

# NIMA: Neural Image Assessment

Hossein Talebi and Peyman Milanfar

**Abstract**— Automatically learned quality assessment for images has recently become a hot topic due to its usefulness in a wide variety of applications such as evaluating image capture pipelines, storage techniques and sharing media. Despite the subjective nature of this problem, most existing methods only predict the *mean* opinion score provided by datasets such as AVA [1] and TID2013 [2]. Our approach differs from others in that we predict the *distribution* of human opinion scores using a convolutional neural network. Our architecture also has the advantage of being significantly simpler than other methods with comparable performance. Our proposed approach relies on the success (and retraining) of proven, state-of-the-art deep object recognition networks. Our resulting network can be used to not only score images reliably and with high correlation to human perception, but also to assist with adaptation and optimization of photo editing/enhancement algorithms in a photographic pipeline. All this is done without need of a “golden” reference image, consequently allowing for single-image, semantic- and perceptually-aware, *no-reference* quality assessment.

## I. INTRODUCTION

Quantification of image quality and aesthetics have been a long-standing problem in image processing and computer vision. While technical quality assessment deals with measuring low-level degradations such as noise, blur, compression artifacts, etc., aesthetic assessment quantifies semantic level characteristics associated with emotions and beauty in images. In general, image quality assessment can be categorized into full-reference and no-reference approaches. While availability of a reference image is assumed in the former (metrics such as PSNR, SSIM [3], etc.), typically blind (no-reference) approaches rely on a statistical model of distortions to predict image quality. The main goal of both categories is to predict a quality score that correlates well with human perception. Yet, the subjective nature of image quality remains the fundamental issue. Recently, more complex models such as deep convolutional neural networks (CNNs) have been used to address this problem [4]–[11]. Emergence of labeled data by human ratings has encouraged these efforts [1], [2], [12]–[14]. In a typical deep CNN approach, weights are initialized by training on classification related datasets (e.g. ImageNet [15]), and then fine tuned on annotated data for perceptual quality assessment tasks.

### A. Related Work

Machine learning has shown promising success in predicting technical quality of images [4]–[7]. Kang et. al. [5] show that extracting high level features using CNNs can result in state-of-the-art blind quality assessment performance. It appears that replacing hand-crafted features with an end-to-end feature

H. Talebi and P. Milanfar are with Google Research, Mountain View, USA, Email: {htalebi, milanfar}@google.com.

learning system is the main advantage of using CNNs for pixel-level quality assessment tasks [5], [6]. The proposed method in [5] is a shallow network with one convolutional layer and two fully-connected layers, and input patches are of size  $32 \times 32$ . Bosse et al. [6] use a deep CNN with 12 layers to improve on image quality predictions of [5]. Given the small input size ( $32 \times 32$  patch), both methods require score aggregation across the whole image. Bianco et al. in [7] propose a deep quality predictor based on AlexNet [15]. Multiple CNN features are extracted from image crops of size  $227 \times 227$ , and then regressed to the human scores.

Success of CNNs on object recognition tasks has significantly benefited the research on aesthetic assessment. This seems natural, as semantic level qualities are directly related to image content. Recent CNN-based methods [8]–[11] show a significant performance improvement compared to earlier works based on hand-crafted features [1]. Murray et al. [1] is the benchmark on aesthetic assessment. They introduce the AVA dataset and propose a technique to use manually designed features for style classification. Later, Lu et al. [8], [16] show that deep CNNs are well suited to the aesthetic assessment task. Their double-column CNN [16] consists of four convolutional and two fully-connected layers, and its inputs are the resized image and cropped windows of size  $224 \times 224$ . Predictions from these global and local image views are aggregated to an overall score by a fully-connected layer. Similar to Murray et al. [1], in [16] images are also categorized to low and high aesthetics based on mean human ratings. A regression loss and an AlexNet inspired architecture is used in [9] to predict the mean scores. In a similar approach to [9], Bin et al. [11] fine-tune a VGG network [17] to learn the human ratings of the AVA dataset. They use a regression framework to predict the histogram of ratings. More recently, [10] uses an adaptive spatial pooling to allow for feeding multiple scales of the input image with fixed size aspect ratios to their CNN. This work presents a multi-net (each network a pre-trained VGG) approach which extracts features at multiple scales, and uses a scene aware aggregation layer to combine predictions of the sub-networks. Similarly, Ma et al. [18] propose a layout-aware framework in which a saliency map is used to select patches with highest impact on predicted aesthetic score. Overall, none of these methods reported correlation of their predictions with respect to ground truth ratings. Recently, Kong et al. in [14] proposed a method to aesthetically rank photos by training on AVA with a rank-based loss function. They trained an AlexNet-based CNN to learn the difference of the aesthetic scores from two input images, and as a result, indirectly optimize for rank correlation. To the best of our knowledge, [14] is the only work that performed a correlation evaluation against AVA ratings.

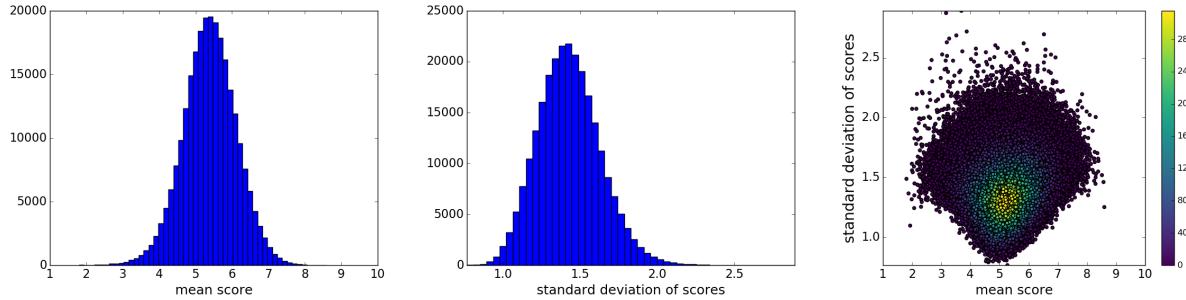


Fig. 1: Histograms of ratings from AVA dataset [1]. Left: Histogram of mean scores. Middle: Histogram of standard deviations. Right: Joint histogram of the mean and standard deviation.

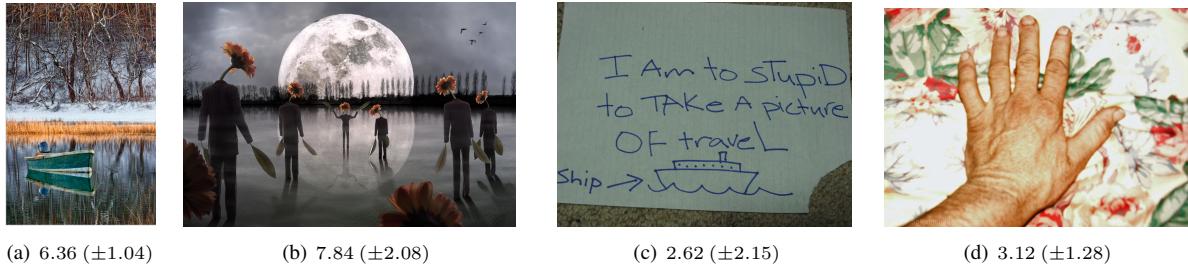


Fig. 2: Some example images from AVA dataset [1] with quality score  $\mu(\pm\sigma)$ , where  $\mu$  and  $\sigma$  represent mean and standard deviation of score, respectively. (a) high aesthetics and low unconventionality (challenge name: “*Best of 2007*”,  $\mu = 6.36$ ,  $\sigma = 1.04$ ), (b) high aesthetics and high unconventionality (challenge name: “*Extreme super moon*”,  $\mu = 7.84$ ,  $\sigma = 2.08$ ), (c) low aesthetics and high unconventionality (challenge name: “*Travel*”,  $\mu = 2.62$ ,  $\sigma = 2.15$ ), (d) low aesthetics and low unconventionality (challenge name: “*Pieces of the human form*”,  $\mu = 3.12$ ,  $\sigma = 1.28$ ).

