Disregarding the Big Picture: Towards Local Image Quality Assessment

Oliver Wiedemann, Vlad Hosu, Hanhe Lin and Dietmar Saupe
Department of Computer and Information Science, University of Konstanz, Germany
Email: {oliver.wiedemann, vlad.hosu, hanhe.lin, dietmar.saupe}@uni-konstanz.de

***Abstract*—Image quality has been studied almost exclusively as a global image property. It is common practice for IQA databases and metrics to quantify this abstract concept with a single number per image. We propose an approach to blind IQA based on a convolutional neural network (patchnet) that was trained on a novel set of 32,000 individually annotated patches of 64×64 pixel. We use this model to generate spatially small local quality maps of images taken from KonIQ-10k, a large and diverse in-the-wild database of authentically distorted images. We show that our local quality indicator correlates well with global MOS, going beyond the predictive ability of quality related attributes such as sharpness. Averaging of patchnet predictions already outperforms classical approaches to global MOS prediction that were trained to include global image features. We additionally experiment with a generic second-stage aggregation CNN to estimate mean opinion scores. Our latter model performs comparable to the state of the art with a PLCC of 0.81 on KonIQ-10k.**

I. Introduction

Digital images pass through an intricate processing pipeline from being captured to being presented to a human observer. Flaws and limitations of the endpoint devices and performance trade-offs in the algorithms used for transport and storage (e.g. compression) may result in a reduced perceived visual quality. Accurate and generally valid objective image quality assessment (IQA) methods have numerous applications in the multimedia domain since manual inspection is costly and time-consuming. For example, media outlets and graphic design companies can simplify their search for usable source materials by filtering for content of sufficient quality, service providers can measure the performance of their products or mitigate ongoing problems with respect to content quality, etc.

Subjective studies are known to yield reliable opinions for both artificially distorted image datasets [1], [2] where the severity of particular degradations is known as well as for in-the-wild collections of images [3], [4] with authentic and unknown mixtures of distortions. The common benchmark for objective quality measures is their ability to estimate mean opinion scores (MOS) acquired from a sufficient large number of observers [5].

It is possible to distinguish objective IQA methods by their requirements regarding additional information besides the image under assessment. Full-reference methods, such as the PSNR, need access to a pristine original. Reduced-reference algorithms only require partial information, e.g. the type of the predominant distortion in the given image. No-reference image quality assessment (NR-IQA) methods do not require additional information.

In this paper, we introduce an approach to local NR-IQA that applies to the wide range of distortion present on images in-the-wild. Quality is generally considered as a property of the entire image, evaluated via the MOS of a group of observers. This is the point of view that previous IQA methods have taken. Some works consider that each part of the image contributes independently [6] to the overall quality score, whereas others assign different weights [7] to build a better global quality estimate.

We hypothesize that quality can be understood as both a local property of an image patch of a sufficiently large size as well as a property of an entire image. In our IQA approach, we intend to rely on the assessed quality of individual patches. To this end, we created a novel dataset of manually quality-annotated RGB patches sampled from *KonIQ-10k* [4]. We build a local patch-level quality prediction CNN architecture and train it on our patch dataset. As far as we are aware of, we are the first to consider to directly predict the quality of individual patches, without making any indirect assumptions about the correspondence between the global and local quality scores.

We expect our predictor to be more representative of the low level technical aspects of quality, without having been influenced by content or other higher level factors, such as aesthetics or image composition. For comparison reasons, we also include two approaches to global MOS prediction: Patchnet is used in a sliding-window fashion to create spatially small quality maps of authentically distorted images taken from *KonIQ-10k*.

The average value of these maps already correlates highly with the global MOS score. Furthermore, we augment our quality maps with two other local low-level indicators, namely the FISH sharpness metric [8] and brightness information in the form of gray-scale version of the original input. We then study the performance of a generic feature aggregator based on a DenseNet-169 CNN [9].

Our results show that the correlations between the mean values of `patchnet` quality maps and global MOS values on *KonIQ-10k* are already comparable to the best-performing global statistical methods that were fine-tuned on the respective dataset. Aggregation of all three of our spatially small feature maps by a second-stage CNN outperforms all classical methods and the naive patch-based deep learning methods. We expected this approach to be falling short of the global performance of models that traded incorporating additional information (e.g. content) for the ability to predict local quality

on small areas.

