Neural Network-Based Full-Reference Image Quality Assessment

Sebastian Bosse*, Dominique Maniry*, Klaus-Robert Müller†‡, *Member, IEEE*,
Thomas Wiegand*†, *Fellow, IEEE*, and Wojciech Samek*, *Member, IEEE*
*Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
‡Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea
†Berlin Institute of Technology, 10587 Berlin, Germany

Abstract—**This paper presents a full-reference (FR) image quality assessment (IQA) method based on a deep convolutional neural network (CNN). The CNN extracts features from distorted and reference image patches and estimates the perceived quality of the distorted ones by combining and regressing the feature vectors using two fully connected layers. The CNN consists of 12 convolution and max-pooling layers; activation is done by a rectifier activation function (ReLU). The overall IQA score is computed by aggregating the patch quality estimates. Three different feature combination methods and two aggregation approaches are proposed and evaluated in this paper. Experiments are performed on the LIVE and TID2013 databases. On both databases linear Pearson correlations superior to state-of-the-art IQA methods are achieved.**

I. Introduction

Images and videos are ubiquitous today. The share of bits representing visual signals is huge and even growing. According to Cisco Visual Networking Index [1], by 2017 80-90% of global internet traffic will be internet video traffic or shared peer-to-peer video. To bring this tremendous amount of visual data to consumers, images and videos need to be compressed in order to allow for the transmission over band-limited channels. With decreasing bit rate, distortions are introduced into the transmitted signal that become visible for the human eye. In order to automatically evaluate and/or optimize the performance of transmission systems or modules of these in terms of rate-distortion costs, a metric for image or video quality is necessary. As humans are typically the ultimate receiver of visual signals, it is crucial for such a metric to relate to human visual perception and to predict the visual distortion perceived by humans reliably. Commonly, image quality metrics (IQM) are categorized by the amount of information used for estimating perceived quality. While full reference (FR) IQM have access to the complete undistorted source reference image and its distorted version, no reference (NR) IQM take only the distorted image as an input in order to estimate the perceived quality of it. Reduced reference (RR) IQM [2][3] lives in the middle of this spectrum as only a set features extracted from the source reference image is given as input to the algorithm.

At sender or encoder side of a transmission system the reference is typically available, which allows for the use of FR IQM. The two most simple FR IQM are mean square error (MSE) and the logarithmically related peak-signal-to-noise-ratio (PSNR), both of which poorly correlate with perceived visual quality [4]. Sophisticated FR IQM typically follow one of two strategies [5]: *Bottom-up* approaches aim at modeling various processing mechanisms of the human visuals system (HVS), such as masking effects [6], contrast sensitivity [7] or just-noticeable-distortion [8], [9]. *Top-bottom* approaches apply assumptions on the general functions of the HVS and try to identify and exploit corresponding features from images in order to estimate perceived quality. Here the structural similarity (SSIM) [10], feature-similarity (FSIM) [11], gradient magnitude similarity deviation (GMSD) [12] and Haar wavelet-based perceptual similarity index (HaarPSI) [13] are examples for this class of approaches. The structural similarity index exploits the sensitivity of the HVS to changes in local structures in order to predict perceived quality. The feature similarity index combines local phase coherency and local gradient magnitude and uses the differences in these spatially local features to predict perceived quality. DOG-SSIM [14] applies SSIM to difference of gaussian filtered images and by this mimics the HVS more explicitly. Recently, methods applying a third strategy have been proposed. These methods operate purely data driven and do not rely on explicit assumptions about the HVS or perceptual image features. Data driven approaches proposed so far mainly deal with the problem of NR IQA. In [15], luminance normalized image patches are k-means clustered in order to learn a general image representation. The distance between the luminance normalized test image patches are soft-thresholded. The resulting code coefficients are pooled and then regressed by a support vector machine (SVM) to predict image quality. A linear SVM regression is used to learn a set of linear image features for NR IQA in [16]. An extension of this approach is presented in [17]. Here, object-like patches are detected in a first step. The identified patches are then given as an input to the method in [16]. In [18], a five layer convolutional neural network (CNN) is trained to jointly learn feature to be extracted and a regression function applied to these features to estimate the quality of luminance normalized image patches. The first layer consists of 50 convolutional kernels used for feature extraction. The extracted feature maps are pooled to one minimum and one maximum feature map. These two feature maps are then given as an input to two fully connected layers in order to learn the regression. The quality of full images is then predicted by averaging the patchwise estimation.