### B. Our Contributions

In this work, we introduce a novel approach to predict both technical and aesthetic qualities of images. We show that models with the same CNN architecture, trained on different datasets, lead to state-of-the-art performance for both tasks. Since we aim for predictions with higher correlation with human ratings, instead of classifying images to low/high score or regressing to the mean score, the distribution of ratings are predicted as a histogram. To this end, we use the squared EMD (earth mover’s distance) loss proposed in [19], which shows a performance boost in classification with ordered classes. Our experiments show that this approach also leads to more accurate prediction of the mean score. Also, as shown in aesthetic assessment case [1], non-conventionality of images is directly related to score deviations. Our proposed paradigm allows for predicting this metric as well. This paper begins with reviewing two widely used datasets for quality assessment. Then, our proposed method is explained in more detail. Finally, performance of this work is quantified and compared to the existing methods.

### C. A Large-Scale Database for Aesthetic Visual Analysis (AVA) [1]

The AVA dataset contains about 255,000 images, rated based on aesthetic qualities by amateur photographers<sup>1</sup>. Each

photo is scored by an average of 200 people in response to photography contests. Each image is associated to a single challenge theme, with nearly 900 different contests in the AVA. The image ratings range from 1 to 10, with 10 being the highest aesthetic score associated to an image. Histograms of AVA ratings are shown in Fig. 1. As can be seen, mean ratings are concentrated around the overall mean score ( $\approx 5.5$ ). Also, ratings of roughly half of the photos in AVA dataset have a standard deviation greater than 1.4. As pointed out in [1], presumably images with high score variance tend to be subject to interpretation, whereas images with low score variance seem to represent conventional styles or subject matter. A few examples with ratings associated with different levels of aesthetic quality and unconventionality are illustrated in Fig. 2. It seems that aesthetic quality of a photograph can be represented by the mean score, and unconventionality of it closely correlates to the score deviation. Given the distribution of AVA scores, typically, training a model on AVA data results in predictions with small deviations around the overall mean (5.5).

It is worth mentioning that the joint histogram in Fig. 1 shows higher deviations for very low/high ratings (compared to the overall mean 5.5, and mean standard deviation 1.43). In other words, divergence of opinion is more consistent in AVA images with extreme aesthetic qualities. As discussed in [1], distribution of ratings with mean value between 2 and 8 can be closely approximated by Gaussian functions, and highly skewed ratings can be modeled by Gamma distributions.

<sup>1</sup> AVA images are obtained from [www.dpchallenge.com](http://www.dpchallenge.com), which is an on-line community for amateur photographers.

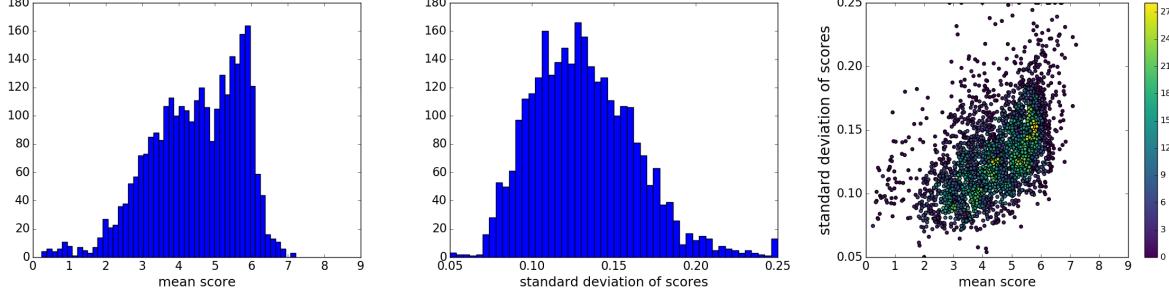


Fig. 3: Histograms of ratings from TID2013 dataset [2]. Left: Histogram of mean scores. Middle: Histogram of standard deviations. Right: Joint histogram of the mean and standard deviation.

#### D. Tampere Image Database 2013 (TID2013) [2]

TID2013 is curated for evaluation of full-reference perceptual image quality. It contains 3000 images, from 25 reference (clean) images (Kodak images [20]), 24 types of distortions with 5 levels for each distortion. This leads to 120 distorted images for each reference image; including different types of distortions such as compression artifacts, noise, blur and color artifacts.

Human ratings of TID2013 images are collected through a forced choice experiment, where observers select a better image between two distorted choices. Set up of the experiment allows raters to view the reference image while making a decision. In each experiment, every distorted image is used in 9 random pairwise comparisons. The selected image gets one point, and other image gets zero points. At the end of the experiment, sum of the points is used as the quality score associated with an image (this leads to scores ranging from 0 to 9). To obtain the overall mean scores, total of 985 experiments are carried out.

Mean and standard deviation of TID2013 ratings are shown in Fig. 3. As can be seen in Fig. 3(c), the mean and score deviation values are weakly correlated. A few images from TID2013 are illustrated in Fig. 4 and Fig. 5. All five levels of JPEG compression artifacts and the respective ratings are illustrated in Fig. 4. Evidently higher distortion level leads to lower mean score<sup>2</sup>. Effect of contrast compression/stretching distortion on the human ratings is demonstrated in Fig. 5. Interestingly, stretch of contrast (Fig. 5(c) and Fig. 5(e)) leads to relatively higher perceptual quality.

Unlike AVA, which includes distribution of ratings for each image, TID2013 only provides mean and standard deviation of the opinion scores. Since our proposed method requires training on score probabilities, the score distributions are approximated through maximum entropy optimization [21].

## II. PROPOSED METHOD

Our proposed quality and aesthetic predictor stands on image classifier architectures. More explicitly, we explore a few different classifier architectures such as VGG16 [17], Inception-v2 [22], and MobileNet [23] for image quality

<sup>2</sup>This is a quite consistent trend for most of the other distortions too (namely noise, blur and color distortions). However, in case of the contrast change (Fig. 5), this trend is not obvious. This is due to the order of contrast compression/stretching from level 1 to level 5)

assessment task. VGG16 consists of 13 convolutional and 3 fully-connected layers. Small convolution filters of size  $3 \times 3$  are used in the deep VGG16 architecture [17]. Inception-v2 [22] is based on Inception module [24] which allows for parallel use of convolution and pooling operations. Also, in the Inception architecture, traditional fully-connected layers are replaced by average pooling, which leads to a significant reduction in number of parameters. MobileNet [23] is an efficient deep CNN, mainly designed for mobile vision applications. In this architecture, dense convolutional filters are replaced by separable depth filters. This simplification results in smaller and faster CNN models.

We replaced the last layer of the baseline CNN with a fully-connected layer with 10 neurons followed by soft-max activations (shown in Fig. 6). Baseline CNN weights are initialized by training on the ImageNet dataset [15], and then an end-to-end training on quality assessment is performed. In this paper, we discuss performance of the proposed model with various baseline CNNs.

In training, input images are rescaled to  $256 \times 256$ , and then a crop of size  $224 \times 224$  crop is randomly extracted. This lessens potential over-fitting issues, especially when training on relatively small datasets (e.g. TID2013). It is worth noting that we also tried training with random crops without rescaling. However, results were not compelling. This is due to the inevitable change in image composition. Another random data augmentation in our training process is horizontal flipping of the image crops.

