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

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller,
Thomas Wiegand, and Wojciech Samek

***Abstract*—This paper presents a deep neural network-based approach to image quality assessment (IQA). The network can be trained end-to-end and comprises 10 convolutional layers and 5 pooling layers for feature extraction, and 2 fully connected layers for regression, which makes it significantly *deeper* than related IQA methods. An unique feature of the proposed architecture is that it can be used (with slight adaptations) in a no-reference (NR) as well as in a full-reference (FR) IQA setting. 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. The network estimates perceived quality patchwise; the overall image quality is calculated as the average of these patchwise scores. In order to consider the locally non-uniform distribution of perceived quality in images, we introduce a spatial attention mechanism which performs a weighted aggregation of the patchwise scores. We evaluate the proposed approach on the LIVE, CISQ and TID2013 databases 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 datasets, indicating a high robustness of the learned features.**

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

I. Introduction

DIGITAL visual information is ubiquitous today in almost every aspect of life, mediated by applications such as high definition television, video chat, or internet video streaming. When an image arrives at the ultimate receiver, typically a human, it has passed a pipeline of processing stages, such as acquisition, digitization, compression and transmission. These different stages introduce distortions into the original image. Such distortions may be visible to human viewers and may exhibit a certain level of annoyance in the viewing experience. For the optimization and evaluation of applications as the mentioned, quantifying *perceived* quality is crucial. However, collecting ratings by psychophysical experiments for evaluation is expensive, slow and cumbersome. Although alternative approaches are currently under study [1], [2], in practical real-time applications human ratings are generally not accessible for optimization. This has motivated research on objective image quality assessment (IQA) and image quality measures (IQMs) for decades.

Generally, different approaches to IQA can be classified by the amount of information about the original, undistorted reference image input to the algorithm: While full-reference (FR) approaches to IQA have full access to the whole reference image, no-reference (NR) IQMs do not make use of any specific information about the reference image. Within this spectrum, reduced-reference (RR) approaches to IQA are located somewhat in the middle as only a set of features is accessible as reference information. For conceptual convenience, IQA methods can also be classified based on the underlying model. Traditionally, bottom-up and top-down approaches are distinguished. While the former are based on a computational system simulating the human visual system (HVS) by modeling its relevant components, the latter treat the HVS as a black box, but implement its general hypothesized properties. With the rise of machine learning, recently a third category of IQA emerged, comprising approaches that are purely data-driven and do not rely on any explicit model. Our approach presented in this paper belongs to this new class of data-driven approaches and employs a deep neural network for IQA.

It was shown that in classification tasks deep convolutional neural networks (CNNs) with more layers outperform shallow network architectures [3]. In terms of complexity of features, quality assessment can be considered a simpler problem than classification, as features do not have to represent objects or other semantic information of the image, but instead should relate to those changes in local image statistics that are relevant to quality perception. This raises the question whether it is advantageous to use *deep* methods for IQA or whether shallow methods[1] suffice? The first contribution of this paper is (1) to train a *deep* neural network with 10 convolutional layers and 5 pooling layers for feature extraction, and 2 fully connected layers for regression, end-to-end for estimating image quality in a NR IQA setting and (2) to show that network depth has a significant impact on performance.

For many applications, such as the optimization of video coding and transmission systems, unconstrained NR IQA is not

SB and DM contributed equally.

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

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

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

[1]To the best of our knowledge a deep neural networks-based approach with multiple convolution layers and end-to-end training have not yet been applied for the IQA task.

a feasible approach — as an example imagine a video codec that reconstructs a noise and blur free version of the movie *Blair Witch Project*. Such a codec would destroy the viewing experience by artificially improving the video quality. Thus, as a second contribution, we show that, following the concept of Siamese networks [4], [5], the proposed architecture can be adapted for FR IQA. Siamese networks are commonly used for classification tasks, whereas IQA is a regression task. Thus, we adopt the Siamese network architecture and introduce a feature pooling stage to allow for a joint regression of the features extracted from the reference and the distorted image. We propose and discuss different strategies for feature pooling.

As the number of parameters to be trained in deep networks is usually very high, the training set has to have enough data samples in order to avoid overfitting. Since publicly available quality-annotated image databases are rather small, training a deep network end-to-end becomes 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 the same quality label of the corresponding image. 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. However, neither local quality nor relative importance of local quality for pooled global quality is uniformly distributed over an image. This leads to a high amount of label noise in the augmented datasets. Thus, as third contribution of this paper, we propose a spatial attention mechanism which assigns a relative weight for its influence on global quality to a specific patch. 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.

The performance of the proposed methods and the influence of different parameters are evaluated extensively on TID2013, LIVE and CISQ image quality databases. Data-driven approaches to classification and regression tasks (such as IQA) are highly dependent on the training set. In order to analyze the generalization ability of the proposed methods in a fair manner, we evaluate its performances in cross-database experiments. Based on the FR network, we further explore the NR space by systematically reducing the amount of information available from the reference image by controlling the number of patches used and the dimensionality of the extracted features. By that we close the gap between FR and NR IQA in the proposed framework. In order to facilitate reproducible research, our implementation is publicly available at https://github.com/dmaniry/deepIQA.

The paper is structured as follows: In Section II we give an overview over related approaches in the field of FR and NR IQA. Section III details the proposed methods for deep neural network-based IQA. Experimental evaluations and comparisons to other state-of-the-art methods as well as two experiments analyzing the influence of the network's depth and the reduction of dimensionality are presented in Section IV. We conclude the paper with a discussion in Section V.

II. Related Work

The most simple and straight-forward image quality metric is the mean square error (MSE), calculated as the average of the ℓ_2-norm of the difference between reference and distortion 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. Although being criticized [8], the structural similarity index (SSIM) [9] is probably the most prominent example of top-down FR approaches. It takes into account 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). Following this basic framework of pooling complementary feature maps, the feature similarity index (FSIM) [10] combines two feature maps derived from the phase congruency measure and the local gradients magnitudes of the reference and the distorted image, respectively. The Haar wavelet-based perceptual similarity index (HaarPSI) [11] employs a similar kind of pooling, as local similarity is computed based on the Haar wavelet representation of reference and distorted image and locally weighted based on a visual activity measure that is calculated from the same filter bank. The difference of Gaussian (DOG)-SSIM 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 to an overall quality estimate. A combination of hand-crafted IQMs can have better (or equal) performance than any single IQM in the set. This is shown in [12] by employing a neural network for regression.

