Supplementary Material

for

Effective Aesthetics Prediction with Multi-level Spatially Pooled Features

Vlad Hosu Bastian Goldlücke Dietmar Saupe
University of Konstanz, Germany
{vlad.hosu, bastian.goldluecke, dietmar.saupe}@uni-konstanz.de

1. Factors limiting model performance

1.1. Information lost when transforming images

We study the effect that two types of distortions have on the performance of two of our best models: "Single-3FC (*)(-aug)" and "Pool-3FC (*)(-aug)". The distortions are proportional rescaling and cropping of the original sized images. Cropping means we take centered crops of the same aspect ratio as the transformed image, however zoomed in by some factor, e.g., zoom = 0.5 means a crop that is half the width and height of the original. Both models work without augmentation, and use the InceptionResNet-v2 base architecture to extract narrow (Single-3FC) and wide (Pool-3FC) MLSP features respectively.

Inception networks do not accept inputs that are too small. Consequently, in order to allow large down-sizing factors, i.e., zoom = 0.3 we select images from the test set that are larger than 400 pixels in both width and height. We compute the SRCC performance metric only for the selected images from the test set ($17,903$ of the original $19,928$). The top performance of Pool-3FC on the subset, at the original image size (zoom = 1) is 0.74 SRCC, which is only slightly different from the 0.75 SRCC on the entire test set. We downscale the original images proportionally and crop at the same aspect ratio as the original, varying the zoom factors from 0.3 to 1. In Fig. 1 we see how the model performances vary with the zoom factor for both transformation types. For the narrow MLSP features, it appears that the performance when we reduce the input size to the feature extraction network depends less on the operation (rescale or crop), and more on the zoom factor. This could mean there is about an equal amount of information lost by any of the transformations. For wide MLSP features the difference between zoom and crop distortions is a bit larger. Even though the difference between the two models is very small when tested at the original image sizes, Pool-3FC performs much better on down scaled images. This is likely due to the spatial component of the wide MLSP features ($5 \times 5 \times b$ compared to $1 \times 1 \times b$ for narrow MLSP). It points to the better generalization ability of wide MLSP.

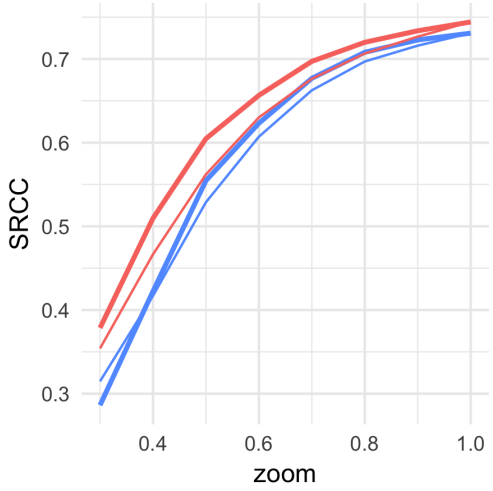

Figure 1: Performance when testing two of our best models at different image sizes from the test set. Blue: "Single-3FC (*)(-aug)", Red: "Pool-3FC (*)(-aug)". Bold line: each image is rescaled by the zoom, Thin line: each image is center cropped such that the size of the crop is the size of the image at the specified zoom factor. The difference between the two models is very small when tested at the original image sizes. However, Pool-3FC generalizes much better to down scaled images, likely due to the spatial component of wide MLSP features.

The sharp drop in performance with lower zoom factors argues for the choice of learning from features taken from original images, without applying rescaling or cropping transformations. This may also be the cause for the small performance improvement when using augmentation by cropping.

1.2. Inconsistent aesthetics scoring in AVA

1.2.1 Failure cases for low quality images

In figures 3 and 4 we show examples of images that have received a low mean opinion score in AVA. The procedure for selecting these images is the same as that presented in the main paper, which shows failure cases and correct predictions for high quality images. In Fig. 3 we notice that users have a tendency to under-rate average images. For instance, the flower images are assigned lower scores than their technical quality would suggest. The preference is noticeable when compared to images in Fig. 4 where the technical quality is low for all the examples.

