Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, *Member, IEEE*, Thomas Wiegand, *Fellow, IEEE*, and Wojciech Samek, *Member, IEEE*

***Abstract*— We present a deep neural network-based approach to image quality assessment (IQA). The network is trained end-to-end and comprises ten convolutional layers and five pooling layers for feature extraction, and two fully connected layers for regression, which makes it significantly deeper than related IQA models. Unique features of the proposed architecture are that: 1) with slight adaptations it can be used in a no-reference (NR) as well as in a full-reference (FR) IQA setting and 2) it allows for joint learning of local quality and local weights, i.e., relative importance of local quality to the global quality estimate, in an unified framework. Our approach is purely data-driven and does not rely on hand-crafted features or other types of prior domain knowledge about the human visual system or image statistics. We evaluate the proposed approach on the LIVE, CISQ, and TID2013 databases as well as the LIVE In the wild image quality challenge database and show superior performance to state-of-the-art NR and FR IQA methods. Finally, cross-database evaluation shows a high ability to generalize between different databases, indicating a high robustness of the learned features.**

***Index Terms*— Full-reference image quality assessment, no-reference image quality assessment, neural networks, quality pooling, deep learning, feature extraction, regression.**

I. Introduction

DIGITAL video is ubiquitous today in almost every aspect of life, and applications such as high definition television, video chat, or internet video streaming are used for information programs, entertainment and private communication. In most applications, digital images are intended to be viewed by humans. Thus, assessing its perceived quality is essential for most problems in image communication and processing. In a communication system, for example, when an image arrives at the ultimate receiver, typically a human viewer, it has passed a pipeline of processing stages, such as digitization, compression or transmission. These different stages introduce distortions into the original image, possibly visible to human viewers, that may exhibit a certain level of annoyance in the viewing experience. Hence, predicting perceived visual quality computationally is crucial for the optimization and evaluation of such a system and its modules.

Generally, image quality measures (IQMs) are classified depending on the amount of information available from an original reference image — if existing at all. While full-reference (FR) approaches have access to the full reference image, no information about it is available to no-reference (NR) approaches. Since for reduced-reference (RR) image quality assessment (IQA) only a set of features extracted from the reference image is available to the algorithm, it lies somewhere in the middle of this spectrum. As no information about the original is exploited, NR IQA is a more difficult problem than FR IQA and potentially the most challenging problem in IQA. NR IQA has a wide range of applications as in practice often no reference is available. However, for many applications, such as the optimization of video coding systems, unconstrained NR IQA is not a feasible approach —imagine a video codec that, as one example, reconstructs a noise and blur free version of the movie *Blair Witch Project* or, as another example, independently of the input, always reconstructs the same image (but this of perfect quality).

Perceptually accurate IQA relies on computational models of the human visual system (HVS) and/or natural image statistics and is not an easy task [1]. The underlying model employed in an IQM allows for conceptual classification, traditionally in bottom-up and top-down approaches. While former approaches are based on a computational system simulating the HVS by modeling its relevant components, latter ones treat the HVS as a black box and track distortion specific deviations in image statistics. In the NR IQA case this is typically done based on a model of general statistical regularities of natural images [2]. Typically, these statistical image models are used to extract features related to perceived quality that

Manuscript received January 13, 2017; revised July 6, 2017 and September 14, 2017; accepted September 27, 2017. Date of publication October 10, 2017; date of current version October 20, 2017. This work was supported in part by the German Ministry for Education and Research as Berlin Big Data Center under Grant 01IS14013A, in part by the Institute for Information and Communications Technology Promotion through the Korea Government under Grant 2017-0-00451, and in part by DFG. The work of K.-R. Müller was supported by the National Research Foundation of Korea through the Ministry of Education, Science, and Technology in the BK21 Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kalpana Seshadrinathan. *(Sebastian Bosse and Dominique Maniry contributed equally to this work.) (Corresponding author: Sebastian Bosse.)*

S. Bosse, D. Maniry, and W. Samek are with the Department of Video Coding and Analytics, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany (e-mail: sebastian.bosse@hhi.fraunhofer.de; wojciech.samek@hhi.fraunhofer.de).

K.-R. Müller is with the Machine Learning Laboratory, Berlin Institute of Technology, 10587 Berlin, Germany, also with the Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, South Korea, and also with the Max Planck Institute for Informatics, 66123 Saarbrücken, Germany (e-mail: klaus-robert.mueller@tu-berlin.de).

T. Wiegand is with the Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany, and also with the Media Technology Laboratory, Berlin Institute of Technology, 10587 Berlin, Germany (e-mail: twiegand@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2760518

are input to trainable regression models. With the rise of machine learning, recently a third category of IQA emerged, comprising approaches that are purely data-driven, do not rely on any explicit model and allow for end-to-end optimization of feature extraction and regression. Our approach presented in this paper belongs to this new class of data-driven approaches and employs a deep neural network for FR and NR IQA. In order to improve prediction accuracy, IQMs can be combined with saliency models. This extension aims at estimating the likelihood of regions to be attended to by a human viewer and to consider this likelihood in the quality prediction.

Until recently, problems in computer vision such as image classification and object detection were tackled in two steps: (1) designing appropriate features and (2) designing learning algorithms for regression or classification. Although the extracted features were input to the learning algorithm, both of these steps were mostly independent from each other. Over the last years, convolutional neural networks (CNNs) have shown to outperform these traditional approaches. One reason is that they allow for joint end-to-end learning of features and regression based on the raw input data without any hand-engineering. It was further shown that in classification tasks deep CNNs with more layers outperform shallow network architectures [3].

This paper studies the use of a deep CNN with an architecture [3] largely inspired by the organization of the primates' visual cortex, comprising 10 convolutional layers and 5 pooling layers for feature extraction, and 2 fully connected layers for regression, in a general IQA setting and shows that network depth has a significant impact on performance. We start with addressing the problem of FR IQA in an end-to-end optimization framework. For that, we adapt the concept of Siamese networks known from classification tasks [4], [5] by introducing a feature fusion stage in order to allow for a joint regression of the features extracted from the reference and the distorted image. Different feature fusion strategies are discussed and evaluated.

As the number of parameters to be trained in deep networks is usually very large, the training set has to contain enough data samples in order to avoid overfitting. Since publicly available quality-annotated image databases are rather small, this renders end-to-end training of a deep network a challenging task. We address this problem by artificially augmenting the datasets, i.e., we train the network on randomly sampled patches of the quality annotated images. For that, image patches are assigned quality labels from the corresponding annotated images. Different to most data-driven IQA approaches, patches input to the network are not normalized, which enables the proposed method to also cope with distortions introduced by luminance and contrast changes. To this end, global image quality is derived by pooling local patch qualities simply by averaging and, for convenience, this method is referred to as Deep Image QuAlity Measure for FR IQA (DIQaM-FR).

However, neither local quality nor relative importance of local qualities is uniformly distributed over an image. This leads to a high amount of label noise in the augmented datasets. Thus, as a further contribution of this paper, we assign a patchwise relative weight to account for its influence on the global quality estimate. This is realized by a simple change to the network architecture and adds two fully connected layers running in parallel to the quality regression layers, combined with a modification of the training strategy. We refer to this method as Weighted Average Deep Image QuAlity Measure for FR IQA (WaDIQaM-FR). This approach allows for a joint optimization of local quality assessment and pooling from local to global quality, formally within the classical framework of saliency weighted distortion pooling.

After establishing our approaches within a FR IQA context, we abolish one of the feature extraction paths in the Siamese network. This adaptation allows to apply the network within a NR context as well. Depending on the spatial pooling strategy used, we refer to the NR IQA models as Deep Image QuAlity Measure for NR IQA (DIQaM-NR) and Weighted Average Deep Image QuAlity Measure for NR IQA (WaDIQaM-NR).

Interestingly, starting with a FR model, our approach facilitates systematic reduction of the amount of information from the reference image needed for accurate quality prediction. Thus, it helps closing the gap between FR and NR IQA. We show that this allows for exploring the space of RR IQA from a given FR model without retraining. In order to facilitate reproducible research, our implementation is made publicly available at https://github.com/dmaniry/deepIQA.

The performance of the IQA models trained with the proposed methods are evaluated and compared to state-of-the-art IQMs on the popular LIVE, TID2013 and CISQ image quality databases. The models obtained for NR IQA are additionally evaluated on the more recent LIVE In the Wild Image Quality Challenge Database (that, for convenience, we will refer to as CLIVE). Since the performance of data-driven approaches largely relies on on the data used for training we analyze the generalization ability of the proposed methods in cross-database experiments.

The paper is structured as follows: In Section II we give an overview over state-of-the-art related to the work presented in this paper. Section III develops and details the proposed methods for deep neural network-based FR and NR IQA with different patch aggregation methods. In Section IV the presented approaches are evaluated and compared to related IQA methods. Further, weighted average patch aggregation, network depth, and reference reduction are analyzed. We conclude the paper with a discussion and an outlook to future work in Section V.

II. Related Work

A. Full-Reference Image Quality Assessment

The most simple and straight-forward image quality metric is the mean square error (MSE) between reference and distorted image. Although being widely used, it does not correlate well with perceived visual quality [6]. This led to the development of a whole zoo of image quality metrics that strive for a better agreement with the image quality as perceived by humans [7].

