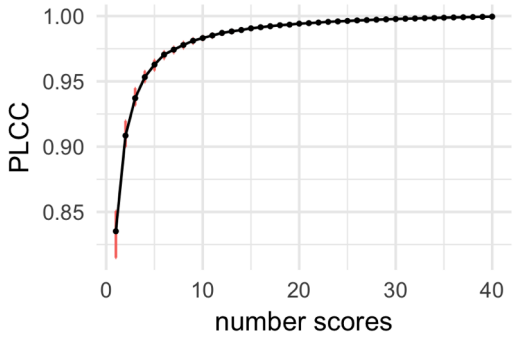


Figure 4: Correlation (PLCC) between the MOS from 'number scores' and the MOS from all 40 scores. The random repeated sub-sampling leads to confidence intervals that are displayed in red.

2.3. Rating Statistic

We compute the histogram of Mean Opinion Scores (MOS) on the entire color similarity dataset. If we were to randomly sample pairs of images, there would be a very low chance that two images were highly similar i.e. a score of 4 and above. In our dataset, we have about 2,500 pairs (6.4% of the dataset) that have ratings of 4 and above. We also have a fair amount of pairs with MOS ratings distributed in the middle of the rating scale (See Fig. 5). This distribution was possible only due to our pair selection procedure employed during the active learning.

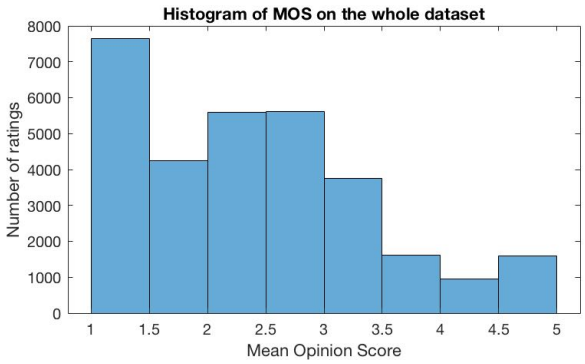


Figure 5: Histogram of Mean Opinion Scores for the entire dataset.

2.4. Binary rating and its contribution

In the binary rating experiment, we asked the participants to rate *Yes* if a pair of images are similar and *No* if they are not. For individual participant, it's a binary rating decision. However, with many participants worked in the same set of data, the experiment resulted the percentage of participants answered *Yes* and *No* for each pair of images. After examining the rating results carefully, we find that these percentages can be interpreted as the confidence scores for the similarity measurement.