II. Related Work

A popular approach in traditional NR-IQA utilizes Natural Scene Statistics (NSS) to estimate the perceptual quality of images. The NSS assumption is that distortions in natural images can be measured by deviations of feature distributions between those observed in distorted images and those on pristine images. NSS is used in multiple works, for instance, Moorthy et al. [10] extract a large number of wavelet coefficients and utilize a two staged support vector classification and regression model to predict quality. Saad et al. [11] fit a generalized Gaussian model on block-wise extracted DCT coefficients and predict quality using a Bayesian model applied to the estimated distribution parameters. Further NSS based methods assess contrast distortions [12] and involve aesthetics measures [13] to augment existing NR-IQA methods. Recent advances in machine learning led to the proposal of fully data-driven IQA regressors. One of the cornerstone implementations was proposed by Kang et. al [14]. It consists of only one convolutional layer with 50 feature maps that are subsequently min and max pooled and fed into a fully-connected network for quality score estimation. Bosse et. al [6] use a streamlined neural network with 10 convolutional layers and a fully-connected predictor at the end. They train their network to estimate MOS values on 32×32 pixel RGB patches sampled at random locations of images taken from the LIVE dataset. Prediction of global scores is implemented by averaging a sufficient number of local scores. Although simple, their proposed aggregation strategy leaves room for improvement. For instance, the overall quality distribution of an image may correlate with the aesthetic sentiment, e.g. image composition following the golden ratio. Bianco et. al [15] report on multiple experiments with different pre-trained network architectures working directly on large image patches. DeepBIQ, their best performing method, uses a CNN to extract local features that are fed into a support vector regressor. The state of the art for global MOS prediction, as set by traditional methods, has been substantially enhanced by deep-learning methods on all commonly used quality benchmark databases such as *LIVE* [1], *LIVE in the Wild* [3] or *TID2013* [16].

III. Approach

The implicit assumption of a uniform quality distribution introduced by measuring quality with a single MOS value per image does not necessarily hold in natural images. For example, background areas are often blurred or underexposed in comparison to the object in focus. Naive training of a patch-based IQA model on such a database is therefore prone to difficulties in capturing quality-indicative areas of an image.

We propose a new method to create training data in an attempt to attenuate the issues of varying quality in natural images. To estimate global MOS more precisely while avoiding to rely on higher-level image content, we augment the quality maps generated by grid-wise application of `patchnet` with maps generated by applying the FISH sharpness metric in the same spatial fashion. As a third layer, we add a downsized gray-scale version of the original input image to convey brightness information to the global predictor.

IV. Patchnet

The `patchnet` model takes RGB images of a fixed size of 64×64 pixel as an input and produces a quality score in $[0, 1]$. The neural network has seven convolutional layers followed by three fully connected layers. Its architecture is depicted in Fig 1. Our contribution lies not so much in the CNN itself, but in the novel way of creating training data and applying it for IQA.

A. Database Creation

We randomly sampled 500 pictures from *KonIQ-10k* as a basis for patch selection. Those were excluded from the remaining dataset to guarantee that MOS estimation experiments are never carried out on images which `patchnet` was partially trained on. From each of the selected images, we sampled 64 patches at random locations. This set of $32,000$ patches was manually annotated in an attempt to flag patches that were *not* indicative of being sampled from a high quality image.

Instead of having the patches rated on an absolute category rating (ACR) scale as usually done for entire images, we simplified the task to a binary classification. In a total of three iterations, a student labeled the data in a controlled lab environment. First, patches were presented one at a time and received an initial rating. In the second and third iterations, the patches were shown in a grid. We displayed only patches of either high or low quality at a time to provide context about the other patches' quality in the same set. This was intended to allow amending the dataset in a consistent way, as each patch could be directly compared to a large number of elements from the same set and relabeled in case the opposing set was a contextually more reasonable choice.

Examples of patches marked as being indicative of high quality are given in Fig. 2, patches marked as being not indicative of high quality are presented in Fig. 3.