Our goal is to predict the distribution of ratings for a given image. Ground truth distribution of human ratings of a given image can be expressed as an empirical probability mass function  $\mathbf{p} = [p_{s_1}, \dots, p_{s_N}]$  with  $s_1 \leq s_i \leq s_N$ , where  $s_i$  denotes the  $i$ th score bucket, and  $N$  denotes the total number of score buckets. In both AVA and TID2013 datasets  $N = 10$ , in AVA,  $s_1 = 1$  and  $s_N = 10$ , and in TID  $s_1 = 0$  and  $s_N = 9$ . Since  $\sum_{i=1}^N p_{s_i} = 1$ ,  $p_{s_i}$  represents the probability of a quality score falling in the  $i$ th bucket. Given the distribution of ratings as  $\mathbf{p}$ , mean quality score is defined as  $\mu = \sum_{i=1}^N s_i \times p_{s_i}$ , and standard deviation of the score is computed as  $\sigma = (\sum_{i=1}^N (s_i - \mu)^2 \times p_{s_i})^{1/2}$ . As discussed in the previous section, one can qualitatively compare images by mean and standard deviation of scores.

Each example in the dataset consists of an image and its ground truth (user) ratings  $\mathbf{p}$ . Our objective is to find the



Fig. 4: JPEG artifact example images from TID2013 dataset [2] with quality score  $\mu(\pm\sigma)$ , where  $\mu$  and  $\sigma$  represent mean and standard deviation of score, respectively. Clean image and 5 levels of JPEG compression artifacts are shown here. (a) clean image, (b) compression artifact level 1,  $\mu = 5.73$ ,  $\sigma = 0.15$ , (c) compression artifact level 2,  $\mu = 5.47$ ,  $\sigma = 0.11$ , (d) compression artifact level 3,  $\mu = 4.86$ ,  $\sigma = 0.11$ , (e) compression artifact level 4,  $\mu = 3.0$ ,  $\sigma = 0.11$ , (f) compression artifact level 5,  $\mu = 1.66$ ,  $\sigma = 0.16$ .

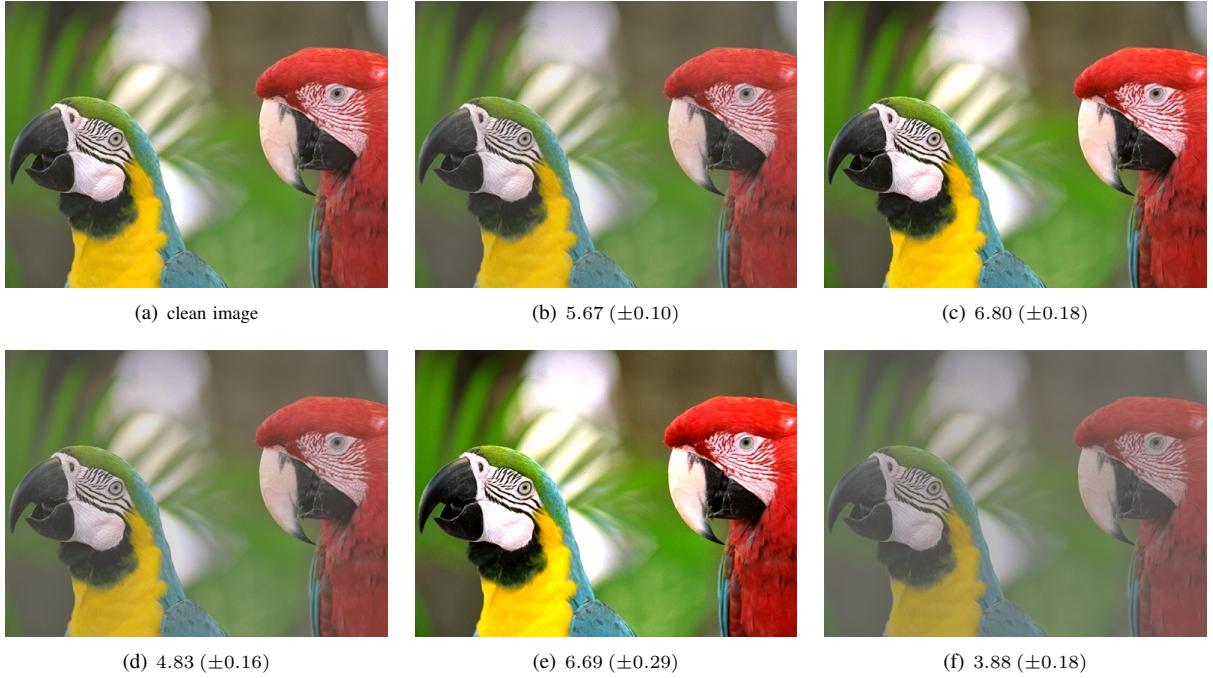


Fig. 5: Some example images from TID2013 dataset [2] with quality score  $\mu(\pm\sigma)$ , where  $\mu$  and  $\sigma$  represent mean and standard deviation of score, respectively. Clean image and 5 levels of contrast change distortions are shown here. (a) clean image, (b) contrast change distortion of level 1,  $\mu = 5.67$ ,  $\sigma = 0.10$ , (c) contrast change distortion of level 2,  $\mu = 6.80$ ,  $\sigma = 0.18$ , (d) contrast change distortion of level 3,  $\mu = 4.83$ ,  $\sigma = 0.16$ , (e) contrast change distortion of level 4,  $\mu = 6.69$ ,  $\sigma = 0.29$ , (f) contrast change distortion of level 5,  $\mu = 3.88$ ,  $\sigma = 0.18$ .

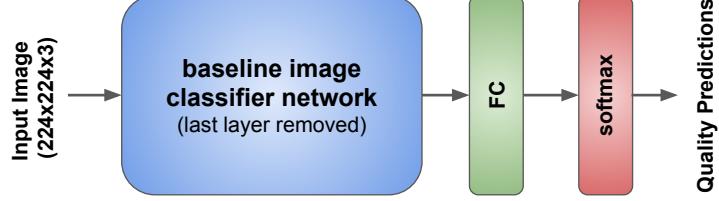


Fig. 6: Modified baseline image classifier network used in our framework. Last layer of classifier network is replaced by a fully-connected layer to output 10 classes of quality scores. Baseline network weights are initialized by training on ImageNet dataset [15], and the added fully-connected weights are initialized randomly.

probability mass function  $\hat{\mathbf{p}}$  that is an accurate estimate of  $\mathbf{p}$ . Next, our training loss function is discussed.

#### A. Loss Function

Soft-max cross-entropy is widely used as training loss in classification tasks. This loss can be represented as  $\sum_{i=1}^N -p_{s_i} \log(\hat{p}_{s_i})$  (where  $\hat{p}_{s_i}$  denotes estimated probability of  $i$ th score bucket) to maximize predicted probability of the correct labels. However, in the case of ordered-classes (e.g. aesthetic and quality estimation), cross-entropy loss lacks the inter-class relationships between score buckets. One might argue that ordered-classes can be represented by a real number, and consequently, can be learned through a regression framework. Yet, it has been shown that for ordered classes, classification frameworks can outperform regression models [19], [25]. Hou et al. [19] show that training on datasets with intrinsic ordering between classes can benefit from EMD-based losses. These loss functions penalize mis-classifications according to class distances.