Most popular quality metrics belong to the class of top-down approaches and try to identify and exploit distortion-related changes in image features in order to estimate perceived quality. These kinds of approaches can be found in the FR, RR, and NR domain. The SSIM [8] is probably the most prominent example of these approaches. It considers the sensitivity of the HVS to structural information by pooling luminance similarity (comparing local mean luminance), contrast similarity (comparing local variances) and structural similarity (measured as local covariance). The SSIM was not only extended for multiple scales to the MS-SSIM [9], but the framework of pooling complementary features similarity maps served as inspiration for other FR IQMs employing different features, such as the FSIM [10], the GMSD [11], the SR-SIM [12] or HaarPSI [13]. DeepSim [14] extracts feature maps for similarity computation from the different layers of a deep CNN pre-trained for recognition, showing that features learned for image recognition are also meaningful in the context of perceived quality. The difference of Gaussian (DOG)-SSIM [15] belongs somewhat to the top-down as well as to the bottom-up domain, as it mimics the frequency bands of the contrast sensitivity function using a DOG-based channel decomposition. Channels are then input to SSIM in order to calculate channel-wise quality values that are pooled by a trained regression model to an overall quality estimate. The MAD [16] distinguishes between supra- and near-threshold distortions to account for different domains of human quality perception.

Combining several hand-crafted IQMs can have better performance than any single IQM in the set [17].

Although feature learning [14] and regression [15], [17] have been employed in FR IQA, to our best knowledge, so far no end-to-end trained method was used for this task.

B. No-Reference Image Quality Assessment

A typical approach to NR IQA is to model statistics of natural images and regress parametric deviations from this model to perceived image degradations. As these parameters and its deviations may depend on the distortion type, the DIIVINE framework [18] identifies the distortion type affecting an image in a first step and uses a distortion-specific regression scheme to estimate the perceived quality in a second step. The statistical features are calculated based on an oriented subband decomposition. BLIINDS-II [19] uses a generalized Gaussian density function to model block DCT coefficients of images. BRISQUE [20] proposes a NR IQA approach that utilizes an asymmetric generalized Gaussian distribution to model images in the spatial domain. The modeled image features here are differences of spatially neighbored, mean subtracted and contrast normalized image samples. NIQE [21] extracts features based on a multivariate Gaussian model and relates them to perceived quality in an unsupervised manner. In order to cope with more complex and authentic distortion types FRIQUEE [22], [23] employs a deep belief network of 4 layers trained to classify bins of 10 different distortion ranges. Input to the network is a set of handcrafted feature maps and the feature representation on the last hidden layer is extracted to be input to support vector regression (SVR) for quality prediction.

CORNIA [24] is one of the first purely data-driven NR IQA methods combining feature and regression training. Here, a codebook is constructed by k-means clustering of luminance and contrast normalized image patches. Soft-encoded distances between visual codewords and patches extracted from distorted images are used as features that are pooled and regressed using SVR for estimating image quality. This approach is refined to the semantic obviousness metric (SOM) [25], where object-like regions are detected and the patches extracted from these regions are input to CORNIA. Similarly to CORNIA, QAF [26] constructs a codebook using sparse filter learning based on image log-Gabor responses. As log-Gabor responses are often considered a low level model of the HVS, conceptually, QAF also belongs to the bottom-up domain.

Motivated by the recent success of CNNs for classification and detection tasks and the notion that the connectivity patterns in these networks resemble those of the primate visual cortex, [27] proposes a shallow CNN consisting of 1 convolutional layer, 1 pooling layer and 2 fully-connected layers, that combines feature extraction and regression. Quality is estimated on contrast normalized image patches and patch-wise quality is pooled to imagewise quality by averaging. BIECON [28] proposes an interesting approach for data augmentation and tackles CNN-based NR IQA in 2 steps: First, a local quality is estimated based on normalized image patches employing a CNN of 2 convolutional, 2 pooling and 5 fully-connected layers. This network is trained to replicate a conventional FR IQM such as SSIM or GMSD within a NR framework. Second, mean values and the standard deviations of the extracted patchwise features are regressed to an imagewise quality estimate employing a perceptron with one hidden layer. Preliminary results on the application of deeper neural networks, trained end-to-end, for IQA have been presented in [29] and [30] and are extended and further evaluated in this paper.

C. Salience and Attention for IQA

Not every region in an image receives the same amount of attention by viewers and generally distortions in regions that attract viewers' attention are assumed to be more disturbing than in other regions. This led to the idea to combine models of visual saliency with IQMs by weighting the local quality y_i of a region i with the corresponding local saliency w_i to the overall image quality Q with

$$Q = \frac{\sum_i w_i y_i}{\sum_i w_i} \tag{1}$$

Various models of saliency have been proposed and combined with different IQMs to improve prediction performance [31]. The Visual Saliency-Induced Index (VSI) [32] takes the local saliency from reference and distorted image as features maps in a similarity formalism (Section II-A) and combines it with similarity from local gradients and local chrominance. The combined similarity maps are then spatially pooled by the local maximal saliency of reference and distorted image.

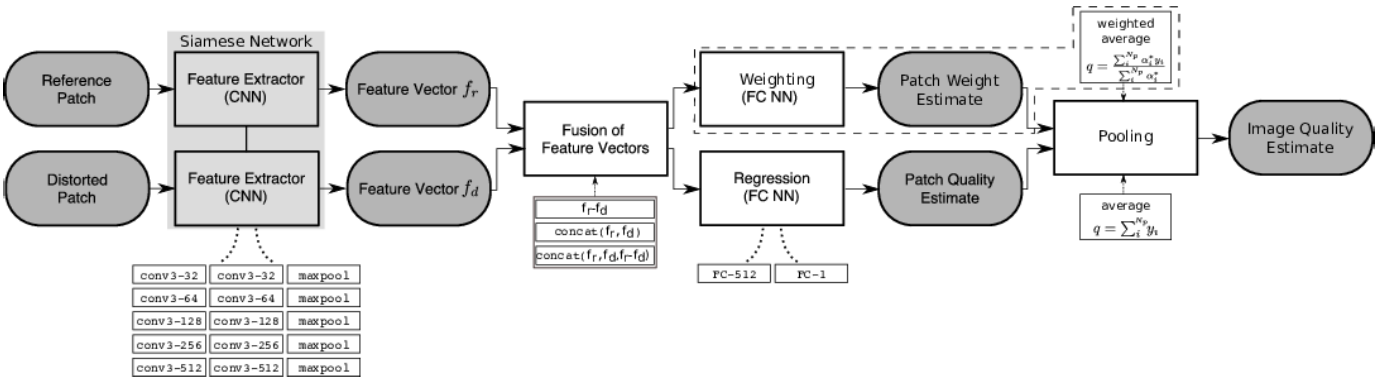


Fig. 1. Deep neural network model for FR IQA. Features are extracted from the distorted patch and the reference patch by a CNN and fused as difference, concatenation or concatenation supplementary with the difference vector. The fused feature vector is regressed to a patchwise quality estimate. The dashed-boxed branch of the network indicates an optional regression of the feature vector to a patchwise weight estimate that allows for pooling by weighted average patch aggregation.

Saliency has usually been extracted from the reference image and, as for VSI, in some cases additionally from the distorted image. Employing saliency estimated from the reference image shifts NR IQA to the RR domain. So far, saliency models incorporated in IQMs were mostly developed to predict viewers' attention to specific regions of an image, but not to predict perceived quality explicitly in a joint optimization approach.

III. Deep Neural Networks for Image Quality Assessment

A. Neural Network-Based FR IQA

Siamese networks have been used to learn similarity relations between two inputs. For this, the inputs are processed in parallel by two networks sharing their synaptic connection weights. This approach has been used for signature [4] and face verification [5] tasks, where the inputs are binarily classified as being of the same category or not. For FR IQA we employ a Siamese network for feature extraction. In order to use the extracted features for the regression problem of IQA, feature extraction is followed by a feature fusion step. The fused features are input to the regression part of the network. The architecture of the proposed network is sketched in Fig. 1 and will be further detailed in the following.

Motivated by its superior performance in the 2014 ILSRVC classification challenge [33] and its successful adaptation for various computer vision tasks [34], [35], VGGnet [3] was chosen as a basis for the proposed networks. While still a straightforward, but deep CNN architecture, VGGnet was the first neural network to employ cascaded convolutions kernels small as 3×3. The input of the VGG network are images of the size 224×224 pixels. To adjust the network for smaller input sizes such as 32×32 pixel-sizes patches, we extend the network by 3 layers (conv3-32, conv3-32, maxpool) plugged in front of the original architecture. Our proposed VGGnet-inspired CNN comprises 14 weight layers that are organized in a feature extraction module and a regression module. The features are extracted in a series of 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 layers.[1] The fused features (see Section III-B) are regressed by a sequence of one FC-512 and one FC-1 layer. This results in about 5.2 million trainable network parameters. All convolutional layers apply 3×3 pixel-size convolution kernels and are activated through a rectified linear unit (ReLU) activation function $g = \max(0, \sum_i w_i a_i)$, where g, w_i and a_i denote the output, the weight and the input of the ReLU, respectively [36]. In order to obtain an output of the same size as the input, convolutions are applied with zero-padding. All max-pool layers have 2×2 pixel-sized kernels. Dropout regularization with a ratio of 0.5 is applied to the fully-connected layers in order to prevent overfitting [37].

For our IQA approach, images are subdivided into 32×32 sized patches that are input to the neural network. Local patchwise qualities are pooled into a global imagewise quality estimate by simple or weighted average patch aggregation. The choice of strategy for spatial pooling affects the training of the network and will be explained in more detail in Section III-C.

For convenience we refer to the resulting models as Deep Image QuAlity Measure for FR IQA (DIQaM-FR) and Weighted Average Deep Image QuAlity Measure for FR IQA (WaDIQaM-FR).