1.2.2 Explanation for inconsistencies

There are two possible explanations for the inconsistent scoring in AVA:

1. Relative ratings per contest: images are posted to topical contests, and thus their ratings are relative to the other images posted to the same contest. Ratings between contests may be inconsistent.

2. Popularity measures are not aesthetics measures: Schifanella et al. [4] have suggested that the popularity of an image in terms of number of favorites it receives on Flickr.com is not highly indicative of aesthetics scores in a traditional sense, such as absolute category ratings (ACR). Schifanella et al. [4] prompt users to rate images based on aesthetics criteria: Unacceptable (1), Flawed: low quality (2), Ordinary: standard quality (3), Professional: professional-quality images (4), Exceptional: very appealing images (5). The 10 point rating scale in dpchallenge.com, and thus AVA, is more akin to a popularity measure, by which users show their preference for an image with respect to a topical contest.

The fact that our model sometimes picks up on these scoring inconsistencies may mean that the two flaws are not prevalent in the entire database. It may well be that without these flaws, our top model's performance could be higher.

1.2.3 Relation between aesthetic and technical quality

The field of image quality assessment (IQA) mainly covers technical aspects of image quality such as identifying the perceived degradation due to noise, artifacts, wrong contrast, exposure, etc. We expect that high technical quality generally supports the aesthetic experience, and influences aesthetic quality assessment (AQA). Judging technical aspects of photography is often difficult for less experienced users [2], leading to ambiguous ratings as we observe for some images in AVA.

We test for the correlation between aesthetic and technical quality ratings by running our AQA model on two of the largest IQA benchmark data sets [3, 1], that contain images in the wild, i.e., distortions are not artificially induced. We apply our best model (Pool-3FC (*)) on the original images in each database (500×500 pixels for LIVE-in-the-wild [1], 1024×768 pixels for KonIQ-10k [3]). We obtain a correlation computed between our predicted AQA ratings and the IQA ground-truth ratings for each database of 0.57 (SRCC) for LIVE-in-the-wild and 0.60 (SRCC) for KonIQ-10k. This suggests that our AQA model has some knowledge about technical quality assessment.

1.3. Effects of subjective scoring

Is the performance we obtain with our best model limited by the subjective nature of the aesthetic scores? We take a simple indicator for subjectivity, the standard deviation of the ground-truth scores (SDS) computed for each image. We find a weak correlation of 0.2 SRCC between the SDS and the absolute error between predicted and actual MOS on the test set. This suggests subjectivity affects our model performance, but weakly.

In Fig. 2 we take a closer look at the errors our best model makes (absolute error between actual and predicted MOS) at different SDS. For sliding windows of $1,000$ images, ordered by SDS, we show the mean absolute error (MAE) as a function of the mean SDS per window. As expected, we notice an increase in errors with the SDS. Figure 2 suggests that our model performance has to suffer because of highly subjective judgments.

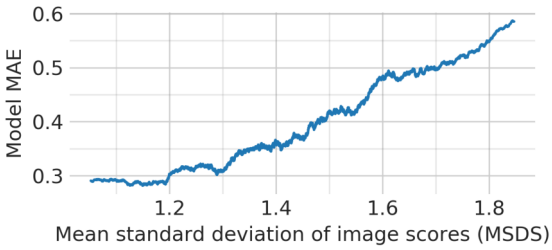

Figure 2: Higher subjectivity (large MSDS) decreases our best method's performance. Test images are sorted according to the SD of their AVA scores. With a sliding window of $1,000$ images, the MAE between the predictions and the ground-truth MOS is computed and shown as a function of the mean standard deviation of the scores (MSDS) for the images in the sliding window. The MAE for the entire test set is 0.39.