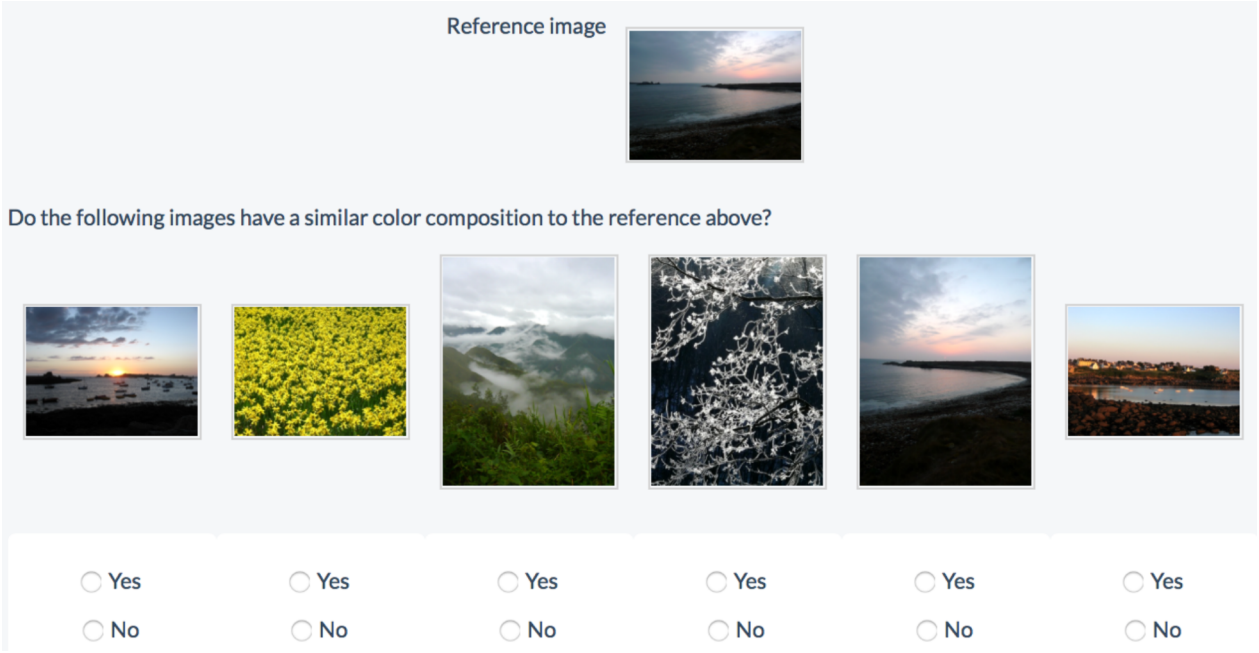


Figure 6: User voting for binary (similar / dissimilar) decision in a small group of 6 evaluated images.

We publish the results of participants' binary evaluation for color composition. One can use our binary dataset to form triplets that are suitable for many methods trained using the triplet loss. One example of a triplet dataset and a method of using triplet data can be found here [3].

3. Color feature based embedding visualization

In this section, we use t-SNE embedding to visualize the Mean Opinion Score (MOS) prediction from our color similarity network (AlexBN). Fig. 7 and Fig. 8 show that MOS predictions reflect the color distances very well, the t-SNE embedding revealing image clusters with similar color compositions. For clear visualizations, we only render images that do not overlap with each other in all the t-SNE embeddings.

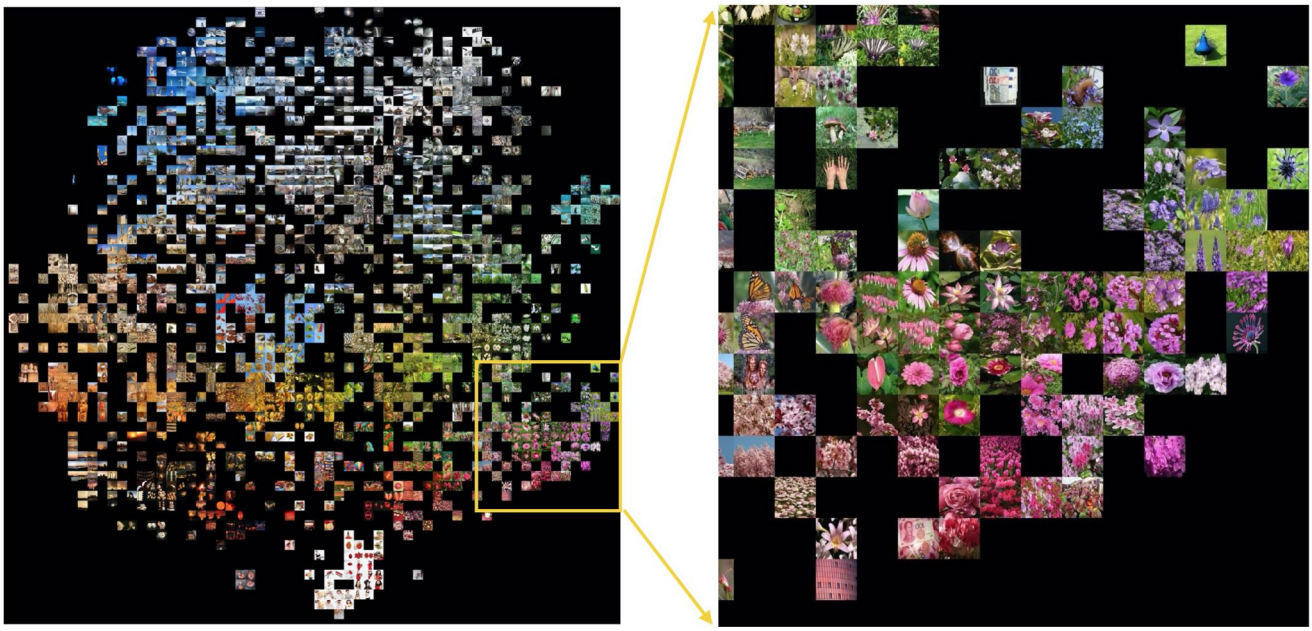


Figure 7: t-SNE embedding on MOS prediction from AlexBN color similarity network for 3K images. The 3K images are mixed selections of Pixabay images and INRIA Holiday dataset. The left image is the overall embedding and the right image is a zoom in of a small region from the left one.