For image quality ratings, classes are inherently ordered as  $s_1 < \dots < s_N$ , and  $r$ -norm distance between classes is defined as  $\|s_i - s_j\|_r$ , where  $1 \leq i, j \leq N$ . EMD is defined as the minimum cost to move the mass of one distribution to another. Given the ground truth and estimated probability mass functions  $\mathbf{p}$  and  $\hat{\mathbf{p}}$ , with  $N$  ordered classes of distance  $\|s_i - s_j\|_r$ , the normalized Earth Mover's Distance can be expressed as [26]:

$$\text{EMD}(\mathbf{p}, \hat{\mathbf{p}}) = \left( \frac{1}{N} \sum_{k=1}^N |\text{CDF}_{\mathbf{p}}(k) - \text{CDF}_{\hat{\mathbf{p}}}(k)|^r \right)^{1/r} \quad (1)$$

where  $\text{CDF}_{\mathbf{p}}(k)$  is the cumulative distribution function as  $\sum_{i=1}^k \mathbf{p}_{s_i}$ . It is worth noting that this closed-form solution requires both distributions to have equal mass as  $\sum_{i=1}^N \mathbf{p}_{s_i} = \sum_{i=1}^N \hat{\mathbf{p}}_{s_i}$ . As shown in Fig. 6, our predicted quality probabilities are fed to a soft-max function to guarantee that  $\sum_{i=1}^N \hat{\mathbf{p}}_{s_i} = 1$ . Similar to [19], in our training framework,  $r$  is set as 2 to penalize the Euclidean distance between the CDFs.  $r = 2$  allows easier optimization when working with gradient descent.

### III. EXPERIMENTAL RESULTS

We train two separate models for aesthetics and technical quality assessment on AVA and TID2013. For each case, we

split the AVA and TID datasets to train and test sets, such that 20% of the data is used for testing. In this section, performance of the proposed models on the test sets are discussed and compared to the existing methods. Then, applications of the proposed technique in photo ranking and image enhancement are explored. Before moving forward, details of our implementation are explained.

The CNNs presented in this paper are implemented using TensorFlow [27], [28]. The baseline CNN weights are initialized by training on ImageNet [15], and the last fully-connected layer is randomly initialized. The weight and bias momentums are set to 0.9, and a dropout rate of 0.75 is applied on the last layer of the baseline network. The learning rate of the baseline CNN layers and the last fully-connected layers are set as  $3 \times 10^{-7}$  and  $3 \times 10^{-6}$ , respectively. We observed that setting a low learning rate on baseline CNN layers results in easier and faster optimization when using stochastic gradient descent. Also, after every 10 epochs of training, an exponential decay with decay factor 0.95 is applied to all learning rates.

#### A. Performance Comparisons

Accuracy, correlation and EMD values of our evaluations on the aesthetic assessment model on AVA are presented in Table I. Most methods in Table I are designed to perform binary classification on the aesthetic scores, and as a result, only accuracy evaluations of two-class quality categorization are reported. In two-class aesthetic categorization task, results from [18], and NIMA(Inception-v2) show the highest accuracy. Also, in terms of rank correlation, NIMA(VGG16) and NIMA(Inception-v2) outperform [14]. *NIMA is much cheaper*: [18] applies multiple VGG16 nets on image patches to generate a single quality score, whereas computational complexity of NIMA(Inception-v2) is roughly one pass of Inception-v2 (see Table III).

Our technical quality assessment model on TID2013 is compared to other existing methods in Table II. While most of these methods regress to the mean opinion score, our proposed technique predicts the distribution of ratings, as well as mean opinion score. Correlation between ground truth and results of NIMA(VGG16) are close to the state-of-the-art results in [29] and [7]. It is worth highlighting that Bianco et al. [7] feed multiple image crops to a deep CNN, whereas our method takes only the rescaled image.

TABLE I: Performance of the proposed method with various architectures in predicting AVA quality ratings [1] compared to the state-of-the-art. Reported accuracy values are based on classification of photos to two classes (column 2). LCC (linear correlation coefficient) and SRCC (Spearman’s rank correlation coefficient) are computed between predicted and ground truth mean scores (column 3 and 4) and standard deviation of scores (column 5 and 6). EMD measures closeness of the predicted and ground truth rating distributions with  $r = 1$  in Eq. 1.

Model	Accuracy (2 classes)	LCC (mean)	SRCC (mean)	LCC (std.dev)	SRCC (std.dev)	EMD
Murray et al. [1]	66.70%	–	–	–	–	–
Kao et al. [9]	71.42%	–	–	–	–	–
Lu et al. [30]	74.46%	–	–	–	–	–
Lu et al. [16]	75.42%	–	–	–	–	–
Kao et al. [31]	76.58%	–	–	–	–	–
Wang et al. [32]	76.80%	–	–	–	–	–
Mai et al. [10]	77.10%	–	–	–	–	–
Kong et al. [14]	77.33%	–	0.558	–	–	–
Ma et al. [18]	81.70%	–	–	–	–	–
NIMA(MobileNet)	80.36%	0.518	0.510	0.152	0.137	0.081
NIMA(VGG16)	80.60%	0.610	0.592	0.205	0.202	0.052
NIMA(Inception-v2)	81.51%	0.636	0.612	0.233	0.218	0.050

### B. Photo Ranking

Predicted mean scores can be used to rank photos, aesthetically. Some test photos from AVA dataset are ranked in Fig. 7 and Fig. 8. Predicted NIMA scores and ground truth AVA scores are shown below each image. Results in Fig. 7 suggest that in addition to image content, other factors such as tone, contrast and composition of photos are important aesthetic qualities. Also, as shown in Fig. 8, besides image semantics, framing and color palette are key qualities in these photos. These aesthetic attributes are closely predicted by our trained models on AVA.

Predicted mean scores are used to qualitatively rank photos in Fig. 9. These images are part of our TID2013 test set, which contain various types and levels of distortions. Comparing ground truth and predicted scores indicates that our trained model on TID2013 accurately ranks the test images.

### C. Image Enhancement

Quality and aesthetic scores can be used to perceptually tune image enhancement operators. In other words, maximizing NIMA score as a prior can increase the likelihood of enhancing perceptual quality of an image. Typically, parameters of enhancement operators such as image denoising and contrast enhancement are selected by extensive experiments under various photographic conditions. Perceptual tuning could be quite expensive and time consuming, especially when human opinion is required. In this section, our proposed models are used to tune a tone enhancement method [37], and an image denoiser [38].

The multi-layer Laplacian technique [37] enhances local and global contrast of images. Parameters of this method control the amount of detail, shadow, and brightness of an image. Fig. 10 shows a few examples of the multi-layer Laplacian with different sets of parameters. We observed that the predicted aesthetic ratings from training on the AVA dataset

can be improved by contrast adjustments. Consequently, our model is able to guide the multi-layer Laplacian filter to find aesthetically near-optimal settings of its parameters. Examples of this type of image editing are represented in Fig. 11, where a combination of detail, shadow and brightness change is applied on each image. In each example, 6 levels of detail boost, 11 levels of shadow change, and 11 levels of brightness change account for a total of 726 variations. The aesthetic assessment model tends to prefer high contrast images with boosted details. This is consistent with the ground truth results from AVA illustrated in Fig. 7.