This paper applies the data driven CNN framework to the domain of FR IQA. A 12-layer CNN is applied to the image patches from source reference and distorted images in parallel. The features extracted from the image patches by the CNNs are

SB and DM contributed equally.

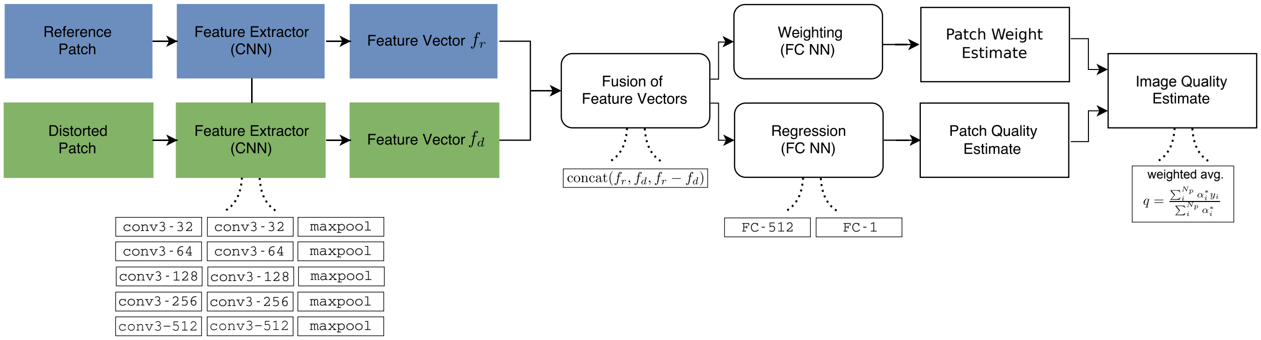

Fig. 1: Layout of the proposed neural network.

fused and fed into a 2-layer fully connected network to learn the regression function for patch-wise quality estimation. This architecture was used in [19] for NR IQA and preliminary results for FR IQA have been presented in [20]. In this work we provide further evaluations on FR IQA and compare two different methods for aggregating local patch quality to global image quality.

This paper is organized as follows. In the next section we present the neural network based FR IQA method and describe three feature combination strategies and two patch aggregation methods. In Section III-A we evaluate our method on the popular LIVE [21] and TID2013 [22] image quality databases. We conclude in Section IV with a brief discussion.

II. Proposed Method

A. Network Layout

As neural networks commonly take data of fixed size as an input, we apply the proposed system to individual unpreprocessed 32×32 RGB patches cropped from the reference and the distorted images. An imagewise estimate of perceived quality can then be obtained by suitable pooling of the patchwise estimated quality, e.g. by weighted or unweighted averaging. Thus, training and testing is performed patchwise. For training, patches are annotated with the quality labels given to the full image that the respective patch was cropped from. A high level layout of the proposed method of patchwise estimation of perceived quality can be subdivided in four modules, as shown in Fig. 1.

In a first step, features are extracted from reference and distorted image patches, respectively, by two identical convolutional neural networks (CNN) that are not interconnected with each other. For feature extraction, we choose a architecture with many layers inspired by [23], as this network achieved very good results in the ILSVRC image classification challenge [24]. Thus, only 3×3 convolution kernels are used, activation is done by a rectifier activation function (ReLU) and feature maps are reduced in size only by max-pooling. Referring to the notation from [23], the 12 weight layers in our architecture are organized as: conv3-32, conv3-32, maxpool, conv3-64, conv3-64, maxpool, conv3-128, conv3-128, maxpool, conv3-256, conv3-256, maxpool, conv3-512, conv3-512, maxpool. In order to ensure the output of the convolution to be of the same size as the input, all convolutions are applied with zero-padding. Maxpool layers have 2×2 kernels. Dropout regularization [25] is applied to the FC layers with dropout ratios 0.3, 0.4, 0.4 and 0.3.

The two feature vectors f_r and f_d that are extracted from the reference image and the distorted image by the CNNs are fused in a second step. In this paper three different fusion strategies are discussed:

(1) subtracting f_r and f_d
(2) concatenating f_r and f_d
(3) concatenating f_r, f_d and $f_r - f_d$

After feature fusion, in the third step the fused feature vector is input to a fully connected neural network (FC-512, FC-1) regressing it to a patch quality estimate.