As no information about the original is available, NR IQA is considered a more difficult problem than FR IQA. A typical approach to FR IQA is to model statistics of natural images and relate the parameters of this model to perceived image degradations. As these parameters and its deviations may depend on the distortion type, the DIIVINE framework [13] 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 [14] uses a generalized Gaussian density function to model block DCT coefficients of images and predicts quality based on the image-specific parameters of the model. BRISQUE [15] 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. Again, deviations in feature space extracted based on the model are regressed to perceived quality.

As mentioned in Section I, recently several methods steering to a new branch of IQA methods were proposed. These approaches are purely data-driven and, as such, do not make any or very few assumptions on the HVS or the natural image statistics. Most of these data-driven approaches proposed so far belong to the NR IQA domain. CORNIA [16] is one of the first data-driven NR IQA methods. Here, a codebook is constructed by k-means clustering of luminance and contrast normalized image patches. Soft-encoded distances between the visual codewords and the patches extracted from distorted images are used as features that are pooled and regressed using a support vector machine to estimate the image quality. This approach is refined to the semantic obviousness metric (SOM) [17], where object-like regions are detected and the patches extracted from these detected regions are input to CORNIA. Similarly to CORNIA, QAF [18] 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 lives in 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 network resemble the primate visual cortex, [19] proposes a shallow CNN consisting of 1 convolutional layer, 1 pooling layer and 2 fully-connected layers, that combines feature extraction and regression and takes contrast normalized image patches as input.

III. Deep Neural Networks for Image Quality Assessment

A. Basic Network Layout for NR IQA

Motivated by its superior performance in the 2014 ILSRVC classification challenge [20] as well as its successful adaptation for various computer vision tasks [21], [22], the proposed networks are inspired by the VGGnet [3] that employs 16-19 weight layers.

Our proposed adaptation of VGGnet consists of 14 weight layers. The layers are organized as conv3-32, conv3-32, max-pool, conv3-64, conv3-64, maxpool, conv3-128, conv3-128, maxpool, conv3-256, conv3-256, maxpool, conv3-512, conv3-512, maxpool, FC-512, FC-1[2]. This results in about 5.2 million trainable parameters in the network. All convolutional layers apply 3×3 pixel-size convolution kernels and are activated through the 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 [23]. 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 in the regression [24]. The network is trained end-to-end, features are extracted by the convolutional layers, while the two fully-connected layers perform regression. The image is subdivided into 32×32 sized patches that are input to the neural network. The overall quality is then estimated by pooling the patchwise quality estimates that are output of the network into a global quality estimate.

The basic architecture of the neural network is illustrated in Fig. 1 and will be explained in more detail in the following subsections.

[2]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

B. Pooling by Simple Averaging

A simple pooling strategy is to just average the patchwise qualities. Then the estimated imagewise quality $\hat{q}$ can be calculated as

$$\hat{q} = \frac{1}{N_p} \sum_{i}^{N_p} y_i, \tag{1}$$

where y_i represents the local quality estimate and N_p denotes the number of patches sampled from the image. As image quality databases typically contain images that are globally, but not locally quality annotated, ground truth quality labels q_t are available only imagewise, but not patchwise.

For training the network, the mean absolute error (MAE)

$$\begin{aligned} E_{patchwise} =& |\hat{q} - q_t| \\ =& \frac{1}{N_p} \sum_{i}^{N_p} |y_i - q_t| \end{aligned} \tag{2}$$

is minimized by backpropagation [25]. Commonly the MSE is used for regression tasks. However, as stated before, the global image quality is implicitly assigned to the local patchwise quality in Eq. 2, but for most of the patches, the locally perceived quality is not identical to the globally perceived quality, which introduces a certain degree of label noise into the training data. Thus, the MAE is used, as it puts less emphasis on outliers. Also, the quadratic growth of the error can lead to destructive updates of the network parameters due to very high magnitudes of the gradients.

In principle, the number of patches N_p can be set 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, to produce reproducible scores.

C. Pooling by weighted average patch aggregation

As already shortly discussed in Section III-B, the perceived quality of local regions in an image does not necessarily reflect the globally perceived quality of the full image. This might be due to the spatial distribution of the distortion, summation effects or combinations of these two. In the simple pooling-by-average approach described above this problem is only addressed by the choice of a less outlier-sensitive loss function. The spatial pooling of local quality estimates by averaging does not consider the effect of spatially varying importance of local quality for global quality (see an example in Section IV-D2).

In order to overcome this limitation, we propose to integrate a spatial attention branch into the network that runs parallel to and has the same dimensionality and properties as the