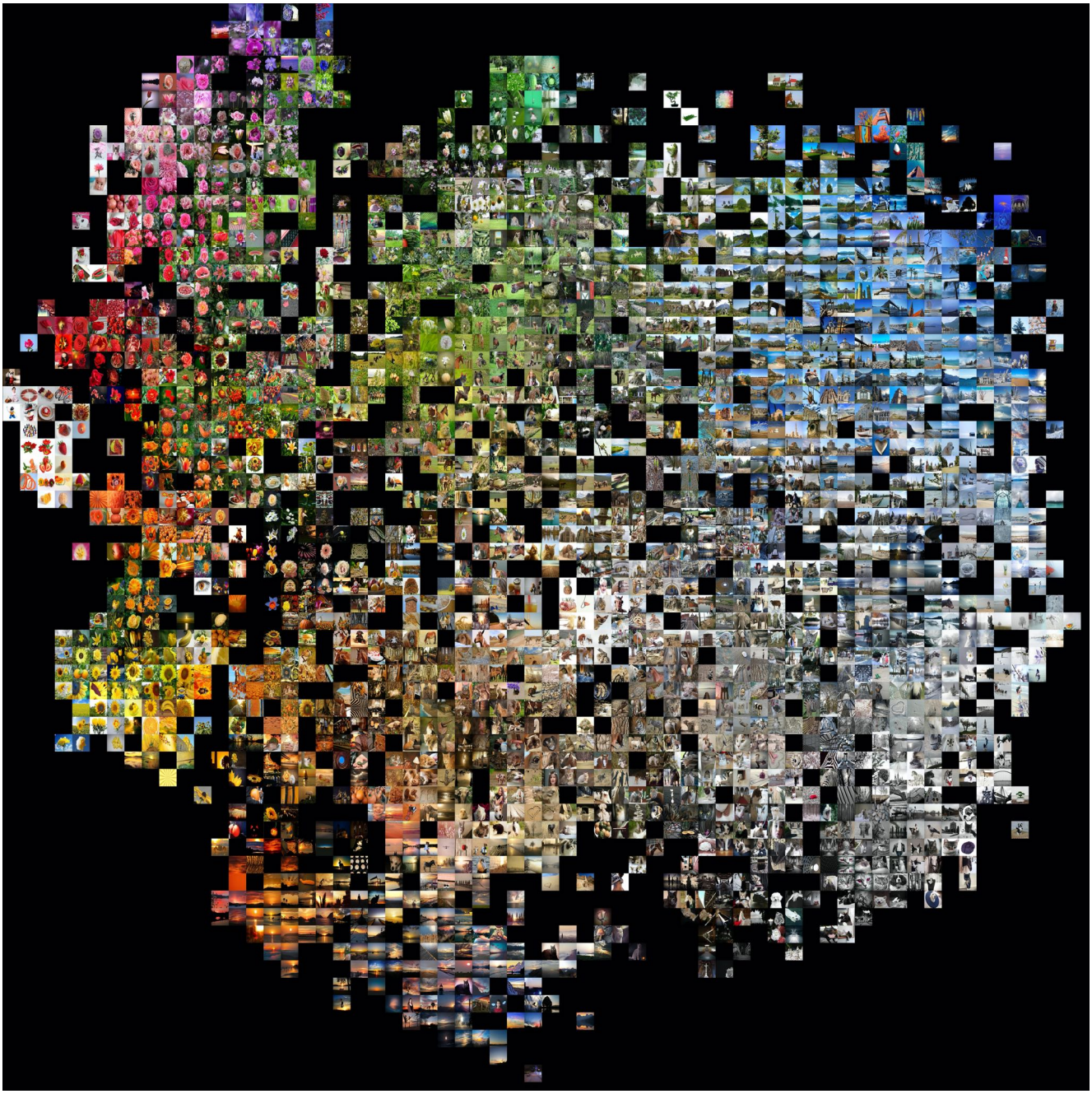


Figure 8: t-SNE embedding based on MOS prediction from AlexBN color similarity network for 5K images from Pixabay. The result shows that images are clustered very well based on colors. For a clear visualization, only images that do not overlap with each other are rendered.