The fourth step aggregates the patch quality estimates to an image quality estimate.

B. Training and Optimization

A given image can be subdivided in N_p patches and annotated with a ground truth quality label q_t collected in subjective tests. The quality q_t of an image is estimated as q and can be calculated by averaging the patchwise quality estimate y_i, output of the neural network: $q = \frac{1}{N_p} \sum_i^{N_p} y_i$. Training is performed by minimizing the mean absolute error (MAE)

$$E_{patchwise} = \frac{1}{N_p} \sum_i^{N_p} |y_i - q_t|$$

MAE puts less emphasis on outlier than mean squared error (MSE) [18]. The optimization is done through the adaptive learning rate optimizer ADAM [26] with $\alpha = 0.0001$.

C. Weighted Patch Aggregation

Image quality ratings given by humans are related to the perceived quality of a full image. Typically, (perceived) distortions are not equally distributed spatially over an image, e.g. due to masking of band-limited noise in textured regions. Moreover, distortions in salient image regions have a stronger influence on global quality than distortions in less salient image regions. Thus, the quality label assigned to a full image does not necessarily reflect the locally perceived quality.

In order to account for the influence of local image and noise properties on the quality of full images, we propose a *weighted average aggregation* of patchwise estimated local quality to global quality. To achieve this, two additional fully connected layers are added to the network that run in parallel to the last two layers (the regression part) of the network proposed in Subsec. II-A and are of the same shape. The output α_i of these layers can be used to weight the estimated local quality of the corresponding patch i. Activating the weight by a ReLU and using a small stability term ϵ ensures it to be positive and non-zero with

$$\alpha_i^* = max(0, \alpha_i) + \epsilon$$

The global quality of a full image can then be calculated as

$$q = \frac{\sum_i^{N_p} \alpha_i^* y_i}{\sum_i^{N_p} \alpha_i^*} \tag{1}$$

.

For end-to-end training, the error of the globally estimated quality

$$E_{weighted} \;=\; |q - q_t| \tag{2}$$

of each image is minimized. In the following we will refer to this patch aggregation strategy as "weighted avg". The simple (non-weighted) averaging will be referred to as "average".

III. Experiments

A. Experimental Setup

The proposed method is evaluated on the LIVE [21] database consisting of quality annotated images that are subject to distortions of different kinds and varying levels. The LIVE database [21] is based on 29 source reference images, subject to 5 different types of distortions at three to five different distortion levels. MOS values were obtained under fairly controlled conditions. The TID2013 [22] comprises 25 colored reference images and 3000 differently distorted images, subject to 24 different distortion types. Subjective ratings were gathered by comparisons. The results from several viewing conditions of experiments in three different labs and on the internet were averaged.

We evaluated the proposed method in terms of prediction accuracy and prediction monotonicity. Prediction accuracy is measured as Pearson linear correlation coefficient (LCC) and prediction monotonicity is measured as Spearman rank order correlation coefficient (SRCC). For evaluating the performance of the proposed FR IQA method, the CNN is trained on 10 random train-test splits: For testing, 6 source reference images and corresponding distorted versions were randomly chosen, 6 of the remaining source reference images and corresponding versions were randomly chosen for validation and the remaining (13 for LIVE, 14 for TID2013) were used for training. In each epoch, 32 random patches are sampled from each image from the training set. Models are trained for 3000 epochs. In our implementation, training takes about 11s per epoch on a Titan X GPU.

B. Results

Table I summarizes the performance the different proposed feature fusion schemes for LIVE and TID2013 databases. The table shows that the relationship between the two feature vectors can be learned by the model, but providing the difference $f_r - f_d$ explicitly leads to better results on both datasets. Simple difference fusion $f_r - f_d$ has the advantage that it becomes zero if the feature vectors coincide (i.e., distorted image equals reference image), but it lacks flexibility, e.g. when only one of the feature vectors is informative. Due to the limited size of the training data set we did not evaluate more complex fusion techniques. However, although the differences in performance are rather marginal for both databases and aggregation methods, for further analysis the proposed method is evaluated based on $\texttt{concat}(f_r, f_d, f_d - f_r)$.