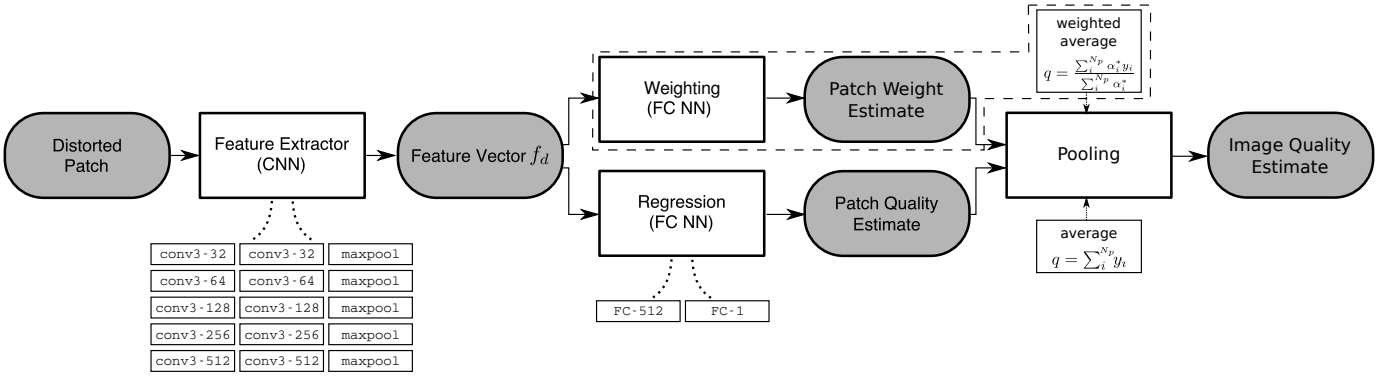


Fig. 1: Graphical representation of the model for deep neural network-based NR IQA. Features are extracted from the distorted patch by a CNN. The feature vector f_d is regressed to a patchwise quality estimate. Patchwise estimates are aggregated to global image quality estimate. The dashed-boxed branch of the network indicates an optional regression of the feature vector to patchwise weight estimate that allows for pooling by weighted-average patch aggregation.

patchwise quality estimation branch (see Fig. 1). The output α_i in this branch is used for weighting the local quality y_i estimated in the other branch for each patch i. By activating the weight α_i through a ReLU and adding a small stability term ϵ

$$\alpha_i^* = \max(0, \alpha_i) + \epsilon \tag{3}$$

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

$$\hat{q} = \sum_i^{N_p} p_i y_i \tag{4}$$

with the normalized weights

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

The stabilization term ϵ is introduced to avoid a division by zero in Eq. 5. For end-to-end training the network weights can now be updated by minimizing the imagewise loss function

$$\begin{aligned} E_{weighted} =& |\hat{q} - q_t| \\ =& \sum_i^{N_p} \Big| \frac{\alpha_i^* y_i}{\sum_j^{N_p} \alpha_j^*} - q_t \Big| \end{aligned} \tag{6}$$

As in Eq. 2, the number of patches N_p can be set arbitrarily.

D. Full-Reference Image Quality Assessment

As outlined in Section I, for some applications NR IQA is conceptually not the optimal approach. Following the concept of Siamese networks [4], [5], the network described in Section III-A can be further modified for FR IQA. Siamese Networks are commonly used to learn a similarity metric of two inputs that are processed in parallel by networks sharing the network weights. This approach has been used for signature [4] and face verification [5], where the two inputs are binarily classified as being of the same category or not. For IQA, not being a classification, but a regression task, this technique has to be adapted and a feature fusion step is introduced between the feature extraction and the regression part of the network.

Fig. 2 sketches the flowchart of the network modified for FR IQA. The network takes a 32×32 pixel sized RGB-image patch from the reference image as an input to the top branch and a 32×32 pixel sized RGB-image patch from the distorted image as an input to the bottom branch. The CNNs of the top and bottom branches are identical in architecture and weight parameters. As for the NR IQA network, the outputs of the top and bottom branch, respectively, are two 512-dimensional feature vectors f_r and f_d extracted from the reference and the distorted image patch. Obviously, as the network in the two branches has exactly the same parameters, f_r and f_d will be identical, if both input patches are identical.

In order to serve as input to the regression part of the network, the feature vectors f_r and f_d have to be combined in a feature fusion step. As f_r and f_d are generated by the same network, both are of equivalent structure and the difference $f_r - f_d$ is a meaningful representation for distance in feature space. Another simple fusion approach is concatenation by combining f_r and f_d to a 1024-dimensional feature vector $\texttt{concat}(f_r, f_d)$ without any further modifications. The following fully-connected layer should be able to learn the difference $f_r - f_d$ from features fused by this strategy and make the explicit formulation obsolete. However, assuming $f_r - f_d$ to be a relevant feature by itself, the explicit formulation might help the actual regression task. Thus, a third feature fusion approach is proposed that combines the former two by combining f_r, f_d and $f_r - f_d$ to the 1536-dimensional vector $\texttt{concat}(f_r, f_d, f_r - f_d)$. This representation combines the strengths of the former two feature fusion approaches, but also introduces redundant information into the regression.

For any of the proposed feature fusion techniques the weighted average patch aggregation can be applied to the FR IQA network as well. In this case the fully-connected layers generating the weight α_i^* take the fused feature representation as input. The relative importance of each patch can then be estimated based on the fused feature vector.

By reducing the number of patches N_p, the FR IQA approach can be tuned towards a reduced reference framework without re-training the network. With the proposed architecture the dimensionality necessary to represent the reference would