We visualize the color composition similarity implied by the L2 distances between color features from the AlexBN color-sim network in Fig. 9. The color features are extracted from the last convolutional layer of our AlexBN color similarity network. The t-SNE visualization produces a different global arrangement, however the local clusters are similar to those derived from the final prediction of the AlexBN network in Fig. 8. Except for a minority of cases the clusters are visually just as correct.

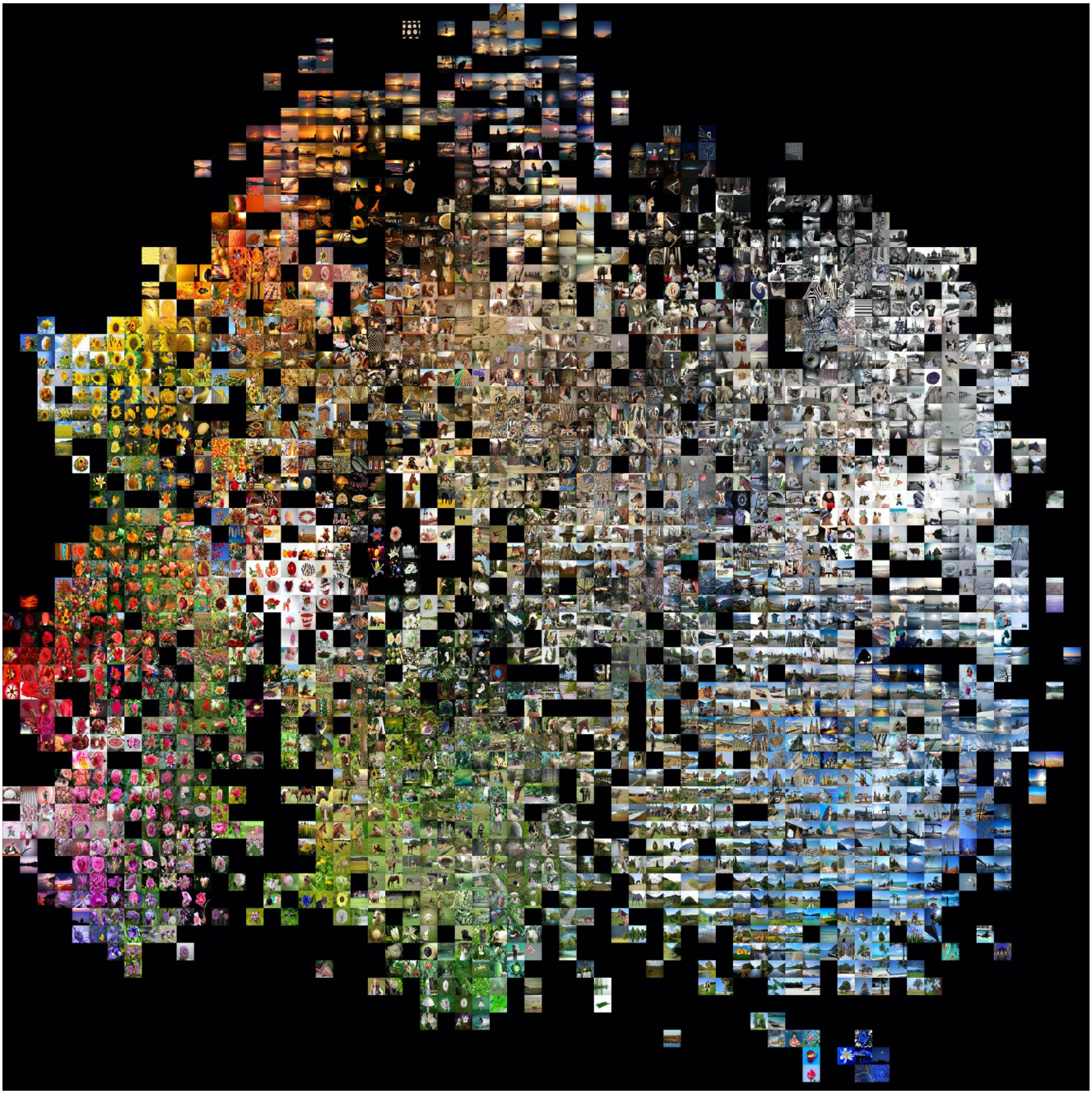


Figure 9: t-SNE embedding based on color features of 5K images from Pixabay. The color features are extracted from the last convolutional layer of our AlexBN color similarity network. For a clear visualization, only images that do not overlap with each other are rendered.

We visualize the L2 distances of fc7 content features from VGG19 in Fig. 10. As expected, we notice that the images are clustered based on category rather than color. For example, the sunset beach images at the lower right of Fig. 10, which contain predominantly oranges and blacks, are grouped together with blue sky beach images. However, in the color similarity embedding in Fig. 8 and Fig. 9, the sunset beach images are grouped together with other images that have dark orange and black colors. Overall, the embeddings on color MOS predictions in Fig. 8 and color feature distances in Fig. 9 show much more clear color clusters than the embedding on fc7 features from VGG19 in Fig. 10.