Dataset	Aggregation	$f_d - f_r$	$\texttt{concat}$ (f_r, f_d)	$\texttt{concat}$ $(f_r, f_d, f_d - f_r)$
LIVE	Average	0.976	0.974	0.976
	Weighted avg	0.982	0.977	0.982
TID2013	Average	0.908	0.893	0.908
	Weighted Avg	0.962	0.958	0.965

TABLE I: Comparison of performance of the two suggested patch aggregation methods. The LCC was computed on the validation set of one random split for each dataset and with $N_p = 1024$ random patches per image.

Database	TID2013		LIVE	
Method	LCC	SROCC	LCC	SROCC
PSNR	0.675	0.687	0.856	0.866
SSIM[10]	0.790	0.742	0.906	0.913
FSIM$_C$[11]	0.877	0.851	0.961	0.965
DOG-SSIM[14]	0.919	0.907	0.963	0.961
Average (proposed)	0.880	0.859	0.977	0.966
Weighted Avg (proposed)	**0.946**	**0.940**	**0.980**	**0.970**

TABLE II: Comparison of different FR IQA methods based on the TID2013 and LIVE database. Highest LCC and SROCC are set in bold. The reported correlations of the proposed method are the average correlation achieved on the test sets of 10 random train-test splits.

Table II compares the proposed methods with three popular IQMs, namely PSNR, SSIM[1], FSIM and DOG-SSIM. Regardless of the patch aggregation strategy, the proposed method achieves higher prediction accuracy and prediction monotonicity on the LIVE database. On TID2013, DOG-SSIM achieves higher LCC and SROCC than the proposed method applying average patch aggregation, but is outperformed by the proposed method if weighted-average patch aggregation is used.

In the left panel of Fig. 2 the estimated patchwise quality estimates y_i are scattered against the patchwise ground truth q_t. The patchwise weight α_i assigned to the specific patch by the neural network in order to calculate the imagewise quality is indicated by color. Some of the patches are assigned with a negative quality estimate. These patches are consistently

[1] Please note that we refer to the correlations of SSIM as reported in [18]. Slightly different values are reported e.g. in [11].

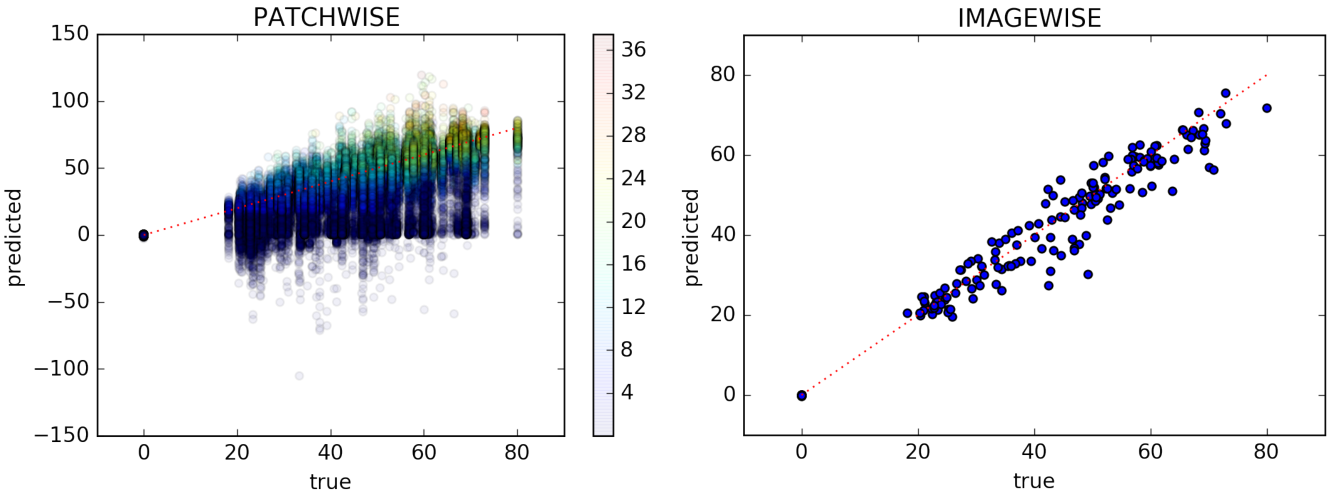

Fig. 2: Scatter plot of patchwise (left) and imagewise (right) scores for the $\texttt{concat}(f_r, f_d, f_r - f_d)$ method with weighted average patch aggregation. The color represents the weight α^* assigned to a specific patch in the left plot.