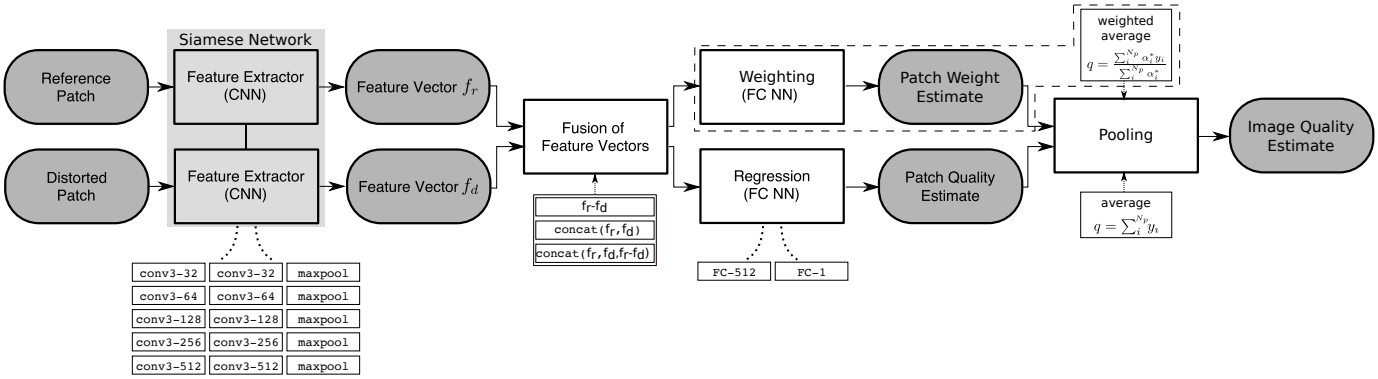


Fig. 2: Graphical representation of the deep neural network model extended for FR IQA. Features are extracted from the distorted patch and the reference patch by a CNN. Feature vectors f_d and f_d are fused either by difference to $f_r - f_d$, by concatenating f_d and f_d or by concatenating f_d and f_d and $f_r - f_d$. The pooled feature vector is regressed to a patchwise quality estimate. As in Fig. 1 patchwise estimates are aggregated to a global image quality estimate. The dashed-boxed branch of the network indicates an optional additional regression of the feature vector to a patchwise weight estimate that allows for pooling by weighted-average patch aggregation.

be $512 \cdot N_p$. Dimensionality reduction on the features (e.g. by PCA) can reduce this number even more. That way, a network trained for FR IQA can be used as a NR IQA method. In this extreme case the reference would be reduced to a dimensionality of 1. This could be done by reducing f_r to the mean of the reference feature vector observed in the training data and allow to use it for IQA.

E. Training

The proposed networks are trained in an iterative process 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 the patchwise method, 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 lead to the effective batch size of 128 patches. For FR training the respective reference patches are included in the mini-batch. The patches are randomly sampled every epoch to ensure that many different image patches are used in training.

The learning rate for the batchwise optimization using backpropagation is controlled per-parameter adaptively using the ADAM method [26] based on the variance of the gradient. The set of all weight parameters θ in the network is updated with

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{7}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{8}$$

$$\theta = \theta_{t-1} - \alpha \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{9}$$

The parameters are chosen as recommended in [26] as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $\alpha = 10^{-4}$. $g_t = \nabla_\theta f_t(\theta)$ denotes the gradient w.r.t θ at a time step t. The update of θ is done for each mini-batch until all images of the training set have been processed. The mean loss over all images in 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 and only the latest model and the one that produced the best validation loss are saved. The final model used in evaluation is the one with the best validation loss. This amounts to early stopping [27], a regularization to prevent overfitting. Note that the two regression branches estimating patch weight and patch quality do not have identical weights. This is because the update of the weight parameters is calculated based on gradients respect to different dimensions of parameters.

IV. Experiments and Results

A. Datasets

Experiments are performed on the LIVE [28], TID2013 [29] and CSIQ [30] image quality databases. The LIVE [28] database comprises 779 quality annotated images and is based on 29 source reference images, subject to 5 different types of distortions at different distortion levels. Distortion types are JPEG2000 compression, JPEG compression, additive white Gaussian noise, Gaussian blur and a simulated fast fading Rayleigh channel. Quality ratings were collected based on a single-stimulus methodology. Scores from different test sessions were aligned and the resulting DMOS quality ratings lie in a range of $[0\ 100]$, where a lower score indicates better visual image quality.

The TID2013 image quality database comprises 3000 quality annotated images. These are 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 double stimulus manner, during which the observers were presented a reference image and two distortion versions simultaneously and the observer was asked to choose the image of higher visual quality. The chosen image won one point and points assigned to each image were accumulated to the final quality score. Each distorted image was presented in nine comparisons, so the obtained MOS values lie in a range $[0\ 9]$, where larger MOS indicate better visual quality.

The CISQ image quality database contains 866 quality annotated images. The 30 reference images are distorted by JPEG compression, JPEG2000 compression, Gaussian blur, Gaussian white noise, Gaussian pink noise or contrast change. For quality assessment, all distorted versions of each reference image were presented on a monitor array and subjects were asked to position these distorted images horizontally according to the visual quality. The horizontal position was then taken as the quality rating. After the alignment and normalization the DMOS values span the range $[0\ 1]$, where a lower value indicates better visual quality.

B. Experimental Setup

For evaluation, the networks are trained either on the LIVE or the TID2013 database. For cross-validation the 29 reference images (and the respective distorted versions) in LIVE are randomly split into 17 training images, 6 validation images and 6 test images. The 25 reference images in TID2013 are analogously split in 15, 5, 5 training, validation and test images, respectively. Thus, no version of an image, e.g., in the test set has been seen by the network during training or validation. Results are reported based on 10 random splits. Models are trained for 3000 epochs. During training the network has seen $\sim$48M patches in the case of LIVE and $\sim$178M patches in the case of TID2013.

To analyze the generalization ability of the proposed method, the CSIQ image database is used for cross-dataset evaluations. For this, models are trained either on LIVE or on TID2013 and tested on CSIQ. For cross-dataset evaluations the LIVE dataset is split into 23 reference images for training and 6 reference images for validation. The TID2013 database is split analogously to 20 training images and 5 validation images. As LIVE and TID2013 have a lot of reference images in common, tests between these two are unsuitable for evaluating generalization for unseen images. However, testing models trained on LIVE with the images of TID2013 can be used to determine how well a model deals with distortions that have not been seen in training. Performance on unseen distortions is important to determine whether a method is truly non-distortion-specific or just many-distortion-specific.

In order to make errors and gradients for LIVE and TID2013 comparable, the MOS values of TID2013 have been inverted and scaled to the same range as the DMOS values in LIVE. For cross-database evaluation the DMOS values of the CSIQ are scaled to the same range as well.

We evaluate the proposed methods by three different commonly use metrics: Prediction accuracy is quantified by Pearson linear correlation coefficient (LCC) and MSE, prediction monotonicity is measured by Spearman rank order coefficient (SROCC). For both correlation metrics a value close to 1, for MSE a value close to 0 indicates high performance of a specific quality measure.

C. No-Reference Image Quality Assessment

TABLE I: Comparison of different NR IQA methods based on the LIVE and TID2013 databases. For further reference, in the first four rows performances of four prominent FR IQA methods are reported. The highest LCC and SROCC for the NR IQA methods are set in bold. The reported correlations are achieved on the test sets of 10 random train-test splits.