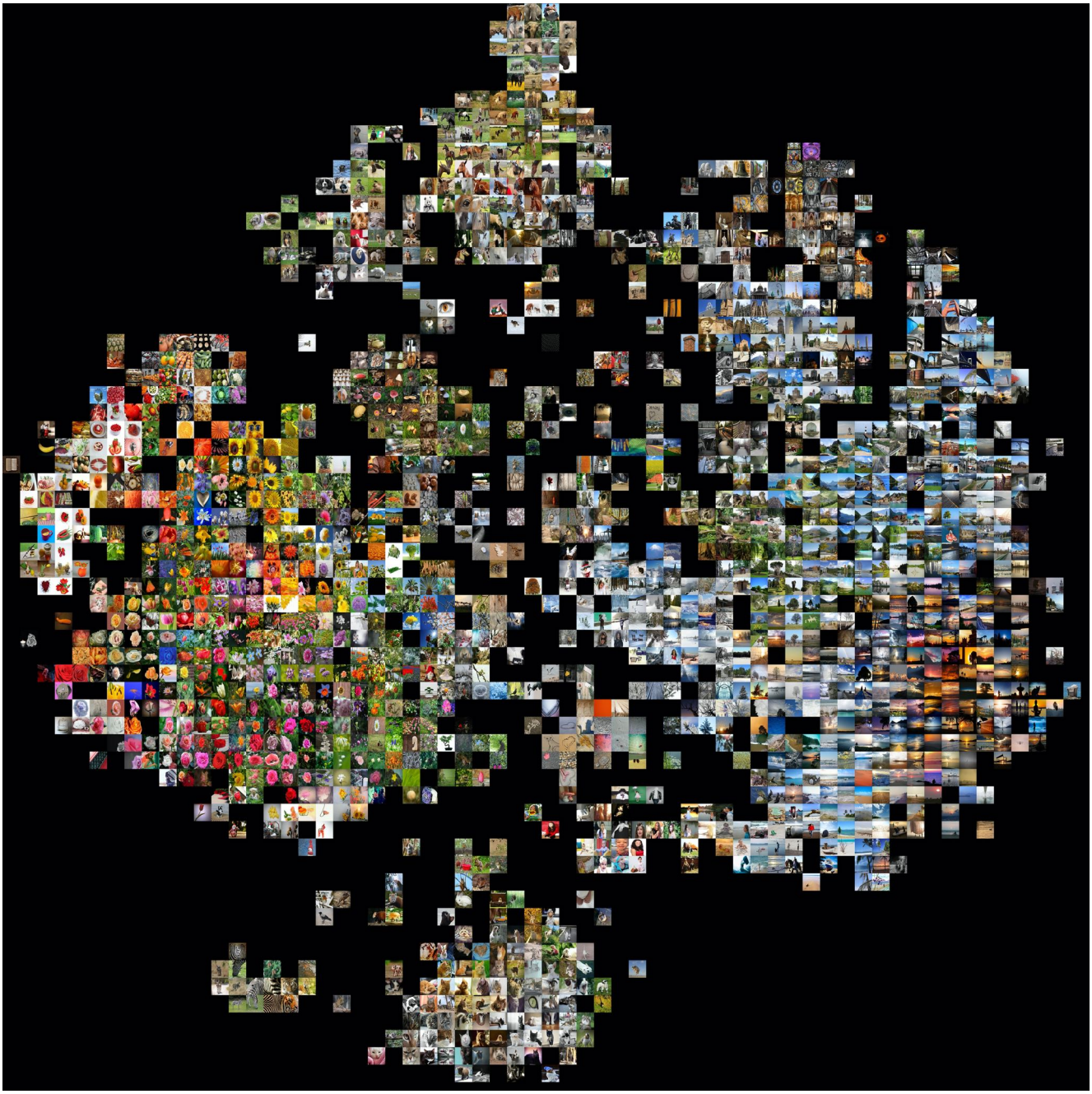


Figure 10: t-SNE embedding based on the L2 distance of fc7 content features extracted from VGG19 for 5,000 images from Pixabay. For a clear visualization, only images that do not overlap with each other are rendered.

4. Comparison between content and color retrieval

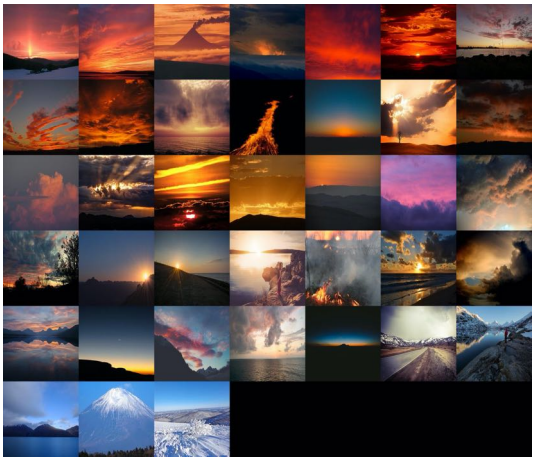
We show retrieval results for query q ranked by content and color similarity. Content similarity is based on L2 for fc7 from VGG19, and color similarity is from our color-sim network. The similarity between the query q and a retrieved image r is $\mathcal{H}(q, r)$. The visualized results show all retrieved images r such that $\mathcal{H}(q, r) > 0.5$. It shows that color similarity features produce better visual similarity ranking.