B. Training the Network

Each of the 500 source images, that together yielded the 32,000 annotated patches, has an ACR derived mean opinion score $mos(\mathcal{I})$, whereas each patch p has been labeled as either low or high quality. We proceeded under the assumption that the quality of patch p sampled from image $\mathcal{I}$ can not exceed the quality of the image it was sampled from. We therefore clamped the quality scores using the following relationship to generate our final scores:

$$\mathcal{S}(p) = \begin{cases} mos(\mathcal{I}) & \text{if } p \text{ was marked as high quality} \\ 0 & \text{otherwise} \end{cases}$$

Each patch was presented to the network in all 4 orientations and also mirrored horizontally and vertically. Our augmented dataset thus consists of a total of 384,000 patches. All convolutional kernels in `patchnet` have a spatial dimension of

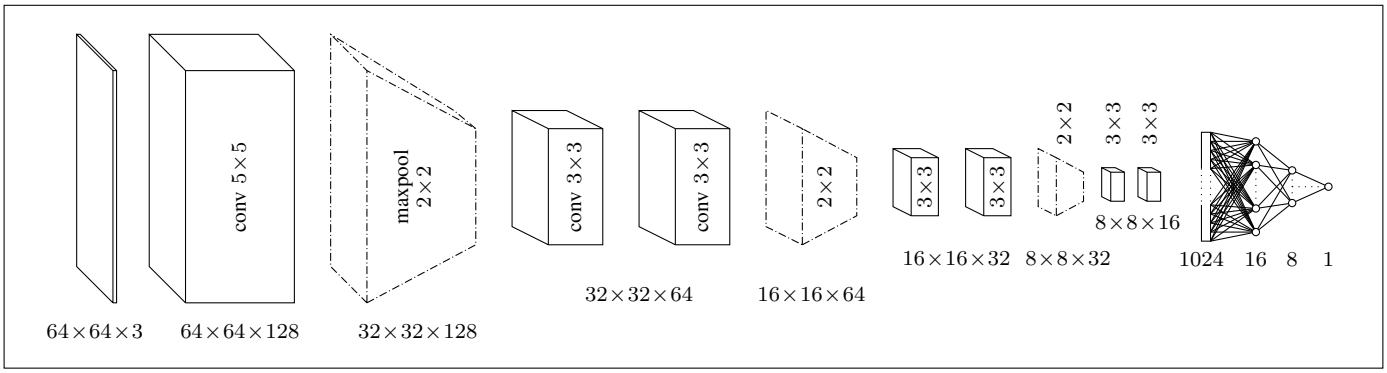

Fig. 1. Architecture of `patchnet`.

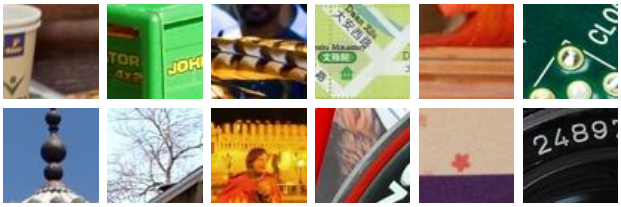

Fig. 2. Patches marked as being indicative of high quality.

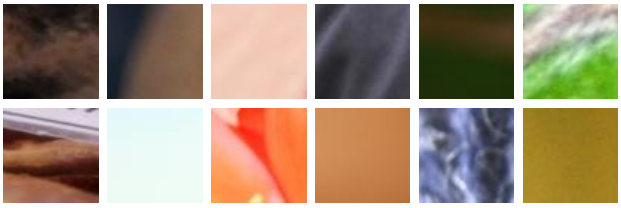

Fig. 3. Patches marked as being not indicative of high quality.

3×3 pixel, except for the first layer, where kernels of 5×5 pixel were used. Max pooling was always performed with a kernel size of 2×2 pixel. The output of the last convolutional layer is flattened into a 1024-dimensional vector which is passed through a fully connected network with 16, 8 and finally 1 node. All nonlinearities were rectified linear units, except for the output node which employs a sigmoid function. We trained a Keras [17] implementation of `patchnet` on a Nvidia K40 GPU in a little less than 12 hours. We used an Adam [18] optimizer and a batch size of 64 patches per iteration.