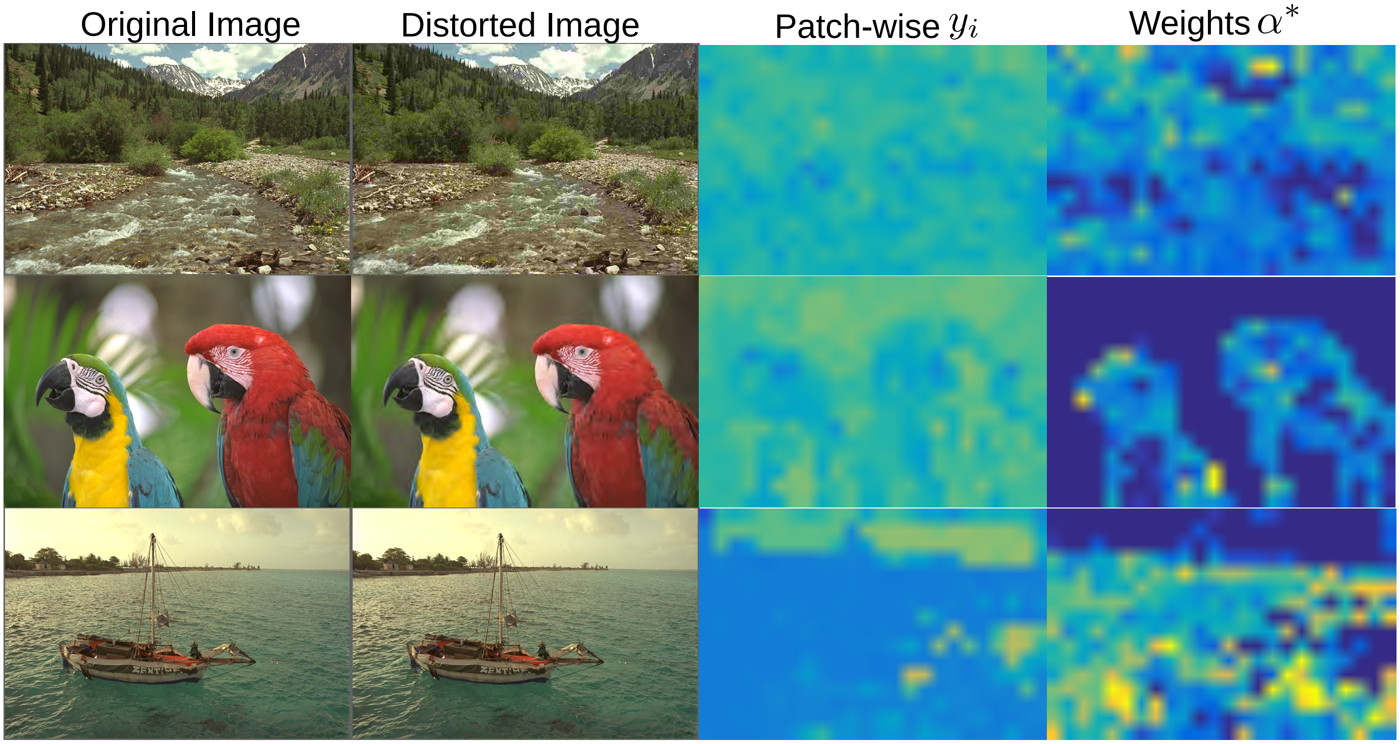

Fig. 3: Blockwise spatial distribution of patchwise predictions y_i and weights α^* for the $\texttt{concat}(f_r, f_d, f_r - f_d)$ method with weighted average patch aggregation. Blue color stands for low and orange color represents high values.

assigned with a very small relative weight α^*. By that, these patches have only little influence on the global quality, calculated by the weighted average, and the imagewise prediction error $|q - q_t|$ of these patches has with Eq. 1 only little influence on the loss function Eq. 2. Thus, the backpropagation for the specific patches gets 'stuck' due to small gradients. The resulting imagewise quality estimate is scattered against ground truth on the right hand side of Fig. 2.