B. Feature Fusion

In order to serve as input to the regression part of the network, the extracted feature vectors f_r and f_d are combined in a feature fusion step. In the FR IQA framework, concatenating f_r and f_d to $\texttt{concat}(f_r, f_d)$ without any further modifications is the simplest way of feature fusion. f_r and f_d are of identical structure, which renders the difference $f_r - f_d$ to be a meaningful representation for distance in feature space. Although the regression module should be able to learn $f_r - f_d$ by itself, the explicit formulation might ease the actual regression task. This allows for two other simple feature fusion strategies, namely the difference $f_r - f_d$, and concatenation to $\texttt{concat}(f_r, f_d, f_r - f_d)$.

[1] Notation is borrowed from [3] where conv⟨receptive field size⟩-⟨number of channels⟩ denotes a convolutional layer and FC⟨number of channels⟩ a fully-connected layer

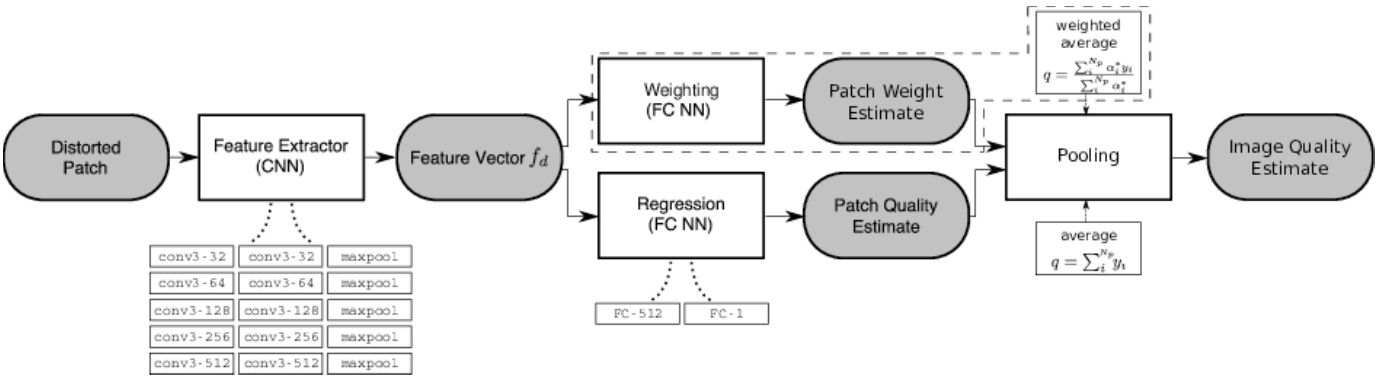


Fig. 2. Deep neural network for NR IQA. Features are extracted from the distorted patch by a CNN. The feature vector f_d is regressed to a patchwise quality estimate. Patchwise estimates are aggregated to a global image quality estimate. The dashed-boxed branch of the network indicates an optional regression of the feature vector to a patchwise weight estimate that allows for pooling by weighted average patch aggregation.

C. Spatial Pooling

1) Pooling by Simple Averaging: The simplest way to pool locally estimated visual qualities y_i to a global imagewise quality estimate $\hat{q}$ is to assume identical relative importance of every image region, or, more specifically, of every image patch i as

$$\hat{q} = \frac{1}{N_p} \sum_i^{N_p} y_i, \tag{2}$$

where N_p denotes the number of patches sampled from the image. For regression tasks, commonly the MSE is used as minimization criterion. However, as simple average quality pooling implicitly assigns the locally perceived quality to be identical to globally perceived quality q_t this approach introduces a certain degree of label noise into the training data. Optimization with respect to mean absolute error (MAE) puts less emphasis on outliers and reduces their influence. As our IQA problem is a regression task, we choose MAE as a less outlier sensitive alternative to MSE. The loss function to be minimized is then

$$E_{simple} = \frac{1}{N_p} \sum_i^{N_p} |y_i - q_t|. \tag{3}$$

In principle, the number of patches N_p can be chosen arbitrarily. A complete set of all non-overlapping patches would ensure all pixels of the image to be considered and, given the same trained CNN model, be mapped to reproducible scores.

2) Pooling by Weighted Average Patch Aggregation: As already shortly discussed in Section II-C, quality perceived in a local region of an image is not necessarily reflected by the imagewise globally perceived quality and vice-versa, e.g. due to spatially non-uniformly distributed distortion, summation or saliency effects or combinations of these influencing factors. In the above described pooling-by-average approach this is accounted for only very roughly by the choice of a less outlier-sensitive loss function. However, spatial pooling by averaging local quality estimates does not consider the effect of spatially varying perceptual relevance of local quality.

We address the spatial variance of relative image quality by integrating a second branch into the regression module of the network that runs in parallel to the patchwise quality regression branch (see Fig. 1) and shares the same dimensionality. This branch outputs an α_i for a patch i. By activating α_i through a ReLU and adding a small stability term ϵ

$$\alpha_i^* = \max(0, \alpha_i) + \epsilon \tag{4}$$

it is guarantied to be $\alpha_i^* > 0$ and can be used to weight the estimated quality y_i of the respective patch i.

With the normalized weights

$$p_i = \frac{\alpha_i^*}{\sum_j^{N_p} \alpha_j^*}. \tag{5}$$

the global image quality estimate $\hat{q}$ can be calculated as

$$\hat{q} = \sum_i^{N_p} p_i y_i = \frac{\sum_i^{N_p} \alpha_i^* y_i}{\sum_i^{N_p} \alpha_i^*}. \tag{6}$$

As in Eq. 3, the number of patches N_p can be set arbitrarily. Comparing Eq. 6 to Eq. 1 shows that the proposed pooling method implements a weighting technique formally equivalent to the framework of linear saliency weighting as described in Section II-C.

For joint end-to-end training the loss function to be minimized is then

$$E_{weighted} = |\hat{q} - q_t|. \tag{7}$$

D. Network Adaptations for NR IQA

Abolishing the branch that extracts features from the reference patch from the Siamese network is a straight forward approach to use the proposed deep network in a NR IQA context. As no features from the reference patch are available anymore, no feature pooling is necessary. However, both spatial pooling methods detailed in Section III-C are applicable for NR IQA as well. The resulting approaches are referred to as Deep Image QuAlity Measure for NR IQA (DIQaM-NR) and Weighted Average Deep Image QuAlity Measure for NR IQA (WaDIQaM-NR). This amounts to the same loss functions as for the FR IQA case. The resulting architecture of the neural network adapted for NR IQA is illustrated in Fig. 2.

E. Training

The proposed networks are trained iteratively by backpropagation [38], [39] over a number of epochs, where one epoch is defined as the period during which each sample from the training set has been used once. In each epoch the training set is divided into mini-batches for batchwise optimization. Although it is possible to treat each image patch as a separate sample in the case of simple average pooling, for weighting average pooling image patches of the same image can not be distributed over different mini-batches, as their output is combined for the calculation of the normalized weights in the last layer. In order to train all methods as similar as possible, each mini-batch contains 4 images, each represented by 32 randomly sampled image patches which leads to the effective batch size of 128 patches. The backpropagated error is the average loss over the images in a mini-batch. For training the FR IQA networks, the respective reference patches are included in the mini-batch. Patches are randomly sampled every epoch to ensure that as many different image patches as possible are used in training.

The learning rate for the batchwise optimization is controlled per parameter adaptively using the ADAM method [40] based on the variance of the gradient. Parameters of ADAM are chosen as recommended in [40] as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $\alpha = 10^{-4}$. The mean loss over all images during validation is computed in evaluation mode (i.e. dropout is replaced with scaling) after each epoch. The 32 random patches for each validation image are only sampled once at the beginning of training in order to avoid noise in the validation loss. The final model used in evaluation is the one with the best validation loss. This amounts to early stopping [41], a regularization to prevent overfitting.

Note that the two regression branches estimating patch weight and patch quality do not have identical weights, as the update of the network weights is calculated based on gradients with respect to different parameters.

IV. Experiments and Results

A. Datasets

Experiments are performed on the LIVE [42], TID2013 [43] and CSIQ [44] image quality databases, the NR IQA approach is also evaluated on the LIVE In the Wild Image Quality Challenge Database [22], [23].

The LIVE [42] database comprises 779 quality annotated images based on 29 source reference images that are subject to 5 different types of distortions at different distortion levels. Distortion types are JP2K compression, JPEG compression, additive white Gaussian noise, Gaussian blur and a simulated fast fading Rayleigh channel. Quality ratings were collected using a single-stimulus methodology, scores from different test sessions were aligned. Resulting DMOS quality ratings lie in the range of [0, 100], where a lower score indicates better visual image quality.

The TID2013 image quality database [43] is an extension of the earlier published TID2008 image quality database [45] containing 3000 quality annotated images based on 25 source reference images distorted by 24 different distortion types at 5 distortion levels each. The distortion types cover a wide range from simple Gaussian noise or blur over compression distortions such as JPEG to more exotic distortion types such as non-eccentricity pattern noise. This makes the TID2013 a more challenging database for the evaluation of IQMs. The rating procedure differs from the one used for the construction of LIVE, as it employed a competition-like double stimulus procedure. The obtained mean opinion score (MOS) values lie in the range [0, 9], where larger MOS indicate better visual quality.

The CISQ image quality database contains 866 quality annotated images. 30 reference images are distorted by JPEG compression, JP2K compression, Gaussian blur, Gaussian white noise, Gaussian pink noise or contrast change. For quality assessment, subjects were asked to position distorted images horizontally on a monitor according to its visual quality. After alignment and normalization resulting DMOS values span the range [0, 1], where a lower value indicates better visual quality.