	LIVE		TID2013	
Method	LCC	SROCC	LCC	SROCC
PSNR	0.856	0.866	0.675	0.687
SSIM [9]	0.906	0.913	0.790	0.742
FSIM$_C$ [10]	0.961	0.965	0.877	0.851
DIIVINE [13]	0.917	0.916	0.654	0.549
BLIINDS-II [14]	0.930	0.931	0.628	0.536
BRISQUE [15]	0.942	0.940	0.651	0.573
CORNIA [16]	0.935	0.942	0.613	0.549
QAF [18]	0.953	0.948	0.662	0.589
CNN [19]	0.953	0.956	-	-
SOM [17]	0.962	**0.964**	-	-
Patchwise (proposed)	**0.972**	0.960	**0.855**	**0.835**
Weighted (proposed)	0.963	0.954	0.787	0.761

1) Single Dataset Evaluations: As the proposed methods estimate image quality based on estimates of local patch quality, their performance depends on the number of patches N_p considered and better performance can be expected for a larger number or patches. This is confirmed by Fig. 3, where the average LCC, SROCC and MSE over 10 random splits is shown for the LIVE test set (top row) and the TID2013 test set (bottom row). The three performance metrics are almost perfectly in agreement and show a clear ranking of the considered pooling methods on both test sets. For both pooling methods and on both datasets all three performance metrics improve monotonically with increasing number of patches N_p until saturation. On LIVE, with only one randomly sampled patch an average linear correlation can be achieved and saturation sets in at about $N_p \approx 16$ to reach its maximal performance, whereas the model employing weighted average patch aggregation 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 low importance rating (i.e., spatial attention mechanism). As the bottom row of Fig. 3 shows, qualitatively the same results are obtained on TID2013. Generally, the achieved correlations on TID2013 are significantly lower than on LIVE over the whole range of N_p. This is expected as IQA on TID2013 is much harder with 24 different distortion types instead of only 5 for LIVE. Again the weighted average patch aggregation performs worse than the simpler alternative, despite some local distortions being present in the dataset. As for the LIVE database, $N_p \approx 16$

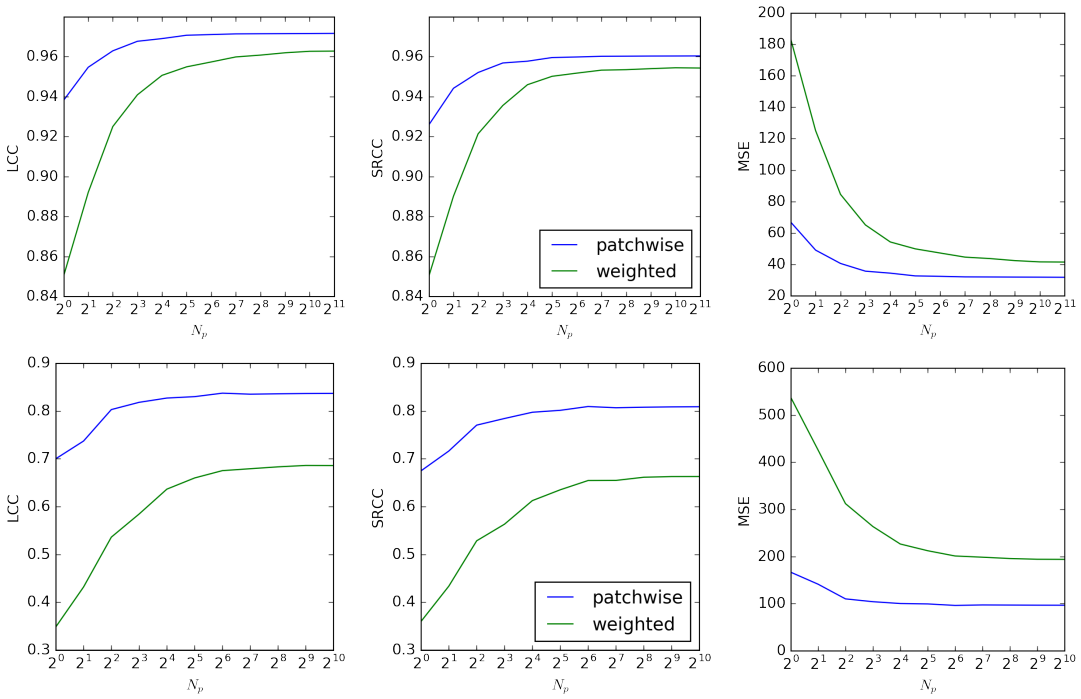


Fig. 3: Average performance of the proposed CNN for NR IQA in terms of LCC, SROCC and MSE in dependence of the number of randomly sampled patches for LIVE (top row) and TID2013 (bottom row).

randomly sampled patches are sufficient for achieving saturation and maximum performance of the patchwise method. Based on these findings the number of patches considered for quality estimation is $N_P = 32$ for further evaluation.

Following the training and testing protocol outlined in Section III-E, the results achieved by the proposed method on LIVE and TID2013 as summarized in comparison to state-of-the-art general purpose NR IQMs (DIIVINE [13], BLINDS-II [14], CORNIA [16], QAF [18], CNN [19] and SOM [17]) in Table I. To allow for a better assessment of the results, performances of popular FR IQA methods PSNR, SSIM, FSIM_C and HaarPSI are also tabularized. For both pooling schemes, the proposed approach performs better than competing methods in terms of LCC on both datasets. In terms of SROCC the proposed methods show competitive performance, only outperformed by SOM [17] on the LIVE dataset[3]. On LIVE, all state-of-the-art NR IQMs perform comparable or superior to PSNR or SSIM. The proposed NR IQM methods even outperform FSIM_C and HaarPSI in terms of LCC with both proposed pooling techniques. On TID2013, none of the state-of-the-art NR IQMs used for comparison outperforms any of the FR IQMs, whereas again the proposed methods achieve performances superior or comparable to PSNR and SSIM and the patchwise model even comes within reach of the state-of-the-art FR IQM FSIM_C and HaarPSI.