In the case of retrieving the sunflower image, all the yellow flowers are pushed up in their rank in the color similarity compared to the content similarity. Moreover, the images that contain sunflowers with a blue sky background are also higher in rank for color similarity.

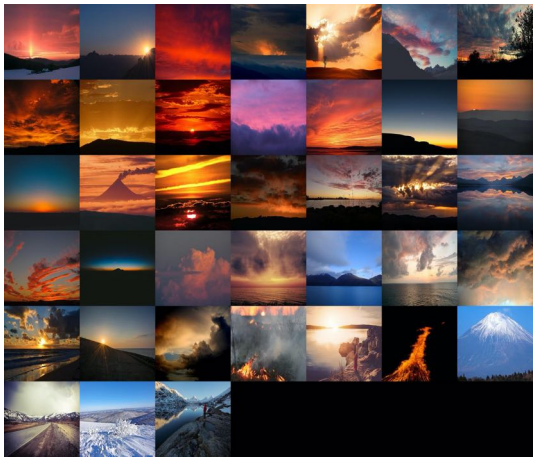
For the sunset image retrieval example, the red-orange tone images are ranked visually more correct in color similarity compared to the content similarity. Similarly, the same consistent results are noticeable for the mushroom retrieval example.

Color similarity

Content similarity



Color similarity



Content similarity

Color similarity

Content similarity

Figure 11: Retrieval results, independently ranked by content and colors. The first image in each figure is the query image.

References

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

[2] 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

[3] 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. 7