Turbo denoising [38] is a technique which uses the domain transform [39] as its core filter. Performance of Turbo denoising depends on spatial and range smoothing parameters, and consequently, proper tuning of these parameters can effectively boost performance of the denoiser. We observed that varying the spatial smoothing parameter makes the most significant perceptual difference, and as a result, we use our quality assessment model trained on TID2013 dataset to tune this denoiser. Application of our no-reference quality metric as a prior in image denoising is similar to the work of Zhu et al. [40], [41]. Our results are shown in Fig. 12. Additive white Gaussian noise with standard deviation 30 is added to the clean image, and Turbo denoising with various spatial parameters is used to denoise the noisy image. To reduce the score deviation, 50 random crops are extracted from denoised image. These scores are averaged to obtain the plots illustrated in Fig. 12. As can be seen, although the same amount of noise is added to each image, maximum quality scores correspond to different denoising parameters in each example. For relatively smooth images such as (a) and (g), optimal spatial parameter of Turbo denoising is higher (which implies stronger smoothing) than the textured image in (j). This is probably due to the relatively high signal-to-noise ratio of (j). In other words, the quality assessment model tends to respect textures and avoid over-

TABLE II: Performance of the proposed method with various architectures in predicting TID2013 quality ratings [2] compared to the state-of-the-art. LCC (linear correlation coefficient) and SRCC (Spearman’s rank correlation coefficient) are computed between predicted and ground truth mean scores (column 2 and 3) and standard deviation of scores (column 4 and 5). EMD measures closeness of the predicted and ground truth rating distributions with  $r = 1$  in Eq. 1.

Model	LCC (mean)	SRCC (mean)	LCC (std.dev)	SRCC (std.dev)	EMD
Moorthy et al. [33]	0.89	0.88	–	–	–
Mittal et al. [34]	0.92	0.89	–	–	–
Saad et al. [35]	0.91	0.88	–	–	–
Kottayil et al. [36]	0.89	0.88	–	–	–
Xu et al. [29]	0.96	0.95	–	–	–
Bianco et al. [7]	0.96	0.96	–	–	–
NIMA(MobileNet)	–	0.782	–	0.209	–
NIMA(VGG16)	0.941	0.944	0.538	0.557	0.054
NIMA(Inception-v2)	0.827	0.750	0.470	0.468	0.064

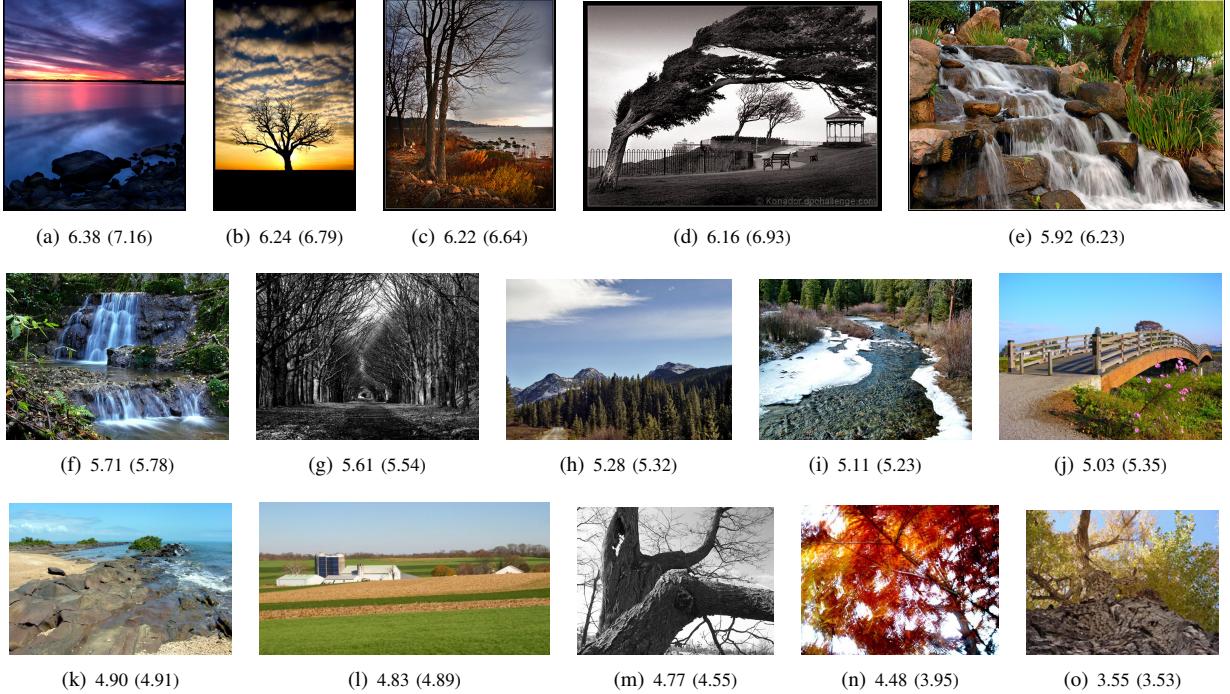


Fig. 7: Ranking some examples labelled with “*landscape*” tag from AVA dataset [1] using our proposed aesthetic assessment model NIMA(VGG16). Predicted (and ground truth) scores are shown below each image.

smoothing of details. Effect of the denoising parameter can be visually inspected in Fig. 13. While the denoised result in Fig. 13 (a) is under-smoothed, (c), (e) and (f) show undesirable over-smoothing effects. The predicted quality scores validate this perceptual observation.

#### D. Computational Costs

Computational complexity of NIMA models are compared in Table III. Our inference TensorFlow implementation is tested on an Intel Xeon CPU @ 3.5 GHz with 32 GB memory and 12 cores, and NVIDIA Quadro K620 GPU. Timings of

one pass of NIMA models on an image of size  $224 \times 224 \times 3$  are reported in Table III. Evidently, NIMA(MobileNet) is significantly lighter and faster than other models. This comes at the expense of a slight performance drop (shown in Table I and Table II).

#### IV. CONCLUSION

In this work we introduced a CNN-based image assessment method, which can be trained on both aesthetic and pixel-level quality datasets. Our models effectively predict the distribution of quality ratings, rather than just the mean scores. This leads