In order to better understand the effect of weighted average patch aggregation we analyze the blockwise spatial distribution of the weights α^* and patch-wise predictions y_i in Fig. 3. The first example shows a relatively uniform patchwise score distribution y_i, but a weighting which focuses more on the forest than the river or the sky (similar to SSIM and FSIM). The effect of the weighting in the second example is even clearer. Here it separates the foreground object from the background. For this image this is reasonable, as the foreground patches, showing e.g. feathers of the parrot, contain finer structure than the blurry background. Therefore the background is relative to the foreground less important for the globally perceived quality of the image. Also in the last example the weightings have a region separation effect, namely they separate the scores assigned to the sky from the scores assigned to the rest. The former ones overestimate the MOS value, thus are downweighted by the proposed method.

IV. Conclusion

We applied a deep CNN with a feature fusion architecture to the problem of FR IQA and showed that it outperforms state-of-the-art IQA methods on the LIVE dataset. We evaluated three feature fusion approaches and two patch aggregation techniques and briefly discussed their advantages and limitations. By proposing weighted average patch aggregation we considered local differences in relative influence to global quality. By this the performance of the proposed method can be improved further.

In future work we will study the influence of the CNN architecture (e.g., depth) on FR IQA and investigate what features are actually learned by the network using explanation methods [27], [28]. Furthermore we will explore the performance of our method in a cross-dataset scenario and separately study the different distortion types.

References

[1] Cisco Visual Networking Index, "Global Mobile Data Traffic Forecast Update 2014–2019 White Paper, Feb 2015," *See: http://www. cisco. com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862. html*, 2015.

[2] J. Wu, W. Lin, G. Shi, and A. Liu, "Reduced-reference image quality assessment with visual information fidelity," *Multimedia, IEEE Transactions on*, vol. 15, no. 7, pp. 1700–1705, 2013.

[3] S. Bosse, Q. Chen, M. Siekmann, W. Samek, and T. Wiegand, "Shearlet-based reduced reference image quality assessment," in *Image Process. (ICIP), 2016 IEEE Int. Conf.* 2016, pp. 2052–2056, IEEE.

[4] B. Girod, "What's Wrong with Mean-squared Error?," in *Digital Images and Human Vision*, pp. 207–220. MIT Press, 1993.

[5] W. Lin and C. C. Jay Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[6] A.B. Watson, R. Borthwick, and M. Taylor, "Image quality and entropy masking," in *SPIE Proceedings*, 1997, vol. 3016, pp. 1–11.

[7] S. J. Daly, "Application of a noise-adaptive contrast sensitivity function to image data compression," *Optical Engineering*, vol. 29, no. 8, pp. 977–987, 1990.

[8] J. Lubin, "A human vision system model for objective picture quality measurements," *International Broadcasting Convention*, pp. 498–503, 1997.

[9] Y. Jia, W. Lin, and A. Kassim, "Estimating just-noticeable distortion for video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 7, pp. 820–829, 2006.

[10] 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.

[11] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[12] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: a highly efficient perceptual image quality index," *Image Processing, IEEE Transactions on*, vol. 23, no. 2, pp. 684–695, 2014.

[13] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A haar wavelet-based perceptual similarity index for image quality assessment," *arxiv preprint*, vol. abs/1607.06140, 2016.

[14] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual dog model fused with random forest," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3282–3292, 2015.

[15] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1098–1105.

[16] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-Time No-Reference Image Quality Assessment Based on Filter Learning," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 987–994, 2013.

[17] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2394–2402, 2015.

[18] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional Neural Networks for No-Reference Image Quality Assessment," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 1733–1740.

[19] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Image Process. (ICIP), 2016 IEEE Int. Conf.* 2016, pp. 3773–3777, IEEE.

[20] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Full-Reference Image Quality Assessment Using Neural Networks," in *Qual. Multimed. Exp. (QoMEX), 2016 8th Int. Conf.* IEEE, 2016.

[21] H. R. Sheikh and A. C. Bovik, "Image information and visual quality.," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 15, no. 2, pp. 430–444, 2006.

[22] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C.J. Kuo, "Color Image Database TID2013 : Peculiarities and Preliminary Results," *Visual Information Processing (EUVIP), 2013 4th European Workshop on*, pp. 106–111, 2013.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Iclr*, pp. 1–14, 2015.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[25] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout : A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research (JMLR)*, vol. 15, pp. 1929–1958, 2014.

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

[27] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. e0130140, 2015.

[28] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2912–2920.