2) Cross-Dataset Evaluations: In order to address the generalization ability of the proposed methods we perform a cross-dataset evaluation in five different experimental settings: In the first two experiments, a model trained on the full LIVE database is evaluated on subsets of CSIQ and TID2013. These subsets only contain the four distortions shared between LIVE and CSIQ, and LIVE and TID2013, respectively, being JPEG compression, JPEG2000 compression, Gaussian blur and white noise. The patchwise method is still able to

[3]Unfortunately, no results are reported for the performance of SOM on TID2013

TABLE II: SROCC results of the cross-dataset evaluations with comparable results taken from [18] and [17]. All models are trained on the full LIVE dataset and evaluated on CSIQ and TID2013. The subsets of CSIQ and TID2013 contain only the 4 distortions that are shared with LIVE.

	subset		full	
Method	CSIQ	TID2013	CSIQ	TID2013
DIIVINE [13]	-	-	0.596	0.355
BLIINDS-II [14]	-	-	0.577	0.393
BRISQUE [15]	0.899	0.882	0.557	0.367
CORNIA [16]	0.899	0.892	0.663	0.429
QAF [18]	-	-	0.701	0.440
CNN [19]	-	0.920	-	-
SOM [17]	-	**0.923**	-	-
Patchwise (proposed)	**0.908**	0.867	0.681	0.392
Weighted (proposed)	0.866	0.872	**0.704**	**0.462**

TABLE III: SROCC results of the cross-dataset evaluations with comparable results taken from [31]. All models are trained on the full TID2013 dataset and evaluated on CSIQ.

Method	CSIQ full
DIIVINE [13]	0.146
BLIINDS-II [14]	0.456
BRISQUE [15]	0.639
CORNIA [16]	0.656
Patchwise (proposed)	0.717
Weighted (proposed)	**0.733**

outperform BRISQUE and CORNIA in this experiment (see 2nd column in Table II) for the CISQ subset. Unfortunately, no results are available for the other state-of-the-art approaches. As shown in the 3rd column of Table II, for the subset of TID2013, the proposed approaches perform worse than the other state-of-the-art methods.

In the other two experiments, a model trained on the full LIVE database is tested on the full CISQ and TID2013 databases. The results obtained on the full CISQ are reported in the fourth column in Table II. The two unseen distortions (i.e. frequency noise and contrast change) are considerably different in their visual results compared to the other four. As such, it is not surprising that all compared IQA methods perform worse in this setting. Despite performing worse on the single dataset experiments, the weighted model adapts better to unseen distortions than the patchwise model. Both models perform better than all of the compared state-of-the-art approaches.

The fourth cross-dataset experiment is based on the ambitious task of predicting MOS values from the complete TID2013 dataset using a model trained on LIVE. With only 4 out of 24 distortions being represented in the training set, this is a particularly hard challenge. 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 dataset is hard or even impossible. Nevertheless, the proposed methods obtain competitive results. Again, the weighted method shows the best performance on unseen distortions.

In a fifth cross-dataset experiment, all models are trained on the full TID2013 dataset and tested on the full CSIQ dataset. Results are shown in Table III. DIIVINE, BLIINDS-II and CORNIA decrease their performance compared to the models trained on LIVE, despite TID2013 being the larger and more diverse training set. Comparing the third column of Table II to Table III reveals that and the proposed methods (as well as BRISQUE) can make use of the larger training size and shows an improved SROCC. 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 even 4 reference images fewer. Thus, training on TID2013 has the same short-comings as training on LIVE when it comes to adaption to unseen images.

All learning-based IQA methods face the challenge of learning about the statistics of natural images with a training set of only 29 references. Some methods like SOM [17] try to solve this through unsupervised learning on unlabeled training data.

D. Full Reference Image Quality Assessment

1) Feature Fusion: In contrast to the proposed NR IQA methods, the FR IQA employs a feature fusion step. Thus, we begin the evaluation of the proposed FR IQA approach with a design choice regarding the feature fusion methods presented in Section III-D. The performances of the three feature fusion schemes are reported for LIVE and TID2013 in Table IV. Mere concatenation of both feature vectors does not fail but it consistently performs worse than the two feature fusion methods that exploiting the explicit difference of both feature vectors. This shows that while the model is able to learn the relation between the two feature vectors, providing that relation explicitly does lead to better results. The results do not provide enough evidence for preferring one over the other two feature fusion method. This might suggest that adding the original feature vectors to the representation does not add useful information. Despite the inconclusive results, $\texttt{concat}(f_r, f_d, f_d - f_r)$ is used for further evaluations.

TABLE IV: LCC results for each combination of the two proposed pooling methods and the three proposed feature fusion schemes. The LCC was computed on the validation set of one specific split for each dataset and with $N_p = 1024$ random patches per image.

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

TABLE V: Comparison of different FR IQA methods based on the LIVE and TID2013 databases. The highest LCC and SROCC are set in bold. The reported correlations are achieved on the test sets of 10 random train-test splits.

	LIVE		TID2013	
Method	LCC	SROCC	LCC	SROCC
PSNR	0.856	0.866	0.675	0.687
SSIM [9]	0.906	0.913	0.790	0.742
FSIM$_C$ [10]	0.961	0.965	0.877	0.851
DOG-SSIM [32]	0.963	0.961	0.919	0.907
HaarPSI[11]	0.967	0.968	0.87	0.863
CNNM [12]	-	-	-	0.93
Patchwise (proposed)	0.977	0.966	0.880	0.859
Weighted (proposed)	**0.980**	**0.970**	**0.946**	**0.940**

TABLE VI: Prediction monotonicity of different FR IQM for the subsets of TID2013. The highest SROCC are set in bold. The reported correlations are the average correlation achieved on the test sets of 10 random train-test splits.