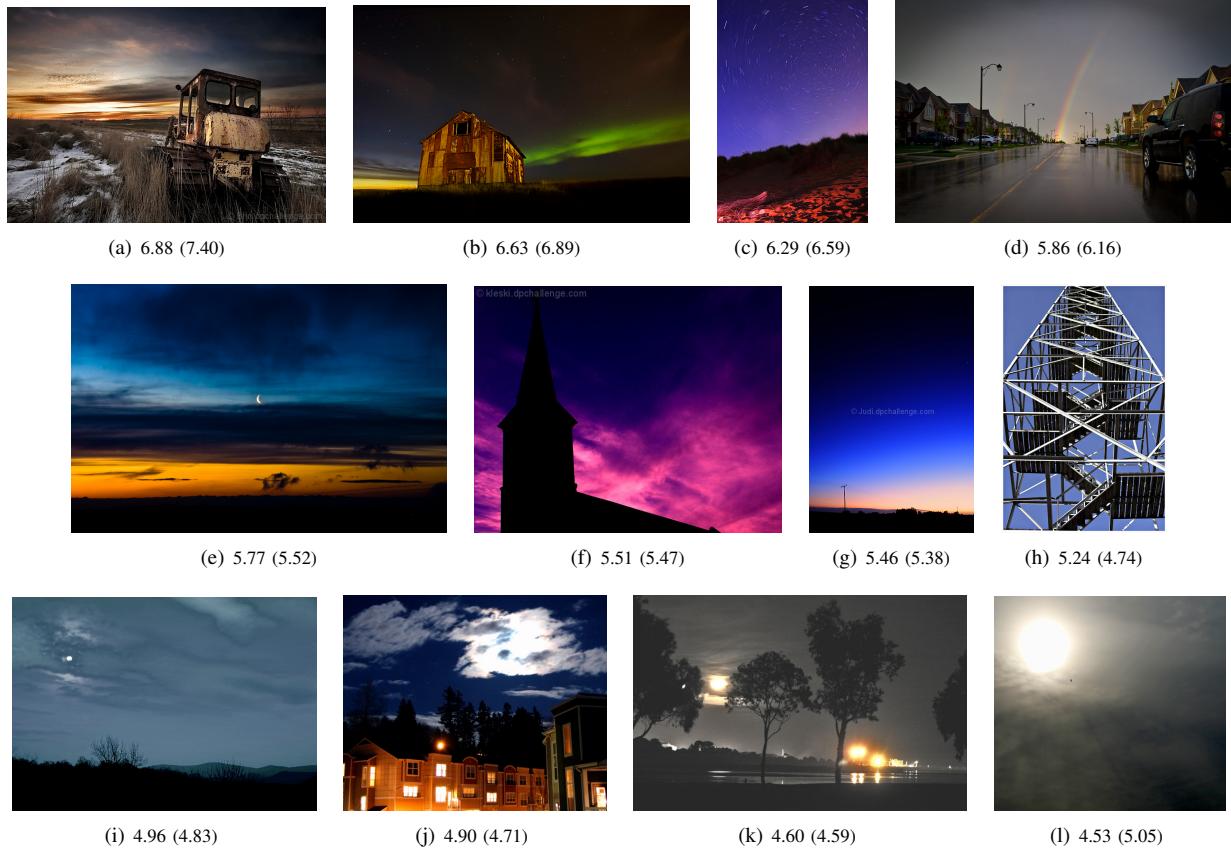


Fig. 8: Ranking some examples labelled with “sky” tag from AVA dataset [1] using our proposed aesthetic assessment model NIMA(Inception-v2). Predicted (and ground truth) scores are shown below each image.

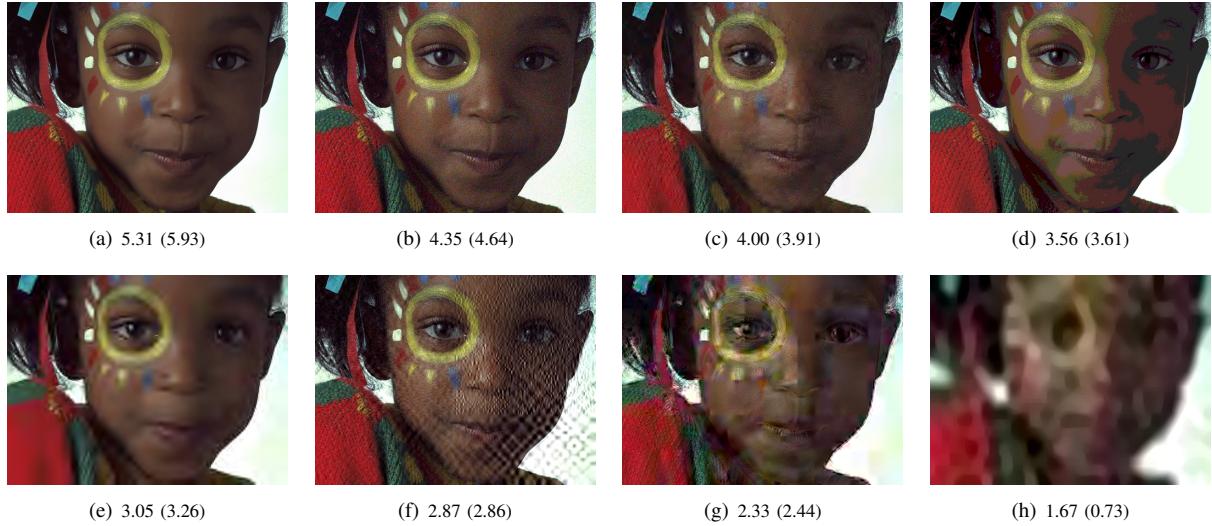


Fig. 9: Ranking some examples from TID2013 dataset [2] using our proposed quality assessment model NIMA(VGG16). Predicted (and ground truth) scores are shown below each image.

to a more accurate quality prediction with higher correlation to the ground truth ratings. We trained two models for high level aesthetics and low level technical qualities, and utilized them

to steer parameters of a few image enhancement operators. Our experiments suggest that these models are capable of guiding denoising and tone enhancement to produce perceptually su-



Fig. 10: Predicted aesthetic score (NIMA(VGG16)) for various parameter settings of multi-layer Laplacian technique [37]. Predicted aesthetic scores are shown below each image.

terior results.

As part of our future work, we will exploit the trained models on other image enhancement applications. Our current experimental setup requires the enhancement operator to be evaluated multiple times. This limits real-time application of the proposed method. One might argue that in case of an enhancement operator with well-defined derivatives, using NIMA as the loss function is a more efficient approach.

#### ACKNOWLEDGMENT

We would like to thank Dr. Pascal Getreuer for valuable discussions and helpful advice on approximation of score distributions.

#### REFERENCES

- [1] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A large-scale database for aesthetic visual analysis,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2408–2415. 1, 2, 6, 7, 8
- [2] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, “Color image database TID2013: Peculiarities and preliminary results,” in *Visual Information Processing (EUVIP), 2013 4th European Workshop on*. IEEE, 2013, pp. 106–111. 1, 3, 4, 7, 8
- [3] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. 1
- [4] W. Xue, L. Zhang, and X. Mou, “Learning without human scores for blind image quality assessment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 995–1002. 1
- [5] L. Kang, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740. 1
- [6] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, “A deep neural network for image quality assessment,” in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3773–3777. 1
- [7] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, “On the use