The LIVE In the Wild Image Quality Challenge Database (CLIVE) [22], [23] comprises 1162 images taken under real life conditions with a large variety of objects and scenes captured under varying luminance conditions using different cameras. In that sense the images are authentically distorted with impairments being the result of a mixture of different distortions, such as over- or underexposure, blur, grain, or compression. As such, no undistorted reference images are available. Quality annotations were obtained in the form of MOS in a crowdsourced online study. MOS values lie in the range [0, 100], a higher value indicates higher quality.

B. Experimental Setup

For evaluation, the networks are trained either on LIVE or TID2013 database. For cross-validation, databases are randomly split by reference image. This guarantees that no distorted or undistorted version of an image used in testing or validation has been seen by the network during training. For LIVE, the training set is based on 17 reference images, validation and test set on 6 reference images each. TID2013 is split analogously in 15 training, 5 validation and 5 test images. CLIVE does not contain versions of different quality levels of the same image, therefore splitting in sets can be done straight forward by distorted image. Training set size for CLIVE is 698 images, validation and test set size 232 images each. Results reported are based on 10 random splits. Models are trained for 3000 epochs. Even though some models converge after less than 1000 epochs a high number is used to ensure convergence for all models. During training the network has seen ~48M patches in the case of LIVE, ~178M patches in the case of TID2013, and ~67M in the case of CLIVE.

To assess the generalization ability of the proposed methods the CSIQ image database is used for cross-dataset evaluating the models trained either on LIVE or on TID2013. For training the model on LIVE, the dataset is split into 23 reference images for training and 6 for validation; analogously, for training the model on TID2013, the dataset is split

TABLE I
PERFORMANCE COMPARISON ON LIVE AND TID2013 DATABASES

	IQM	LIVE		TID2013	
		LCC	SROCC	LCC	SROCC
Full-Reference	PSNR	0.872	0.876	0.675	0.687
	SSIM [8]	0.945	0.948	0.790	0.742
	FSIM_C [10]	0.960	0.963	0.877	0.851
	GMSD [11]	0.956	0.958	-	-
	DOG-SSIM [15]	0.963	0.961	0.919	0.907
	DeepSim [14]	0.968	**0.974**	0.872	0.846
	DIQaM-FR (proposed)	0.977	0.966	0.880	0.859
	WaDIQaM-FR (proposed)	**0.980**	0.970	**0.946**	**0.940**
No-Reference	BLIINDS-II[19]	0.916	0.912	0.628	0.536
	DIIVINE [18]	0.923	0.925	0.654	0.549
	BRISQUE [20]	0.942	0.939	0.651	0.573
	NIQE [21]	0.915	0.914	0.426	0.317
	BIECON [28]	0.962	0.961	-	-
	FRIQUEE [22]	0.930	0.950	-	-
	CORNIA [24]	0.935	0.942	0.613	0.549
	CNN [27]	0.956	0.956	-	-
	SOM [25]	0.962	**0.964**	-	-
	DIQaM-NR (proposed)	**0.972**	0.960	**0.855**	**0.835**
	WaDIQaM-NR (proposed)	0.963	0.954	0.787	0.761

TABLE II
PERFORMANCE COMPARISON FOR DIFFERENT SUBSETS OF TID2013

	Noise	Actual	Simple	Exotic	New	Color
PSNR	0.822	0.825	0.913	0.597	0.618	0.535
SSIM [8]	0.757	0.788	0.837	0.632	0.579	0.505
FSIM_C [10]	0.902	0.915	0.947	0.841	0.788	0.775
DOG-SSIM [15]	0.922	0.933	0.959	0.889	0.908	0.911
DIQaM-NR	0.938	0.923	0.885	0.771	0.911	0.899
WaDIQaM-NR	**0.969**	**0.970**	**0.971**	**0.925**	**0.941**	**0.934**

TABLE III
PERFORMANCE EVALUATION FOR NR IQA ON CLIVE

IQM	PLCC	SROCC
FRIQUEE [22]	**0.706**	**0.682**
BRISQUE [20]	0.610	0.602
DIIVINE [18]	0.5577	0.5094
BLIINDS-II [19]	0.4496	0.4049
NIQE [21]	0.4776	0.4210
DIQaM-NR (proposed)	0.601	0.606
WaDIQaM-NR (proposed)	0.680	0.671

into 20 training images and 5 validation images. LIVE and TID2013 have a lot of reference images in common, thus, tests between these two are unsuitable for evaluating generalization for unseen images. For cross-distortion evaluation models trained on LIVE are tested on TID2013 in order to determine how well a model deals with distortions that have not been seen during training in order to evaluate whether a method is truly non-distortion or just many-distortion specific.

Note that different to many evaluations reported in the literature, we use the full TID2013 database and do not ignore any specific distortion type. To make errors and gradients comparable for different databases, the MOS values of TID2013 and CLIVE and the DMOS values of CSIQ have been linearly mapped to the same range as the DMOS values in LIVE. Note that by this mapping high values of y_i represent high local distortion. For evaluation, prediction accuracy is quantified by Pearson linear correlation coefficient (LCC), prediction monotonicity is measured by Spearman rank order coefficient (SROCC). For both correlation metrics a value close to 1 indicates high performance of a specific quality measure.

C. Performance Evaluation

Evaluations presented in this subsection are based on image quality estimation considering $N_P = 32$ patches. Other values of N_p will be discussed in Section IV-F. Performances of the FR IQA models are reported for features fused by $\texttt{concat}(f_r, f_d, f_r - f_d)$; the influence of the different feature fusion schemes are examined in Section IV-G.

1) Full-Reference Image Quality Assessment: The upper part of Table I summarizes the performance of the proposed FR IQA models in comparison to other state-of-the-art methods on the full LIVE and full TID2013 database in terms of LCC and SROCC. With any of the two presented spatial pooling methods, the proposed approach obtains superior performance to state-of-the-art on LIVE, except for DeepSim evaluated by SROCC. On TID2013 DIQaM-FR performs better than most evaluated state-of-the-art methods, but is outperformed by DOG-SSIM.[2] Here, employing weighted average patch aggregation clearly improves the performance and WaDIQaM-FR performs better than any other evaluated IQM. This effect can be observed as well in Table II for the groups of different distortion of TID2013 defined in [43]. While DIQaM-FR performs comparable to the state-of-the-art method, on some groups better, on some worse, WaDIQaM-FR shows superior performance for all grouped distortion types.

2) No-Reference Image Quality Assessment: Performances of the NR IQMs are compared to other state-of-the-art NR IQA methods in the lower part of Table I. The proposed model employing simple average pooling (DIQaM-NR) performs best in terms of LCC among all methods evaluated, and in terms of SROCC performs slightly worse than SOM. Evaluated on TID2013, DIQaM-NR performs superior among all other methods in terms of LCC and SRCC[2]. Although no results were reported for BIECON [28] on the TID2013 dataset, this method achieves a relatively high SROCC $= 0.923$ on the older TID2008 database when 5 distortion types (non-eccentricity pattern noise, local block-wise distortions, mean shift, contrast change) are removed from the analysis. Future investigations will show how BIECON performs on the challenging TID2013 database with all distortions included. In contrast to our FR IQA models for NR IQA the weighted average patch aggregation pooling decreases the prediction performance.

A comparison of performances for the CLIVE database is shown in Table III. Quality assessment on CLIVE is much more difficult than on LIVE or TID2013, thus performances of all methods evaluated are much worse than for the legacy databases. WaDIQaM-NR shows prediction performance superior to most other models, but is clearly outperformed by FRIQUEE. Interestingly and contrasting to the results on LIVE and TID2013, on CLIVE WaDIQaM-NR performs clearly better than DIQaM-NR.

[2]Unfortunately, for many state-of-the-art FR and NR IQA methods no results are reported on TID2013.

TABLE IV
PLCC ON SELECTED DISTORTION OF TID2013

	GB	JEPG	J2K	LBDDI
DIQaM-FR	0.884	0.965	0.900	0.634
WaDIQaM-FR	**0.963**	**0.978**	**0.975**	**0.683**
DIQaM-NR	**0.872**	**0.946**	**0.872**	0.479
WaDIQaM-NR	0.618	0.726	0.816	**0.664**

In order to analyze these apparent contradictory results a little deeper, Table IV shows the performance of (Wa)DIQaM-FR and (Wa)DIQaM-NR for four selected distortion types from TID2013 (Gaussian blur, JPEG compression, JP2K compression and local block-wise distortions of different intensity). While for (WA)DIQaM-FR we see the same behavior on single distortion types as on aggregated distortion types, in the NR IQA case weighted patch aggregation pooling decreases performance for GB, JPEG and JP2K, but increases performance for LBDDI. We conjecture that for most distortions information from the reference image is crucial to assign local weights for pooling, but if distortions are strongly inhomogeneous as LBDDI, the distorted image is sufficient to steer weighting. One of the reasons for CLIVE being so challenging for IQA is that distortions and scenes are spatially much more inhomogeneous than in LIVE or TID2013, which the weighted patch aggregation can compensate for. This also explains the huge increase in performance by weighted patch aggregation pooling for the *exotic* subset of the TID2013 in Table IV as this subset contains a larger amount of inhomogeneous distortion types. The resulting local weights will be examined more closely in the next section.

D. Local Weights

The previous sections showed that the weighted average patch aggregation scheme has an influence that depends on the distortion type and the availability of a reference.