	Noise	Actual	Simple	Exotic	New	Color
PSNR	0.822	0.825	0.913	0.597	0.618	0.535
SSIM [9]	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 [32]	0.922	0.933	0.959	0.889	0.908	0.911
Patchwise (proposed)	0.938	0.923	0.885	0.771	0.911	0.899
Weighted (proposed)	**0.969**	**0.970**	**0.971**	**0.925**	**0.941**	**0.934**

2) Single Dataset Evaluations: As well as for the NR IQA methods, the performance of the proposed FR IQA methods depends on the number of patches N_p used for evaluation. This behavior is plotted in Fig. 4 for the models trained and tested on LIVE (top row) and TID2013 (bottom row). As for the NR IQA approach, all three metrics are showing a monotonic increase towards saturation of performance for a larger N_p.

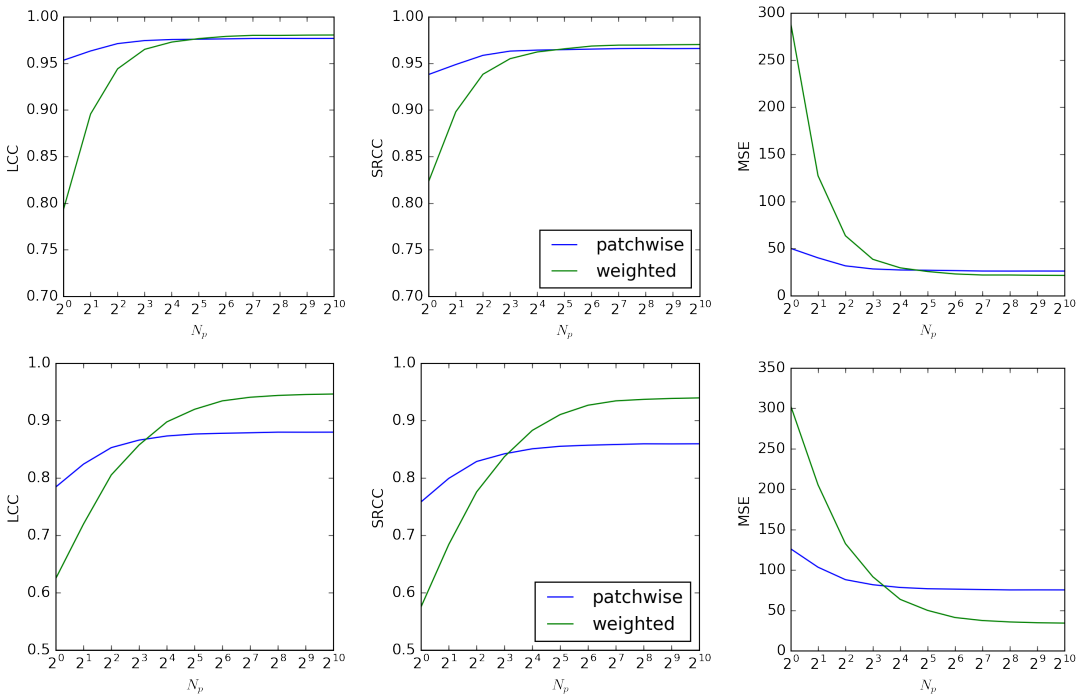


Fig. 4: Performance of the proposed CNN for FR IQA in terms of LCC, SROCC and MSE in dependence of the number of randomly sampled patches evaluated on LIVE (top row) and TID2013 (bottom row).

In contrast to the NR IQA models, in the FR setting the weighted average patch aggregation improves the predictions over the simple patchwise approach across all three evaluation metrics. The weighted model saturates at the maximum performance with $N_p \approx 32$, whereas the unweighted model saturates already at $N_p \approx 16$. The proposed patch aggregation scheme improves performance in the FR setting, but not in the NR setting. This suggests that the reference image is particularly important for estimating relative patch importance. The fact that the two proposed models saturate at almost the same performance on the LIVE database might suggest that these results are very close to the achievable maximum. The difference between the proposed methods is more pronounced in the TID2013 experiment shown in the bottom row of Fig. 4. Here, the models using weighted average patch aggregation perform considerably better than the simple patchwise model when $N_p > 8$ and saturates at a higher maximum performance. In the NR setting this effect of increasing the performance by weighted average patch aggregation does not show in the results. This suggests that the estimation of patchwise weights mostly relies on information from the reference, but not from the distorted image.

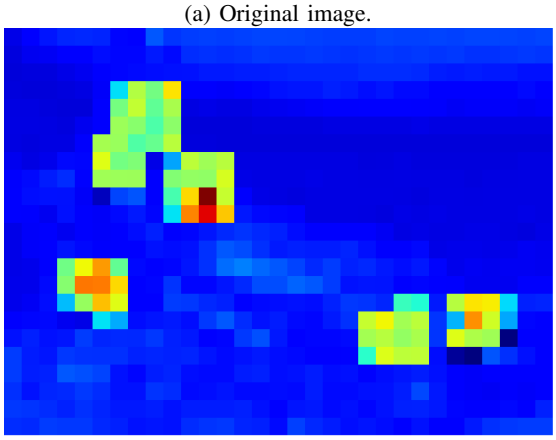
Table V summarizes the performance of the proposed FR IQM in comparison to state-of-the-art methods for LIVE and TID2013. On both datasets the proposed method applying weighted average patch aggregation obtains correlations superior to state-of-the-art methods. The superior performance of the weighted average patch aggregation compared to simple averaging is exemplified in Fig. 5. In this example most of the image is undistorted but for some blocks the content is replaced by a constant value. The weights α_i are close to zero for undistorted patches, making them far less important for the final decision. For the model applying a patchwise unweighted average, the distortions of the different patches contribute equally to the final decision. This leads to an underestimation of the DMOS at 28.42, whereas the ground truth DMOS of the image is 47. The weighted average patch aggregation however compensates for the spatially unequal distribution of patchwise quality and thus achieves a more accurate estimate of the overall quality.