2. Top and bottom rated images, according to users and our DNN model

In figures 5 and 7 we show the first and last 35 images respectively, ranked by their AVA MOS values (from users). In figures 6 and 8 we do the same, but based on scores produced by our model (Pool-3FC (*)). To be fair, only images from the test set are considered.

Top images according to user ratings (Fig. 5) show a wider diversity of styles, while our model (Fig. 6) prefers dramatic pictures, including high contrast city-scapes, portraits, and natural landscapes. Some top images w.r.t. MOS do not appear to have a high technical quality, but show an interesting subject, e.g., dragonfly head macro, child at fruit seller, snail on red background. There are 9 (25.7%) common results in the top 35. If we look at the top $1,000$, the overlap increases to 37.8%.

For the bottom retrieval results, 7 of 35 are common (20%), while among the bottom $1,000$ there are 47.8% common images.

References

[1] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *Transactions on Image Processing (TIP)*, 25(1):372–387, 2016. 2

[2] V. Hosu, H. Lin, and D. Saupe. Expertise screening in crowdsourcing image quality. In *Quality of Multimedia Experience (QoMEX)*, 2018. 2

[3] H. Lin, V. Hosu, and D. Saupe. Koniq-10k: Towards an ecologically valid and large-scale iqa database. *arXiv preprint arXiv:1803.08489*, 2018. 2

[4] R. Schifanella, M. Redi, and L. M. Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *International AAAI Conference on Web and Social Media (ICWSM)*, pages 397–406, 2015. 2

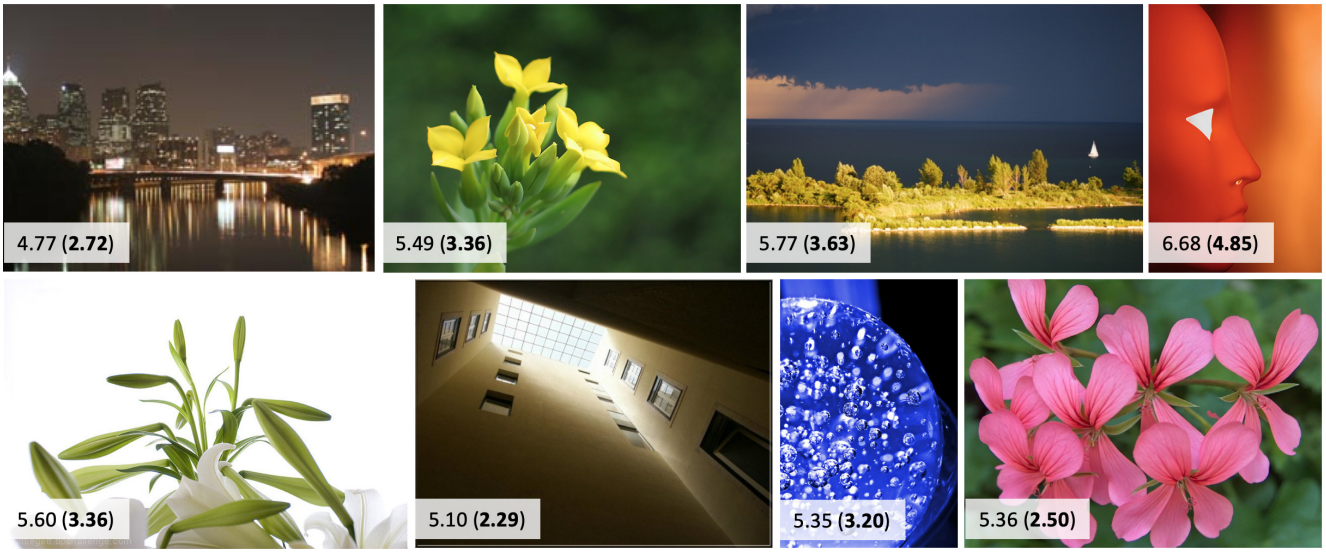

Figure 3: Low quality images from the test set, for which our best model's assessment errors are some of the highest (using Pool-3FC with wide MLSP features from InceptionResNet-v2). Our predicted score is the number on the left in each image while the ground-truth MOS is shown in brackets. Users have a tendency to under-rate average images. For instance, the flower images are assigned lower scores than their technical quality would suggest. The preference is more noticeable when compared to images in Fig. 4 where examples of a lower technical quality are shown, but which have been assigned higher user ratings.