Fig. 3 shows the local qualities y_i and weights α_i^* for an image subject to JP2K compression from TID2013. The MOS value of the distorted image is 34; the relation between prediction accuracies of the four different models are as expected from the previous evaluations (DIQaM-FR: 54, WaDIQaM-FR: 42, DIQaM-NR:60, WaDIQaM-NR: 70). The left column shows the quality and weight maps computed by the proposed FR models, the right column the maps from the NR models. The DIQaM-FR/NR assign higher distortion values to the background of the image than to the two foreground objects (Figs. 3b and 3c). In the FR case, the local weights provide some rough image segmentation as higher weights are assigned to image regions containing objects (Fig. 3f). This fails in the NR case (Fig. 3g), which explains the performance drop from DIQaM-NR to WaDIQaM-NR observed in Section IV-C2.

The local quality and weight maps resulting from an image subject to spatially highly variant distortions, in this example LBDDI from TID2013, is shown in Fig. 4. Here, for WaDIQaM-FR as well as for WaDIQaM-NR the network is able to assign higher weights to the distorted image regions and by that improve prediction accuracy compared

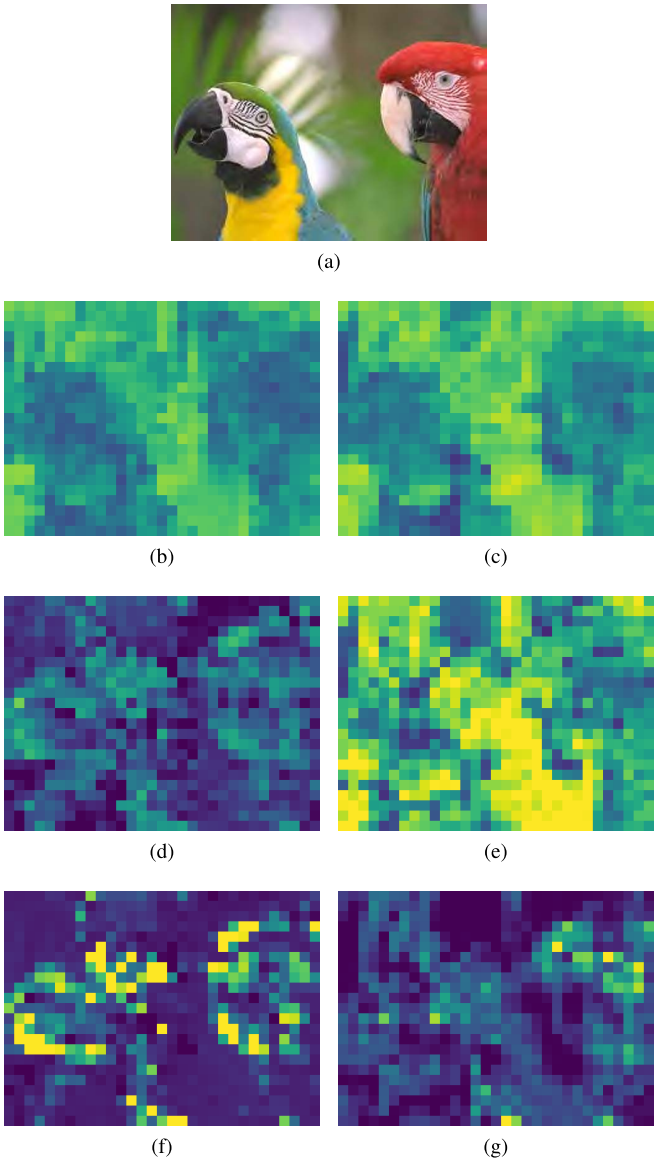


Fig. 3. Local qualities y_i and weights α_i^* for a JP2K distorted image from TID2013. Blue indicates low, yellow high values of local distortions and weights, respectively. The MOS value is 34, predicted qualities are 54 by DIQaM-FR, 42 by WaDIQAM-FR, 60 by DIQaM-NR, and 70 by WaDIQam-NR. (a) Distorted Image. (b) y_i, DIQaM-NR. (c) y_i, DIQaM-FR. (d) y_i, WaDIQaM-FR. (e) y_i, WaDIQaM-NR. (f) y_i, WaDIQaM-NR. (g) α_i^*, WaDIQaM-FR. (h) α_i^*, WaDIQaM-NR.

to the models employing simple average pooling. Note that, as in Fig. 3, WaDIQaM-FR is again able to roughly segment the image, whereas WaDIQaM-NR again fails at segmentation. However, for this extreme distortion type the structure of the image is of less importance.

In Section IV-C2 we conjectured that one reason for WaDIQaM-NR to improve prediction performance over DIQaM-NR for CLIVE, but to decrease performance on LIVE and TID2013 is the higher amount of spatial variance in CLIVE. Fig. 5 exemplifies this effect for two images from CLIVE.

The top row shows the test images, where the left-hand sided one (Fig. 5a) suffers rather globally from underexposure,

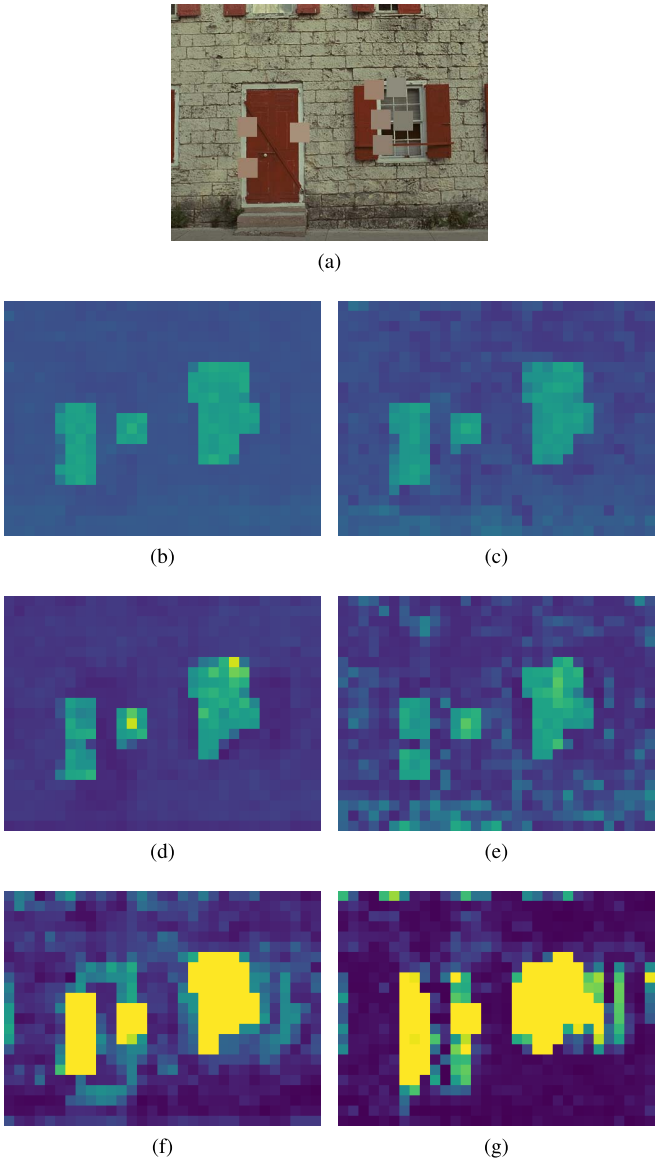


Fig. 4. Local qualities y_i and weights α_i^* for a LBDDI distorted image from TID2013. Blue indicates low, yellow high values of local distortions and weights, respectively. The MOS value is 59, predicted qualities are 30 by DIQaM-FR, 51 by WaDIQAM-FR, 27 by DIQaM-NR, and 53 by WaDIQam-NR. (a) Distorted Image. (b) y_i, DIQaM-FR. (c) y_i, DIQaM-NR. (d) y_i, WaDIQaM-FR. (e) y_i, WaDIQaM-NR. (f) α_i^*, WaDIQaM-FR. (g) α_i^*, WaDIQaM-NR.

rendering identification of a certain area of higher impairment difficult, while the right-hand sided one (Fig. 5b) contains clear regions of interest that are rather easy to identify against the black background. The lower rows show the corresponding quality and weight maps. Fig. 5h shows that for this spatially highly concentrated scene, WaDIQaM-NR is able to identify the patches contributing the most to the overall image structure. However, as Fig. 5g shows, it fails to do so for the homogeneously impaired image.

Another important observation from Fig. 3, Fig. 4 and Fig. 5 is that weighted average patch aggregation has an influence also on the quality maps. Thus, the joint optimization introduces an interaction between y_i and α_i^* that is adaptive to the specific distortion.