V. Indicator Map Generation

In accordance with our dual understanding of image quality, we experimented with two second stage models to estimate mean opinion scores of images solely based on local quality information. It is nontrivial to separate subjectively perceived quality from aesthetic aspects and personal or even cultural preferences for specific content when judging an entire natural image.

For a given image of width w and height h, we created three local indicator maps of width $\tilde{w} = \lfloor (w-\Delta)/\delta \rfloor + 1$ and height $\tilde{h} = \lfloor (w-\Delta)/\delta \rfloor + 1$ where $\Delta = 64$ is the edge length of our patches and $\delta = 4$ is the sub-sampling stride in both horizontal and vertical directions, as shown in Fig. 4. The three indicator maps are a quality map generated by `patchnet`, a sharpness map generated by FISH [8] and a brightness map that is generated by down-scaling a gray-scale version of the input image.

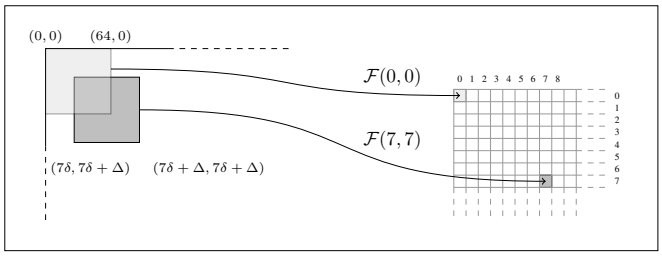

Fig. 4. Subsampling with $\Delta = 64$ pixel and $\delta = 4$ pixel.

VI. Indicator Map Aggregation

Recent publications in the image classification community propose very deep convolutional neural networks with shortcut connections as a mitigation for the common problems of gradient decay and overfitting [19], [20]. We chose a headless DenseNet-169 [9] architecture that was pretrained on the ImageNet ILSVRC dataset [21] as a basis for MOS regression. The final scores are generated from the output feature maps by applying a single 1×1 kernel followed by global max pooling. An overview of the structure is given in Figure 5.

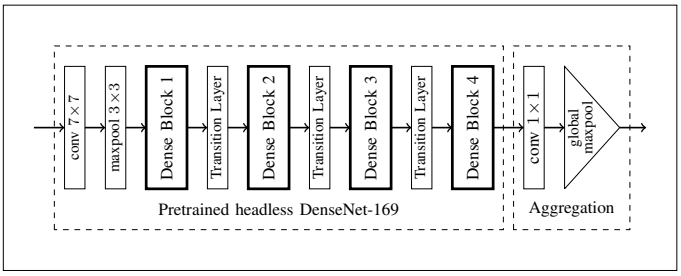

Fig. 5. Coarse architecture of the aggregation network.

Within a Dense Block, the results of a convolutional layer are directly passed to each of the following convolutional

layers. In comparison to ResNets [19], the maps are not added component-wise, but concatenated. Each block in DenseNet-169 consists of sub-blocks of 1×1 convolutions, followed by 3×3 convolutions, with 32 kernels each. There are 6, 12, 32 and 32 of these sub-blocks in the model we utilize. To further validate our approach, we split the remaining $9,500$ unused images from *KonIQ-10k* into training, validation and test sets, according to the usual $60/20/20$ split. The training data set was artificially augmented as follows: Each of the $5,700$ images was taken once unmodified and in three randomly rotated versions between $+10°$ and $-10°$. After rotating, we cropped the image to a rectangle of maximal valid size, cutting off any undefined regions that were introduced by the rotation. All training images were flipped both horizontally and vertically. The resulting set of $68,400$ feature maps was binned by resolution into 82 sets that were fed to the optimizer iteratively. Each one was used for four consecutive optimization steps to reduce the impact of the hardware bottleneck of transferring data to the GPUs. We left the validation and test sets, containing $1,900$ images each, unmodified to avoid any influence rotating and flipping may have on the perceived visual quality. We trained the model using the same optimizer and batch-size as for `patchnet`.

VII. Experimental Results

A. Averaging Patchnet Quality Maps

As shown in Table I, the baseline aggregation strategy of taking the average of `patchnet` predictions already produces results that correlate highly with global MOS scores. Note that training was done solely on patches with individually clamped mos scores. Those were sampled from a disjoint set of images that were excluded from these experiments.