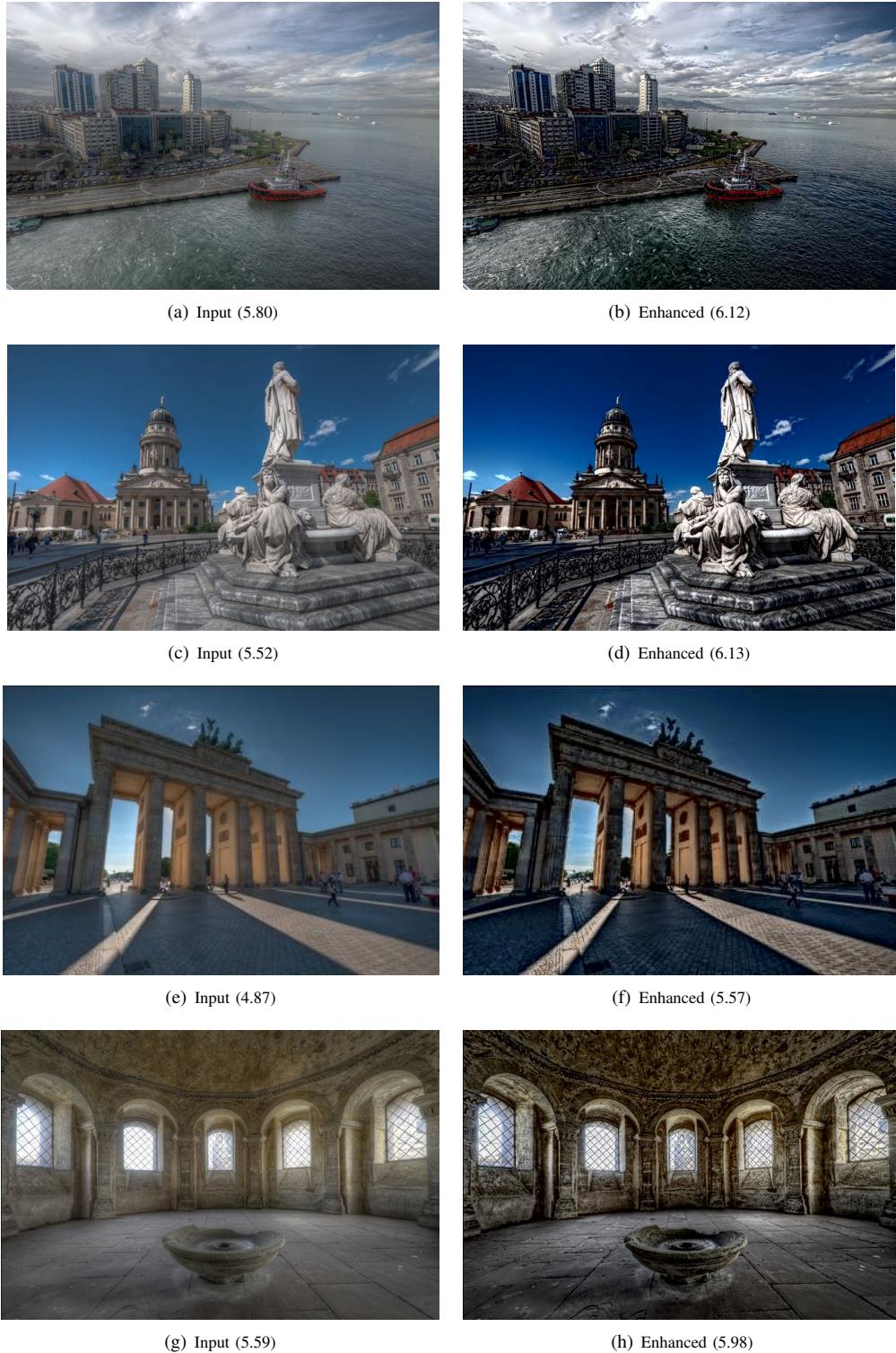


Fig. 11: Tone enhancement by multi-layer Laplacian technique [37] along with our proposed aesthetic assessment model NIMA(VGG16). Predicted aesthetic scores are shown below each image. (Input photos are downloaded from [www.farbspiel-photo.com](http://www.farbspiel-photo.com))

- of deep learning for blind image quality assessment,” *arXiv preprint arXiv:1602.05531*, 2016. 1, 5, 7
- [8] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, “Deep multi-patch aggregation network for image style, aesthetics, and quality estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 990–998. 1

- [9] Y. Kao, C. Wang, and K. Huang, “Visual aesthetic quality assessment with a regression model,” in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1583–1587. 1, 6
- [10] L. Mai, H. Jin, and F. Liu, “Composition-preserving deep photo aesthetics assessment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 497–506. 1, 6

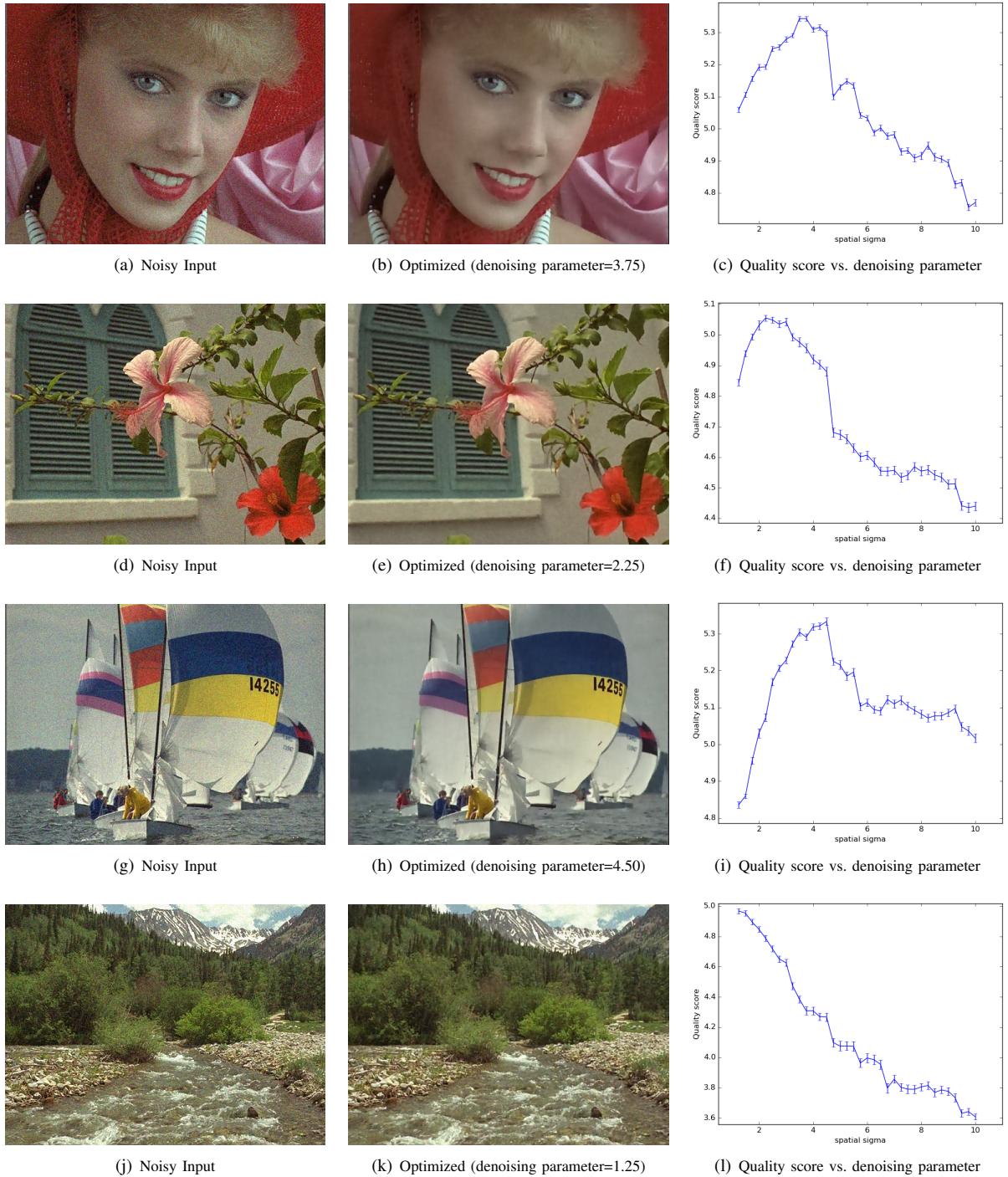


Fig. 12: Tuning spatial parameter of Turbo denoising [38] by using our proposed quality assessment model NIMA(VGG16). Standard deviation of the additive white Gaussian noise is set as 30. Denoised results are shown for maximum quality score.