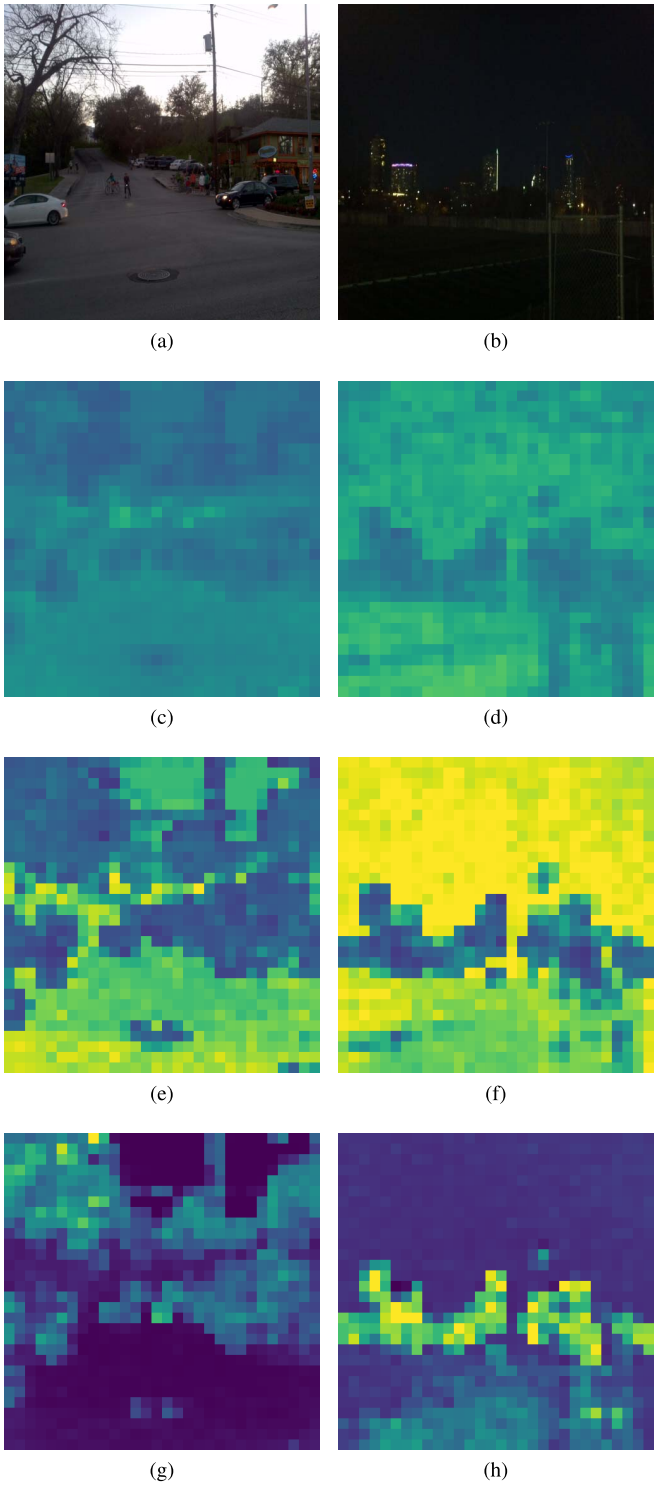


Fig. 5. Local qualities y_i and weight maps α_i^* for two image from CLIVE. Left column: MOS value is 43, predicted qualities are 42 by DIQaM-NR, 34 by WaDIQam-NR. Right column: MOS value is 73, predicted qualities are 56 by DIQaM-NR, 66 by WaDIQam-NR. (a) Input Image. (b) Input Image. (c) y_i, DIQaM-NR. (d) y_i, DIQaM-NR. (e) y_i, WaDIQaM-NR. (f) y_i, WaDIQaM-NR. (g) α_i^*, WaDIQaM-NR. (h) α_i^*, WaDIQaM-NR.

E. Cross-Database Evaluation

1) Full-Reference Image Quality Assessment: Table V shows the results for models trained on LIVE and tested on TID2013 and CSIQ, and for models trained on TID2013 and

TABLE V

SROCC COMPARISON IN CROSS-DATABASE EVALUATION. ALL MODELS ARE TRAINED ON FULL LIVE OR TID2013, RESPECTIVELY, AND TESTED ON EITHER CSIQ, LIVE OR TID2013

Trained on:	LIVE		TID2013	
Tested on:	TID2013	CSIQ	CSIQ	LIVE
DOG-SSIM [15]	**0.751**	**0.914**	0.925	**0.948**
DIQaM-FR (proposed)	0.437	0.660	0.863	0.796
WaDIQaM-FR (proposed)	**0.751**	0.909	**0.931**	0.936

TABLE VI

SROCC IN CROSS-DATABASE EVALUATION FROM. ALL MODELS ARE TRAINED ON THE FULL LIVE DATABASE AND EVALUATED ON CSIQ AND TID2013. THE SUBSETS OF CSIQ AND TID2013 CONTAIN ONLY THE 4 DISTORTIONS SHARED WITH LIVE

	subset		full	
Method	CSIQ	TID2013	CSIQ	TID2013
DIIVINE [18]	-	-	0.596	0.355
BLIINDS-II [19]	-	-	0.577	0.393
BRISQUE [20]	0.899	0.882	0.557	0.367
CORNIA [24]	0.899	0.892	0.663	0.429
QAF [26]	-	-	0.701	0.440
CNN [27]	-	0.920	-	-
SOM [25]	-	**0.923**	-	-
DIQaM-NR (proposed)	**0.908**	0.867	0.681	0.392
WaDIQaM-FR (proposed)	0.866	0.872	**0.704**	**0.462**

tested on LIVE and CSIQ. Results are compared to DOG-SSIM, as most other FR IQA methods compared to do not rely on training. In all combinations of training and test set the DIQaM-FR model shows insufficient generalization capabilities, while WaDIQaM-FR performs best among the two proposed spatial pooling schemes and comparable to DOG-SSIM. The superior results of the model trained on TID2013 over the model trained on LIVE when tested on CISQ indicate that a larger training set may lead to better generalization.

2) No-Reference Image Quality Assessment: For evaluating the generalization ability of the proposed NR IQA models, we extend cross-database experiments presented in [26] with our results. For that, a model trained on the full LIVE database is evaluated on subsets of CSIQ and TID2013, containing only the four distortions types shared between the three databases (JPEG, JP2K, Gaussian blur and white noise). Results are shown in Table VI. While DIQaM-NR shows superior performance compared to BRISQUE and CORNIA on the CISQ subset, the proposed models are outperformed by the other state-of-the-art methods when cross-evaluated on the subset of TID2013. As for the full CSIQ database, the two unseen distortions (i.e. pink additive noise and contrast change) are considerably different in their visual appearance compared to the ones considered during training. Thus, it is not surprising that all compared IQA methods perform worse in this setting. Despite performing worse on the single database experiments, WaDIQaM-NR seems to be able to adapt better to unseen distortions than DIQaM-NR. This is in line with the results on CLIVE, as for CLIVE a specific picture's mixture of distortions is less likely to be in the training set than e.g. for LIVE. Although being a vague comparison as TID2008 contains less distorted images per distortion type, is it worth to note that BIECON obtains a SROCC $= 0.923$ in a similar experiment (trained on LIVE, tested on the 4 distortions types of the smaller TID2008 and excluding one image).

Given the relatively wide variety of distortions types in TID2013 and with only 4 out of 24 distortions being contained in the training set, a model trained on LIVE can be expected to perform worse if tested on TID2013 than if tested on CSIQ. Unsurprisingly, none of the learning-based methods available for comparison is able to achieve a SROCC over 0.5. These results suggest that learning a truly non-distortion-specific IQA metric using only the examples in the LIVE database is hard or even impossible. Nevertheless, the proposed methods obtain competitive, but still very unsatisfactory results.

TABLE VII

SROCC COMPARISON IN CROSS-DATABASE EVALUATION. ALL MODELS ARE TRAINED ON THE FULL TID2013 DATABASE AND EVALUATED ON CSIQ

Method	CSIQ full
DIIVINE [18]	0.146
BLIINDS-II [19]	0.456
BRISQUE [20]	0.639
CORNIA [24]	0.656
DIQaM-NR (proposed)	0.717
WaDIQaM-FR (proposed)	**0.733**

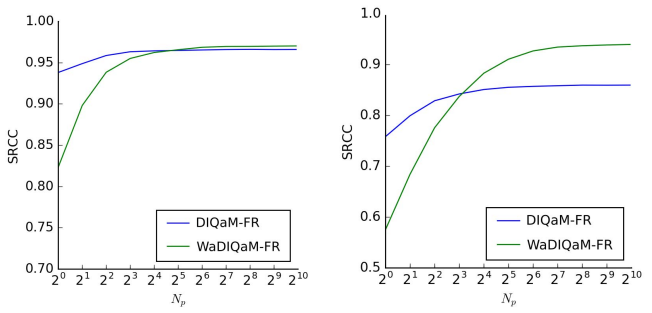
Table VII shows the performance in terms of SROCC for models trained on full TID2013 and tested on full CSIQ. Performance of DIIVINE, BLIINDS-II and CORNIA trained on TID2013 is decreased compared to being trained on LIVE, despite TID2013 being the larger and more diverse training set, while BRISQUE and the proposed models are able to make use of the larger training set size. This follows the notion that the generalization capabilities of deep neural networks depend on the size and diversity of the training set.

Even though the proposed methods outperform comparable methods, a SROCC of 0.733 on the CSIQ dataset is still far from being satisfactory. Despite having more images in total and more distortions than LIVE, the TID2013 has 4 reference images fewer. Thus, training on TID2013 has the same shortcomings as training on LIVE when it comes to adaption to unseen images.

Note that for the evaluation of learning based IQA methods databases are split into training, testing and (for some approaoches) validation sets. Models are the trained and evaluated for a number of different random splits and the performances of these splits need to be pooled for comparison. This renders evaluation a random process. Performances of different models as reported in the literature are obtained based on different split ratios, different numbers of splits and different ways of pooling. This may have a slight influence on the comparison competing methods.