Relying solely on this rudimentary information, it was unexpected that the correlation with the global MOS (0.67 SROCC) on the validation database is very close to that of finely crafted features such as the best performing traditional IQA method BRISQUE [22] (0.7). Moreover, the evaluation is done at a disadvantage to our method, with BRISQUE having been trained and tested on the same scoring database utilizing features that aggregate global image characteristics.

TABLE I
Correlation coefficients between average feature map values and MOS

Database	Measure	FISH	patchnet	gray-scale
KonIQ-10k	SROCC	0.560	0.667	0.342
	PLCC	0.513	0.573	0.354
Live in the Wild	SROCC	0.500	0.512	0.213
	PLCC	0.503	0.527	0.206

B. DenseNet Aggregation

We trained our model on the feature maps generated from *KonIQ-10k* and evaluated its performance on a test set of 1900 pristine images. We have analyzed a variety of existing methods on both of the largest existing annotated databases *KonIQ-10k* and *LIVE In the Wild*. This enables to position the performance of our local features in the broader context of global IQA prediction. Two of the more recent global methods [15], [23], having been trained and tested on the same *KonIQ-10k* database, clearly out-perform our approach. This is not surprising, considering the heavy restrictions we impose on our model inputs. We however achieve close to state of the art performance on *KonIQ-10k* and outperform all of the classical and the naive patch-based models.

TABLE II
Correlation coefficients for recent IQA methods.

	KonIQ-10k		LIVE In the Wild	
	SROCC	PLCC	SROCC	PLCC
BIQI [24]	0.54	0.61	0.29	0.38
BLIINDS-II [11]	0.57	0.58	0.44	0.48
BRISQUE [22]	0.70	0.70	0.59	0.63
DIIVINE [10]	0.58	0.62	0.43	0.46
SSEQ [25]	0.59	0.61	0.45	0.50
BosICIP [6]	0.65	0.67	0.70	0.70
KangCNN [14]	0.63	0.67	0.71	0.73
DeepBIQ [15]	0.90	0.92	0.89	0.91
DeepRN [23]	0.92	0.95	0.91	0.93
Our Model	0.79	0.81	0.60	0.62

The correlation coefficients for *KonIQ-10k* are taken from [23] for BosICIP, KangCNN, DeepBIQ, and DeepRN and the remaining methods are implemented by ourselves.

VIII. Conclusion

We have only broken the ice on this research track towards a better understanding of local, low-level technical quality. We proposed a new approach to train a quality indicator working on small image regions on a novel dataset of individually annotated patches. `Patchnet` was shown to correlate highly with mean opinion scores on a large and authentic in-the-wild dataset of natural images. The amount of information presented to the second-stage global MOS predictor is significantly smaller in comparison to other existing methods. The spatial size of the feature maps is roughly 6.25% of the input image. Still, the performance of our model is close to the state of the art and suggests that stacking feature maps is a promising direction in aggregating local quality indicators. Deliberately trying to disregard higher-level features to avoid influences based on content and aesthetic aspects found in images is a new direction within the domain of IQA and against the general trend of just applying deeper networks on larger databases. Future work includes extending the database on which we trained `patchnet` both regarding the number of patches as well as the number of votes on each patch to remove potential bias. As an alternative alley, one could choose to combine the approach of a content-aware global predictor such as DeepRN [23] with a larger, non-subsampled quality map generated by our local indicator. We expect benefits regarding the required training time if a local precursor already indicates regions of high quality, which should help a content-aware model to identify relevant concepts.

IX. Example Images

Here, we list examples of images and their feature map representations to convey a feel for the indicators' behaviour in different scenarios.