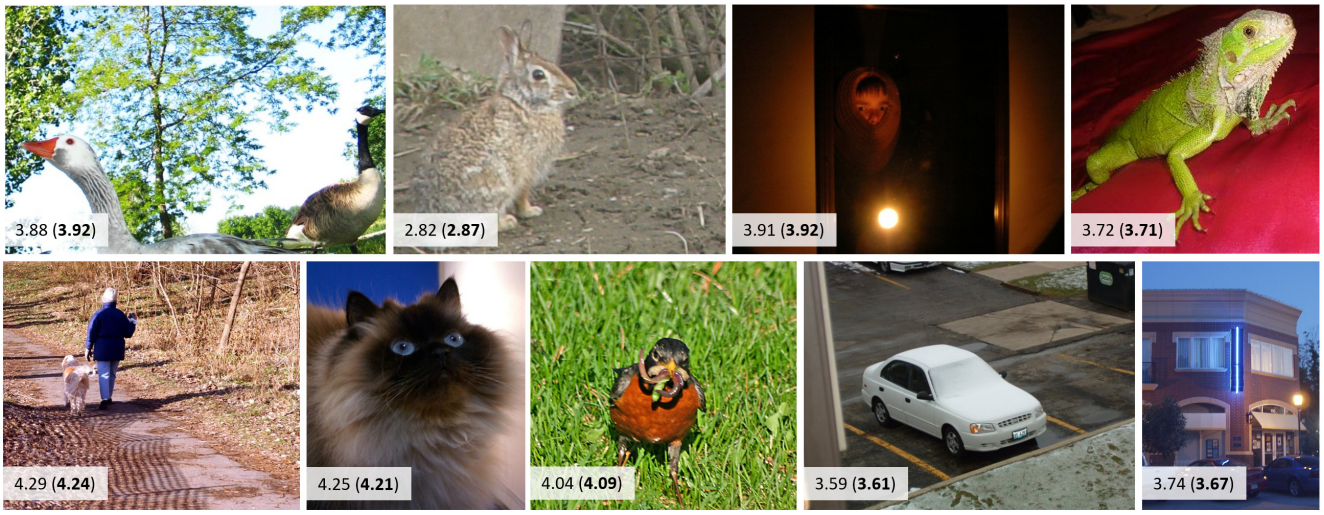

Figure 4: Low quality images from the test set, for which our best model's assessment errors are small (using Pool-3FC with wide MLSP features from InceptionResNet-v2). Our predicted score is the number on the left in each image while the ground-truth MOS is shown in brackets.