- [11] B. Jin, M. V. O. Segovia, and S. Süsstrunk, “Image aesthetic predictors based on weighted CNNs,” in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2291–2295. 1
- [12] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, “Live image quality assessment database release 2,” 2005. 1
- [13] E. C. Larson and D. M. Chandler, “Most apparent distortion: full-reference image quality assessment and the role of strategy,” *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006–011006, 2010. 1
- [14] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, “Photo aesthetics ranking network with attributes and content adaptation,” in *European Conference on Computer Vision*. Springer, 2016, pp. 662–679. 1, 5, 6
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 1, 3, 5
- [16] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, “Rating image aesthetics using deep learning,” *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015. 1, 6
- [17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. 1, 3
- [18] S. Ma, J. Liu, and C. W. Chen, “A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment,” in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017. 1, 5, 6
- [19] L. Hou, C.-P. Yu, and D. Samaras, “Squared earth mover’s distance-based loss for training deep neural networks,” *arXiv preprint*

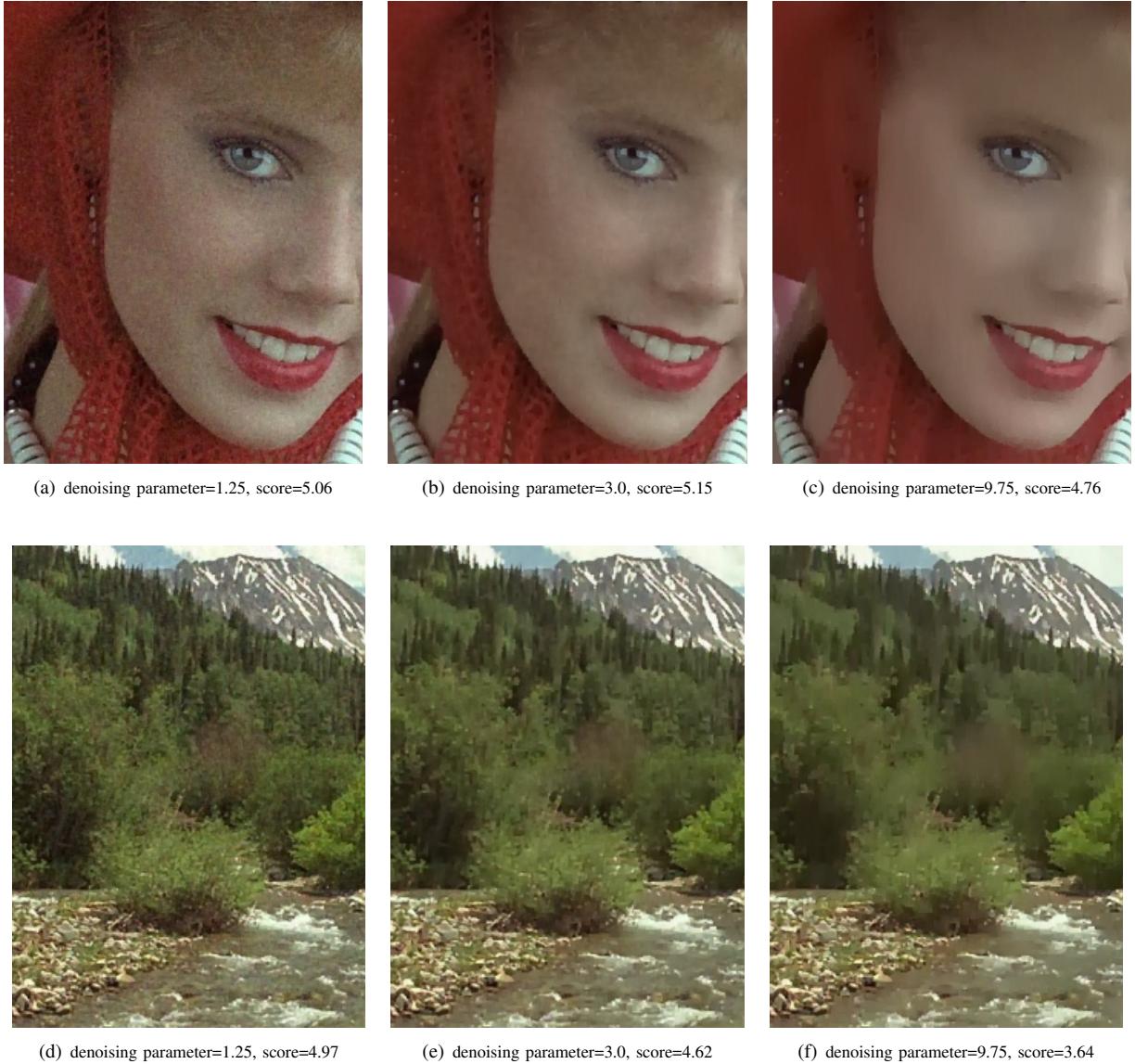


Fig. 13: Effect of Turbo denoising [38] on our predicted quality scores. Input noisy images are shown in Fig. 12.

TABLE III: Comparison of the proposed quality assessment technique with various CNN architectures. Average timings are reported in ms for Xeon Intel CPU @ 3.5 GHz, and NVIDIA Quadro K620 GPU. Timings are reported for applying NIMA models on images of size  $224 \times 224 \times 3$ .

Model	Million Parameters	Billion Flops	CPU Timing (ms)	GPU Timing (ms)
NIMA(MobileNet)	3.22	1.29	30.45	20.23
NIMA(Inception-v2)	10.16	4.37	70.49	39.11
NIMA(VGG16)	134.30	31.62	150.34	85.76

*arXiv:1611.05916*, 2016. 2, 5

- [20] Kodak, “Kodak lossless true color image suite,” <http://r0k.us/graphics/kodak/>. 3
- [21] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012. 3
- [22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the*

*IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826. 3

- [23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017. 3
- [24] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep

- network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, 2015, pp. 448–456. 3
- [25] P. Golik, P. Doetsch, and H. Ney, “Cross-entropy vs. squared error training: a theoretical and experimental comparison,” in *Interspeech*, 2013, pp. 1756–1760. 5
- [26] E. Levina and P. Bickel, “The earth mover’s distance is the Mallows’ distance: Some insights from statistics,” in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 251–256. 5
- [27] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, “TensorFlow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016. 5
- [28] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “TensorFlow: A system for large-scale machine learning,” in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI). Savannah, Georgia, USA*, 2016. 5
- [29] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, “Blind image quality assessment based on high order statistics aggregation,” *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016. 5, 7
- [30] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, “Rapid: Rating pictorial aesthetics using deep learning,” in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 457–466. 6
- [31] Y. Kao, R. He, and K. Huang, “Visual aesthetic quality assessment with multi-task deep learning,” *arXiv preprint arXiv:1604.04970*, 2016. 6
- [32] Z. Wang, S. Chang, F. Dolcos, D. Beck, D. Liu, and T. S. Huang, “Brain-inspired deep networks for image aesthetics assessment,” *arXiv preprint arXiv:1601.04155*, 2016. 6
- [33] A. K. Moorthy and A. C. Bovik, “Blind image quality assessment: From natural scene statistics to perceptual quality,” *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011. 7
- [34] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012. 7
- [35] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind image quality assessment: A natural scene statistics approach in the DCT domain,” *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012. 7
- [36] N. K. Kottayil, I. Cheng, F. Dufaux, and A. Basu, “A color intensity invariant low-level feature optimization framework for image quality assessment,” *Signal, Image and Video Processing*, vol. 10, no. 6, pp. 1169–1176, 2016. 7
- [37] H. Talebi and P. Milanfar, “Fast multilayer Laplacian enhancement,” *IEEE Transactions on Computational Imaging*, vol. 2, no. 4, pp. 496–509, 2016. 6, 9, 10
- [38] T.-S. Wong and P. Milanfar, “Turbo denoising for mobile photographic applications,” in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 988–992. 6, 11, 12
- [39] E. S. Gastal and M. M. Oliveira, “Domain transform for edge-aware image and video processing,” in *ACM Transactions on Graphics (ToG)*, vol. 30, no. 4. ACM, 2011, p. 69. 6
- [40] X. Zhu and P. Milanfar, “A no-reference sharpness metric sensitive to blur and noise,” in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*. IEEE, 2009, pp. 64–69. 6
- [41] ———, “Automatic parameter selection for denoising algorithms using a no-reference measure of image content,” *IEEE Transactions on Image Processing*, vol. 19, no. 12, pp. 3116–3132, 2010. 6