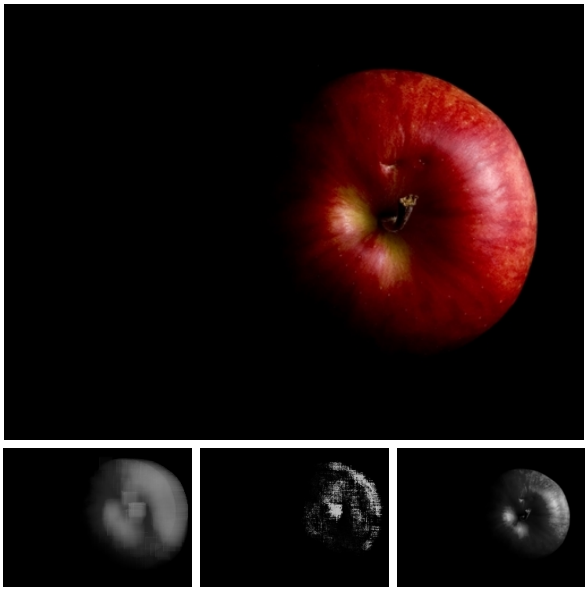

Fig. 6. Image with highest SROCC (0.97) between FISH and `patchnet` maps from KIQ10k. The original image on top with the FISH, `patchnet` and gray-scale maps below (from left to right).

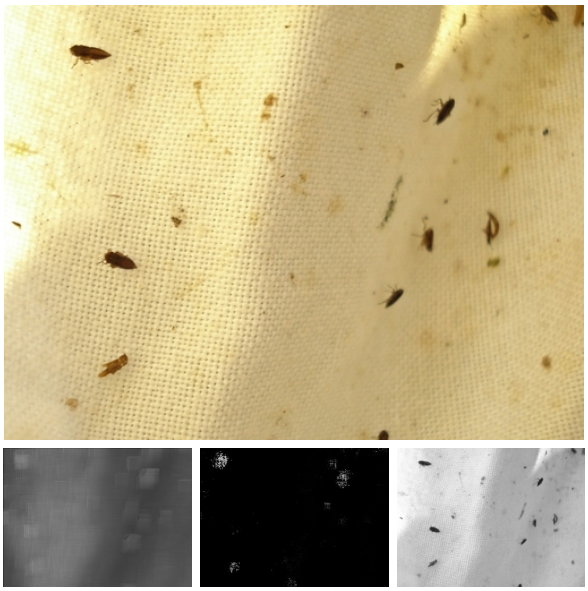

Fig. 7. Image with lowest SROCC (-0.55) between FISH and `patchnet` maps from KIQ10k. The original image on top with the FISH, `patchnet` and gray-scale maps below (from left to right).

Acknowledgment

We thank the German Research Foundation (DFG) for financial support within project A05 of SFB/Transregio 161.

Fig. 8. Example image contrasting the difference in range-normalized responsiveness between FISH and `patchnet`: Mind the clearer distinction of foreground and background between FISH (left) and `patchnet` (middle).

References

[1] H. R. Sheikh, "Live image quality assessment database," *http://live.ece.utexas.edu/research/quality*, 2003.

[2] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[3] D. Ghadiyaram and A. Bovik, "Live in the wild image quality challenge database," *http://live.ece.utexas.edu/research/ChallengeDB/index.html*, 2015.

[4] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10K: Towards an ecologically valid and large-scale IQA database," *arXiv preprint arXiv:1803.08489*, 2018.

[5] Z. Wang and A. C. Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006.

[6] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3773–3777.

[7] 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. [Online]. Available: http://www.cv-foundation.org/openaccess/content_iccv_2015/html/Lu_Deep_Multi-Patch_Aggregation_ICCV_2015_paper.html

[8] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 423–426, 2012.

[9] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[10] 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.

[11] 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.

[12] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 838–842, 2015.

[13] M. Jenadeleh and M. E. Moghaddam, "Blind image quality assessment through wakeby statistics model," in *International Conference Image Analysis and Recognition*. Springer, 2015, pp. 14–21.

[14] 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.

[15] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *arXiv preprint arXiv:1602.05531*, 2016.

[16] 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.

[17] F. Chollet *et al.*, "Keras," 2015.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[20] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[22] 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.

[23] D. Varga, T. Szirányi, and D. Saupe, "DeepRN: A content preserving deep architecture for blind image quality assessment," in *Multimedia and Expo (ICME), 2018 IEEE International Conference on*. IEEE, 2018.

[24] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal processing letters*, vol. 17, no. 5, pp. 513–516, 2010.

[25] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.