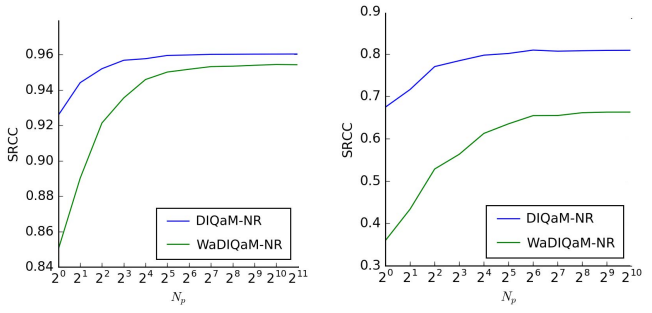
F. Convergence Evaluation

Results presented in the previous sections were obtained when $N_p = 32$ patches are considered for quality estimation. However, the prediction performance and accuracy depends on the number of patches used. This subsection examines the influence of N_p and justifies the choice of $N_p = 32$.



(a) (b)

Fig. 6. SROCC of the proposed CNN for FR IQA in dependence of the number of randomly sampled patches evaluated on LIVE and TID2013.



(a) (b)

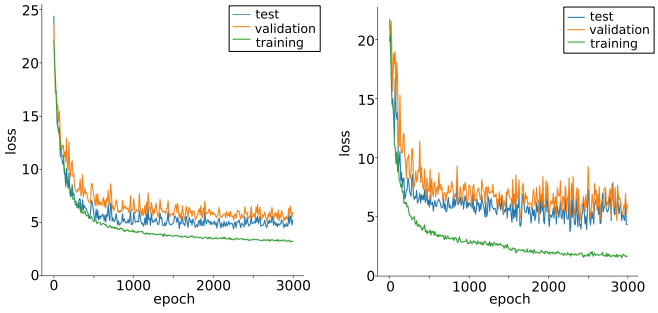
Fig. 7. SROCC of the proposed CNN for NR IQA in dependence of the number of randomly sampled patches evaluated on LIVE and TID2013.

1) Full-Reference Image Quality Assessment: Fig. 6 plots the performance for the models trained and tested on LIVE and TID2013 in dependence of N_p in terms of SROCC and shows a monotonic increase of performance towards saturation for larger N_p. As noted before, weighted average patch aggregation improves the prediction performance over simple averaging, here we see that this holds only for a sufficient number of patches $N_p > 8$. The WaDIQaM-FR saturates at the maximum performance with $N_p \approx 32$, whereas the DIQaM-FR saturates already at $N_p \approx 16$ in all three evaluation metrics.

2) No-Reference Image Quality Assessment: The influence of the number of patches N_p on the prediction performance of DIQaM-NR and WaDIQaM-NR is plotted in Fig. 7. For both pooling methods and on both databases SROCC improves monotonically with increasing number of patches N_p until saturation.

On LIVE, for DIQaM-NR saturation sets in at about $N_p \approx 16$ to reach its maximal performance, whereas WaDIQaM-NR reaches its maximal performance at $N_p \approx 256$. Over the whole range of N_p the performance of average patch aggregation is superior to the performance of weighted average patch aggregation and the difference is largest for small numbers N_p. This is because the weighted average acts as a filter that ignores patches with lower weights and thereby reduces the number of patches considered in quality estimation. Qualitatively the same results are obtained on TID2013.

Fig. 8 shows the course of optimization loss (effectively the MAE) during training, validation and testing in dependence of the number of epochs of training exemplified for DIQaM-NR and WaDIQaM-NR, one random split each, trained on LIVE. For both spatial pooling methods, the loss shows the typical behavior for iterative gradient descent minimization. Interestingly, WaDIQaM-NR achieves a lower loss than DIQaM-NR during training, but is less capable to maintain this level of loss during validation and testing.



(a) (b)

Fig. 8. Loss in training, validation and testing vs. number of epochs in training for DIQaM-NR and WaDIQaM-NR.

TABLE VIII

LCC COMPARISON ON DIFFERENT FEATURE FUSION SCHEMES

Dataset	Method	$f_d - f_r$	concat (f_r, f_d)	concat $(f_r, f_d, f_d - f_r)$
LIVE	DIQaM-FR	0.976	0.974	0.976
	WaDIQaM-FR	0.982	0.977	0.982
TID2013	DIQaM-FR	0.908	0.893	0.908
	WaDIQaM-FR	0.962	0.958	0.965

G. Feature Fusion

Results presented for the FR models in the previous subsections were obtained employing $\texttt{concat}(f_r, f_d, f_d - f_r)$ as feature fusion scheme. The performances of the three feature fusion schemes are reported for LIVE and TID2013 in Table VIII. Mere concatenation of both feature vectors does not fail but consistently performs worse than both of the fusion schemes exploiting the explicit difference of both feature vectors. This suggests that while the model is able to learn the relation between the two feature vectors, providing that relation explicitly helps to improve the performance. The results do not provide enough evidence for preferring one over the other feature fusion methods. This might suggest that adding the original feature vectors explicitly to the representation does not add useful information. Note that the feature fusion scheme might affect the learned features as well, thus, other things equal, it is not guaranteed the extracted features f_r and f_d are equal for different fusion methods.

H. Network Depth

Comparing the proposed NR IQA approach to [27] (see Table I) suggests that the performance of a neural network-based IQM can be increased by making the network deeper. In order to evaluate this observation in a FR context as well, the architecture of the WaDIQaM-FR network was modified by removing several layers and by reducing the intermediate feature vector dimensionality from 512 to 256. This amounts to the architecture conv3-32, conv3-32, maxpool, conv3-64, maxpool, conv3-128,

maxpool, conv3-256, maxpool, FC-256, FC-1 with in total ~790k parameters instead of ~5.2M parameters in the original architecture. When tested on the LIVE database, the smaller model achieves a linear correlation of 0.980, whereas the original architecture achieves 0.984. The same experiment on TID2013 shows a similar result as the shallow model obtains a linear correlation of 0.949, compared to 0.953 obtained by the deeper model. To test whether the decrease in model complexity leads to less overfitting and better generalization, the models trained on TID2013 are additionally tested on CSIQ. The smaller model achieves a SROCC of 0.911, and is outperformed by the original architecture with a SROCC of 0.927. The differences are rather small, but it shows that the deeper and more complex model does lead to a more accurate prediction. However, when computational efficiency is important, small improvements might not justify the five-fold increase in the number of parameters.

I. Bridging From Full- to No-Reference Image Quality Assessment

If argued strictly, the (Wa)DIQaM-Fr as used here is not a FR, but a RR IQM, as only $N_p = 32$ 32×32 patches but not the full image is used for IQA. As shown in Fig. 6, reference information could be reduced even further by reducing the number of patches N_p considered. This can be done without re-training the model. In the proposed architecture the feature vector extracted from one reference patch is 512-dimensional.

The information available from the reference patch can not only be controlled by steering N_p, but also by reducing the dimensionality of the extracted feature vector. A straight forward approach to do so would be to systematically change the network architecture from (Wa)DIQaM-FR to (Wa)DIQaM-NR by reducing the dimensionality of the number of neurons in the last layer of reference branch of the feature extraction module. However, this approach would result in a magnitude of models, each trained for a specific patchwise feature dimensionality.

Another approach is to start with a trained FR model and to linearly reduce the dimensionality of the reference patch feature vector f_r using principal component analysis (PCA). That way, a network trained for FR IQA could be used as a NR IQA method. In this extreme case the reference feature vector f_r is reduced to the mean of the reference feature vectors observed in the training data.

PCA is estimated based on the feature vectors of 4000 reference patches sampled from the training set and used for patchwise dimensionality reduction of f_r during testing. Fig. 9 shows the performance of this RR IQM on one TID2013 test split for increasing dimensionality of the reference patch feature vectors. While the dimensionality reduced version of DIQaM-FR is still able to make useful predictions even without any reference information, this is not the case for the dimensionality reduced version of WaDIQaM-FR method. This corroborates the previous conjecture that weighted average patch aggregation, i.e. the reliable estimation of the weights, is more depending on information from the reference image, at least for homogeneous distortions.

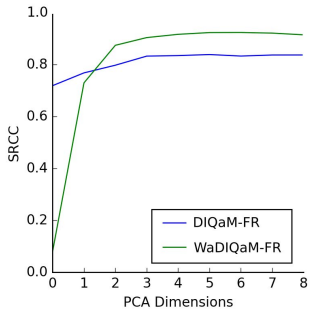


Fig. 9. Average performance on TID2013 for patchwise dimensionality reduced (Wa)DIQaM-FR in terms of SROCC in dependence of the number of principal components of the reference patch feature vector. ($N_p = 32$).

However, for both DIQaM-FR and WaDIQaM-FR 3 principal components (dimensions) are already enough to recover the approximate performance obtained with the 512-dimensional original feature vector. Note that this is the patchwise, not the imagewise feature dimensionality.

Although there are studies analyzing the influence of the feature dimensionality on the performance of RR IQM systematically [46], we are not aware of any unified framework to study NR to FR approaches. However, albeit being a promising framework, a fair comparison, e.g. to the RRED indices [46] would require an analysis of the interaction between patchwise feature dimensionality and number of patches N_p considered.

V. Discussion & Conclusion

This paper presented a neural network-based approach to FR and NR IQA that allows for feature learning and regression in an end-to-end framework. For this, novel network architectures were presented, incorporating an optional joint optimization of weighted average patch aggregation implementing a method for pooling local patch qualities to global image quality. To allow for FR IQA, different feature fusion strategies were studied. The experimental results show that the proposed methods outperforms other state-of-the-art approaches for NR as well as for FR IQA and achieve generalization capabilities competitive to state-of-the-art data-driven approaches. However, as for all current data-driven IQA methods generalization performance offers considerable room for improvement. A principle problem for data-driven IQA is the relative lack of data and significantly larger databases are hopefully to be expected any time soon, potentially using new methods of psychophysiological assessment [47]–[49]. Until then, following the BIECON [28] framework, networks could be pre-trained unsupervised by optimizing on reproducing the quality prediction of a FR IQM, and a pre-trained network employing patch-wise weighting could be refined by end-to-end training. This combines the advantages of the two approaches.