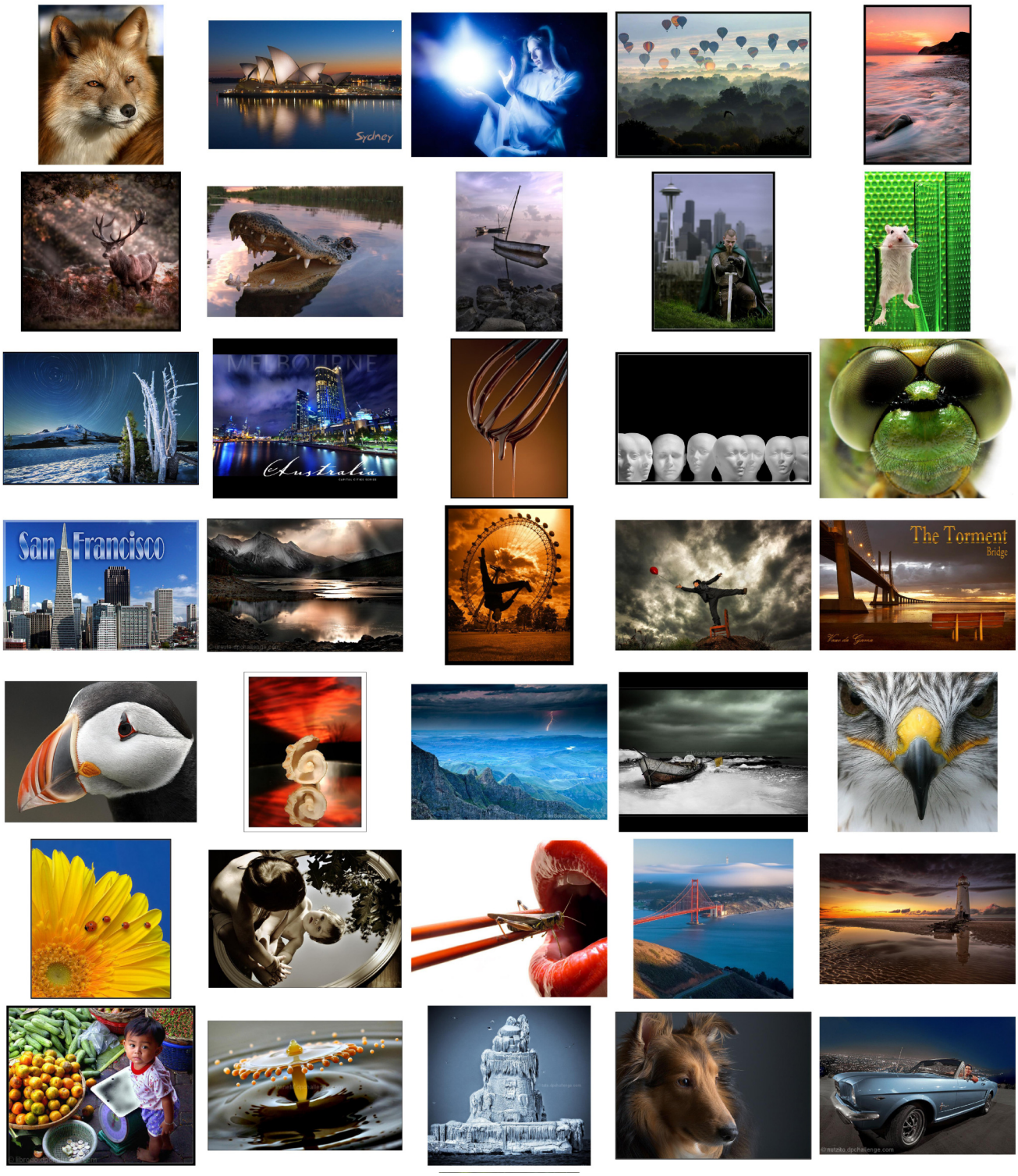

Figure 5: Top 35 images from the test set based on AVA mean opinion scores (from user ratings). Best scoring image is shown in the top left. The images show a wide diversity of styles. Some top images do not appear to have a high technical quality, but show an interesting subject, e.g., dragonfly head macro, child at fruit seller, snail on red background.

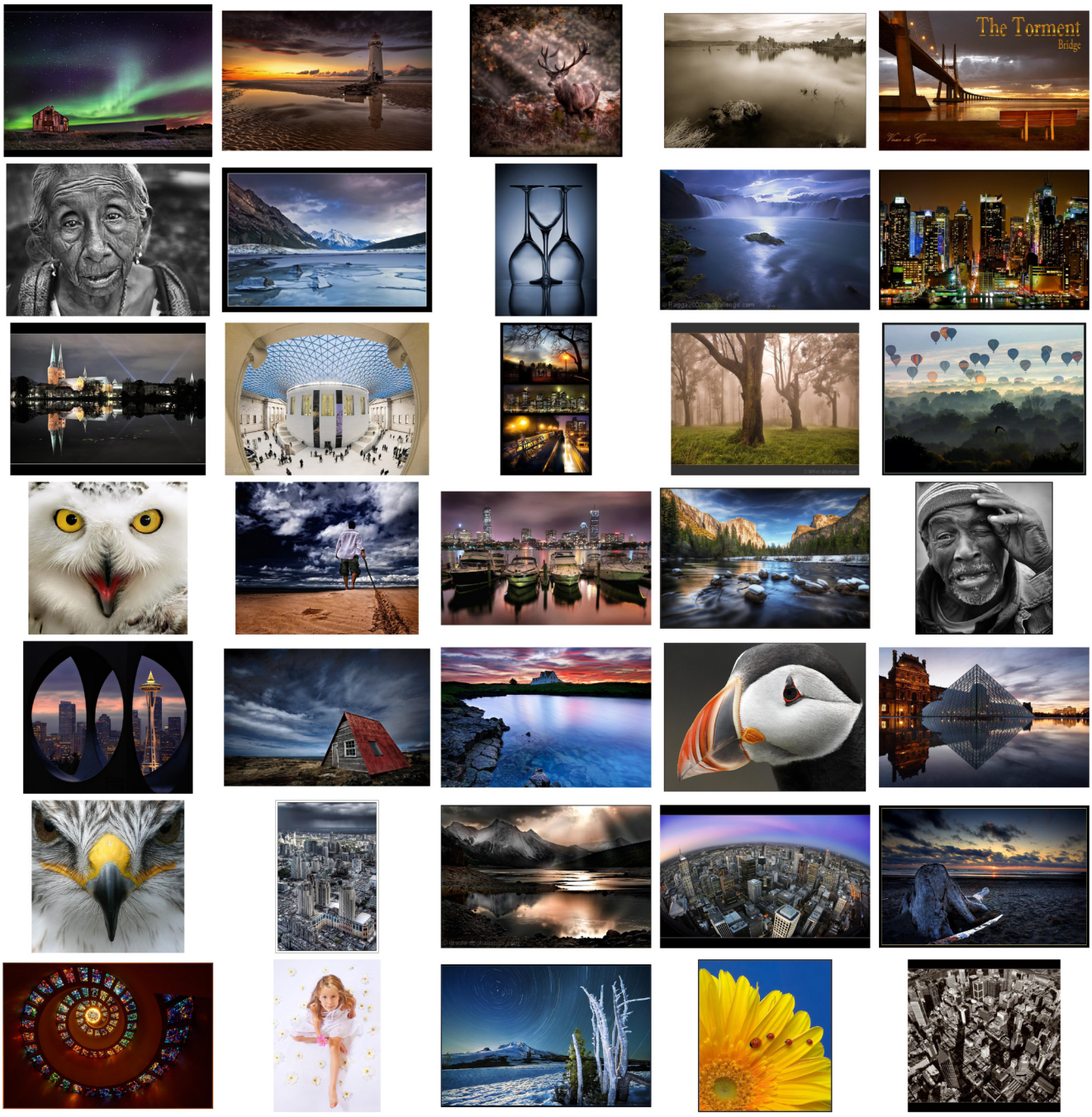

Figure 6: Top 35 scoring images from the test set as predicted by or best model (Pool-3FC with wide MLSP features from InceptionResNet-v2). Best scoring image is shown in the top left. Our model seems to prefer dramatic pictures, such as high contrast portraits, city-scapes and natural landscapes. There are 9 (25.7%) common results with the user MOS based ranking, in the top 35. If we look at the top $1,000$, the overlap increases to 37.8%.