Even though a relative generic neural network is able to achieve high prediction performance, incorporating IQA specific adaptations to the architecture may lead to further improvements. Our results show that there is room for optimization in terms of feature dimensionality and balancing the ratio between network parameters. Here, prediction performance and generalization ability is important to be studied.

In this work, we optimized based on MAE. However, IQA methods are commonly evaluated in terms of correlation and other loss function might be more feasible. We sketched how the proposed framework could be used for RR IQA. We did not present a full-fledged solution, but believe that the results indicate an interesting starting point.

Local weighting of distortion is not a new concept. Classical approaches usually compute the local weights based on a saliency model from the reference image [31], or the reference and the distorted image [32]. Selecting an appropriate saliency model is crucial for the success of this strategy and models best at predicting saliency are not necessarily best for IQA [31]. The proposed weighted average patch aggregation method allows for a joint, end-to-end optimization of local quality estimation and weight assignment.

The equivalence of our weighting scheme to Eq. 1 allows us to interpret the two learned maps as a weighting map and a quality map. Thus, the formal equivalence with the classical approach of linear distortion weighting suggests that local weights capture local saliency. However, this is not necessarily true, as the optimization criterion is not saliency, but image quality. In fact, we showed that the local weights not only depend on the structural (and potentially semantical) organization of the reference image, but also on the distortion type and the spatial distribution of the distortion. This is a fundamental problem for IQA and future work should address the conceptual similarity between the learned weights and saliency models in greater detail.

The proposed network could be adapted for end-to-end learning local weights only in order to improve the prediction performance of any given IQM. This could be combined with conventional regression [50]. Also explanation methods [51], [52] can be applied to better understand what features were actually learned by the network. From an application-oriented perspective the proposed method may be adapted and evaluated for quality assessment of 3D images and 2D and 3D videos.

Although there are still many obstacles and challenges for purely data-driven approaches, the performance of the presented approach and, given the fact that no domain knowledge is necessary, its relative simplicity suggests that neural networks used for IQA have lots of potential and are here with us to stay.

REFERENCES

[1] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4. May 2002, pp. IV-3313–IV-3316.

[2] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ImageNet Challenge*, 2014, pp. 1–10.

[4] J. Bromley *et al.*, "Signature verification using a 'Siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.

[5] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 349–356.

[6] B. Girod, "What's wrong with mean-squared error?" in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA, USA: MIT Press, 1993, pp. 207–220.

[7] W. Lin and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[9] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2. Nov. 2003, pp. 1398–1402.

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

[11] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 668–695, Feb. 2014.

[12] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1473–1476.

[13] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar wavelet-based perceptual similarity index for image quality assessment," *arXiv preprint arXiv:1607.06140*, 2016.

[14] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, Sep. 2017.

[15] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual DOG model fused with random forest," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3282–3292, Nov. 2015.

[16] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011006-1–011006-21, 2010.

[17] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," *Proc. SPIE*, vol. 9394, pp. 93940K, Mar. 2015.

[18] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[19] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[20] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[21] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[22] D. Ghadiyaram and A. C. Bovik. (2015). *LIVE in the Wild Image Quality Challenge Database*. [Online]. Available: http://live.ece.utexas.edu/research/ChallengeDB/index.html

[23] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[24] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

[25] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2394–2402.

[26] L. Zhang, Z. Gu, X. Liu, H. Li, and J. Lu, "Training quality-aware filters for no-reference image quality assessment," *IEEE MultiMedia*, vol. 21, no. 4, pp. 67–75, Oct./Dec. 2014.

[27] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1733–1740.

[28] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

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

[30] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Neural network-based full-reference image quality assessment," in *Proc. Picture Coding Symp. (PCS)*, 2016, pp. 1–5.

[31] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016.

[32] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Aug. 2014.
[33] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[34] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[36] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, vol. 3. 2010, pp. 807–814.
[37] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.
[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[39] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient BackProp," in *Neural Networks: Tricks of the Trade* (Lecture Notes in Computer Science), vol. 7700. Springer, 2012, pp. 9–48.
[40] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[41] L. Prechelt, "Early stopping—But when?" in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 53–67.
[42] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[43] N. Ponomarenko *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *Proc. 4th Eur. Workshop Vis. Inf. Process. (EUVIP)*, 2013, pp. 106–111.
[44] E. C. Larson and D. M. Chandler. (2009). *Consumer Subjective Image Quality Database*. [Online]. Available: http://vision.okstate.edu/index.php
[45] N. Ponomarenko *et al.*, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Mod. Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.
[46] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, Feb. 2012.
[47] S. Bosse *et al.*, "Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2017.2694807.
[48] S. Bosse, K.-R. Müller, T. Wiegand, and W. Samek, "Brain-computer interfacing for multimedia quality assessment," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2016, pp. 003742–003747.
[49] U. Engelke *et al.*, "Psychophysiology-based QoE assessment: A survey," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 6–21, Feb. 2017.
[50] S. Bosse, M. Siekmann, T. Wiegand, and W. Samek, "A perceptually relevant shearlet-based adaptation of the PSNR," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 315–319.
[51] 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, p. e0130140, 2015.
[52] G. Montavon, W. Samek, and K.-R. Müller. (2017). "Methods for interpreting and understanding deep neural networks." [Online]. Available: https://arxiv.org/abs/1706.07979

Sebastian Bosse received the degrees in electrical engineering from RWTH Aachen University, Germany, and the Polytechnic University of Catalonia, Barcelona, Spain, and the Dipl.Ing. degree in electrical engineering and information technology from RWTH Aachen University, Germany. He was a Visiting Researcher with Siemens Corporate Research, Princeton, USA, and Stanford University, USA. He is currently with the Machine Learning Group and the Image Video Coding Group, Video Coding and Analytics Department, Fraunhofer Heinrich Hertz Institute, Berlin, Germany. His major research interests include image and video compression, human visual perception for image communications, and neural correlates of perceived visual quality, and also signal processing and machine learning.

Dominique Maniry received the B.Sc. and M.Sc. degrees in computer engineering from Technische Universität Berlin. He is currently with the Machine Learning Group and the Image Video Coding Group, Video Coding and Analytics Department, Fraunhofer Heinrich Hertz Institute, Berlin, Germany. His major research interests are in machine learning and computer Vision.

Klaus-Robert Müller (M'12) received the degree in physics from Karlsruhe from 1984 to 1989 and the Ph.D. degree in computer science from Technische Universität Karlsruhe in 1992. He has been a Professor of computer science with Technische Universität Berlin since 2006 and has been the Director of Bernstein Focus on Neurotechnology Berlin. After completing a post-doctoral position at GMD FIRST, Berlin, he was a Research Fellow with The University of Tokyo from 1994 to 1995. In 1995, he founded the Intelligent Data Analysis Group, GMD-FIRST (later Fraunhofer FIRST) and directed it until 2008. From 1999 to 2006, he was a Professor with the University of Potsdam. He received the 1999 Olympus Prize by the German Pattern Recognition Society, DAGM, the SEL Alcatel Communication Award in 2006, and the Science Prize of Berlin from the Governing Mayor of Berlin in 2014. He was elected to be a member of the German National Academy of Sciences-Leopoldina and the Berlin Brandenburg Academy of sciences in 2012 and 2017, respectively. His research interests are intelligent data analysis, machine learning, signal processing, and brain-computer interfaces.

Thomas Wiegand is a Professor with the Department of Electrical Engineering and Computer Science, Technical University of Berlin, and jointly heading the Fraunhofer Heinrich Hertz Institute, Berlin, Germany. He received the Dipl.Ing. degree in electrical engineering from the Technical University of Hamburg-Harburg, Germany, in 1995, and the Dr.Ing. degree from the University of Erlangen-Nuremberg, Germany, in 2000. As a student, he was a Visiting Researcher with Kobe University, Japan, the University of California at Santa Barbara, and Stanford University, USA, where he also returned as a Visiting Professor. He was a Consultant with Skyfire Inc., Mountain View, CA. He is currently a Consultant with Vidyo Inc., Hackensack, NJ, USA. Since 1995, he has been an active participant in standardization for multimedia with many successful s submissions to ITU-T and ISO/IEC. In 2000, he was appointed as an Associated Rapporteur of ITU-T VCEG and he was the Co-Chair of ISO/IEC MPEG Video from 2005 to 2009. The projects that he co-chaired for the development of the H.264/MPEG-AVC standard have been recognized by an ATAS Primetime Emmy Engineering Award and a pair of NATAS Technology and Engineering Emmy Awards. For his research in video coding and transmission, he received numerous awards, including the Vodafone Innovations Award, the EURASIP Group Technical Achievement Award, the Eduard Rhein Technology Award, the Karl Heinz Beckurts Award, the IEEE Masaru Ibuka Technical Field Award, and the IMTC Leadership Award. He received multiple best paper awards for his publications. Thomson Reuters named him in their list of The World's Most Influential Scientific Minds 2014 as one of the most cited researchers in his field. He is a recipient of the ITU150 Award.

Wojciech Samek (M'13) received the Diploma degree in computer science from the Humboldt University of Berlin, Germany, in 2010, and the Ph.D. degree in machine learning from the Technische Universität Berlin in 2014. In 2014, he founded the Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, Germany, which he currently directs. He is associated with the Berlin Big Data Center. He was a Scholar of the German National Academic Foundation and a Ph.D. Fellow with the Bernstein Center for Computational Neuroscience.