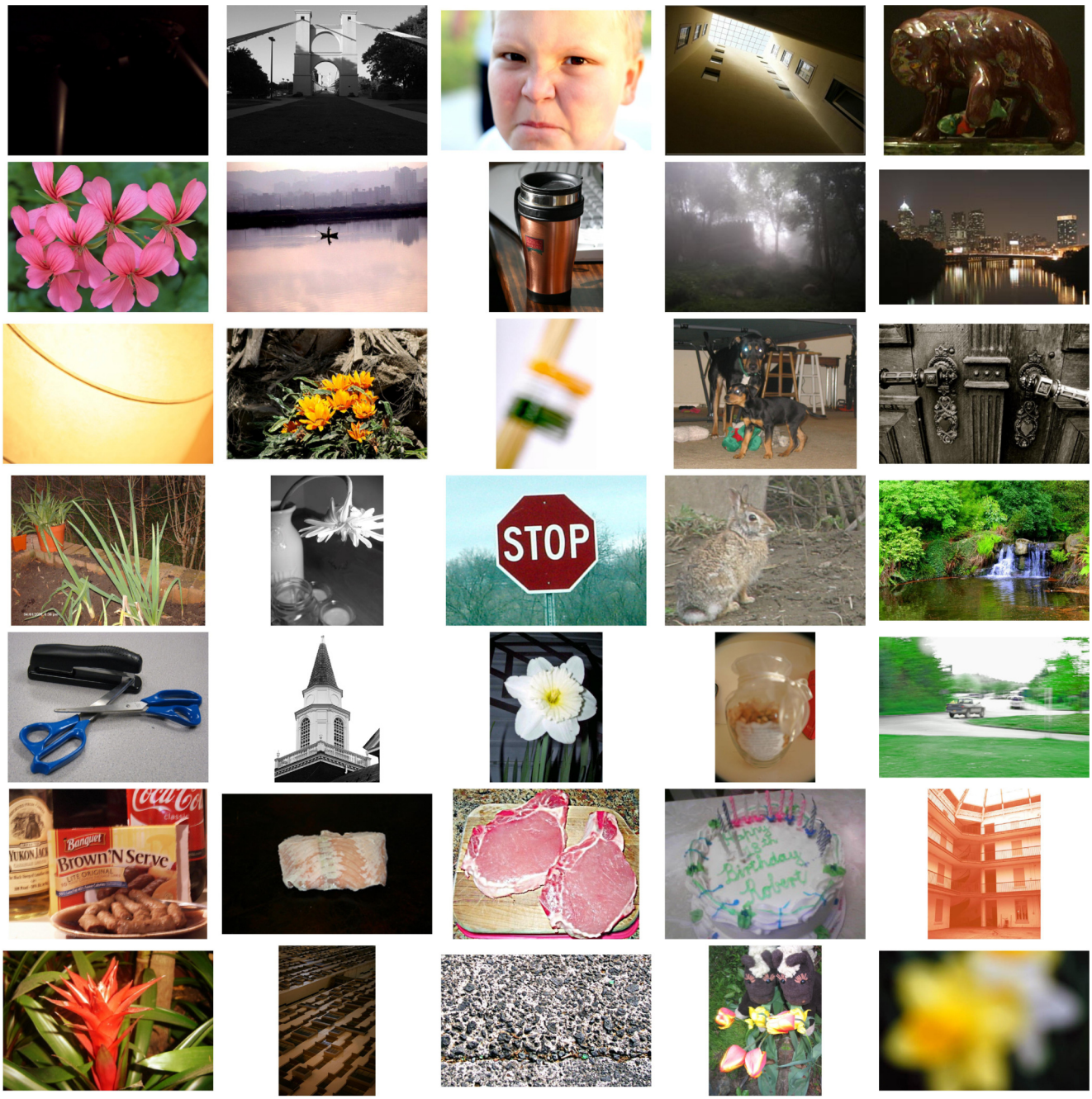

Figure 7: Bottom 35 images from the test set based on AVA mean opinion scores (from users). Lowest scoring image is shown in the top left.

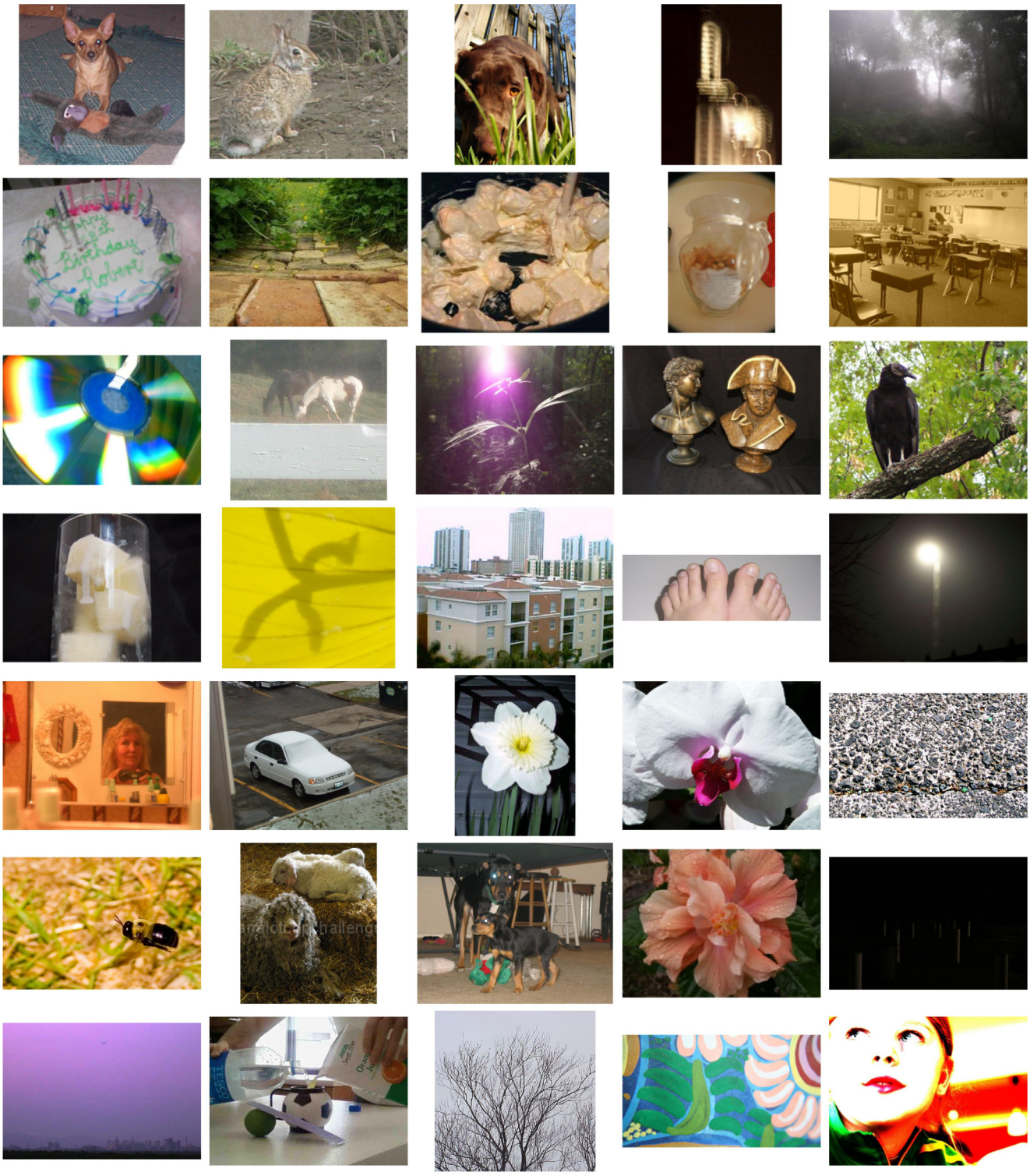

Figure 8: Bottom 35 scoring images from the test set as predicted by or best model (Pool-3FC with wide MLSP features from InceptionResNet-v2). Lowest scoring image is shown in the top left. 7 (20%) of the images shown are common with the user ranking (based on MOS), while among the bottom $1,000$ there are 47.8% common images.