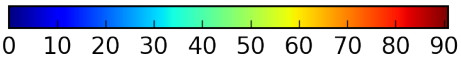
This effect can also be studied by evaluating different groups of distortions separately. In [29] the distortions of the TID2013 dataset are divided into the groups *Noise*, *Actual*, *Simple*, *Exotic*, *New* and *Color*. Table VI lists the prediction monotonicity in terms of SROCC for these groups and different state-of-the-art FR IQM. The patchwise model has a weak point in the *Exotic* subset, probably because it contains the local distortions present in the dataset. The weighted model consistently outperforms other methods on all subsets. Unsurprisingly, the *Exotic* subset is still the most challenging part of the test set. But the proposed method is not only superior in handling difficult distortions. The high SROCC of 0.97 in the *Actual* subset shows that it is well suited to predict the perceptual effects of distortions that appear in relevant application domains.

3) Cross-Database Performance: Analogously to the NR IQA methods in Section IV-C2 the generalization ability of the proposed FR IQA methods are evaluated by training and testing on different datasets. Table VII shows the results for models trained on LIVE and tested on TID2013 and CSIQ, and models trained on TID2013 and tested on LIVE and CSIQ. Results are compared to DOG-SSIM. In all four settings the

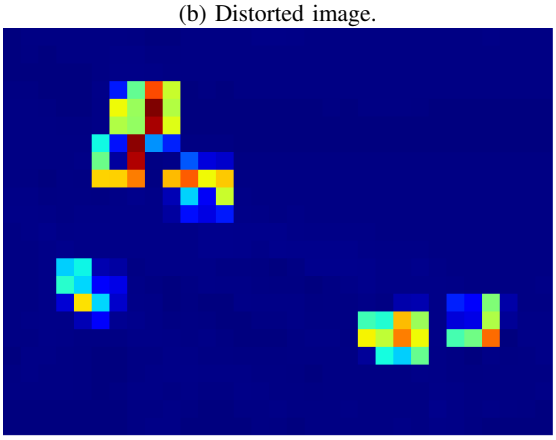
(a) Original image.

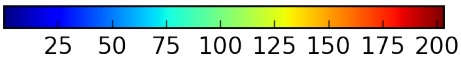
(b) Distorted image.



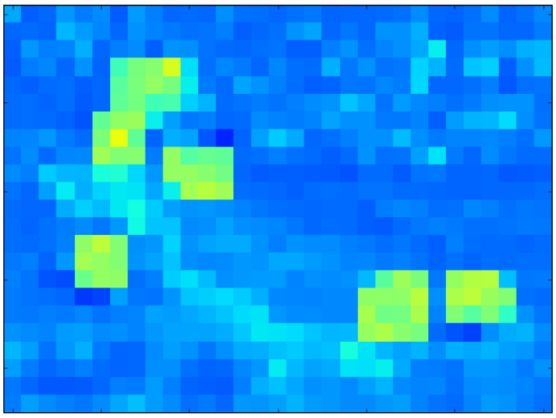


(c) Local patchwise distortion estimate y_i from weighted model.





(d) Local patchwise weight α_i^* from weighted model.



(e) Local patchwise distortion estimate y_i from non-weighted model.

Fig. 5: Example image with local distortion from TID2013 test set with a ground truth DMOS value of 47. Below the reference and distorted images, the network outputs of the weighted model y_i and α_i are shown for each image patch. The weighted average prediction is 45.87. For comparison the network output for the simple patchwise model is shown below. It underestimated the DMOS at 28.42, because it gives the same weight to each patch.

patchwise model shows insufficient generalization capabilities. The weighted model shows the best generalization among the two pooling schemes and performs comparable to DOG-SSIM in these four experiments. The two experiments on the CSIQ dataset show that increasing the training size leads to better generalization. Generalization could probably improved with an even larger training dataset with a larger set of more diverse reference images.

4) Network Depth: The comparison of the proposed NR IQA approach to [19] in Section IV-C suggests that the

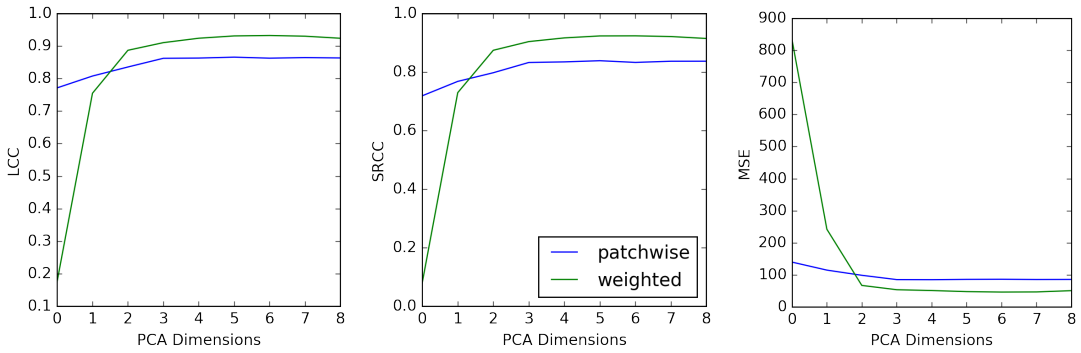


Fig. 6: Average performance on one TID2013 test set of the proposed methods for FR IQA in terms of LCC, SROCC and MSE in dependence of the number of principal components of the reference patch feature vector. ($N_p = 32$)

TABLE VII: Comparison of prediction monotonicity in cross-dataset experiments. 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 [32]	**0.751**	**0.914**	0.925	**0.948**
Patchwise (proposed)	0.437	0.660	0.863	0.796
Weighted (proposed)	**0.751**	0.909	**0.931**	0.936

performance of a neural network-based IQM can be increased by adding layers and making the network deeper. In order to evaluate this observation in a FR context as well, the architecture of the 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. For simplicity only the best performing pooling method employing weighted average patch aggregation is compared. When tested on one split of the LIVE dataset, 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, which is lower than the SROCC of 0.927 when using the original architecture. 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.

5) Reducing the Reference Information: RR IQA only requires a limited number of features extracted from the reference images. As such, it is conceptually living between NR and FR IQA. Although there are studies analyzing the influence of the numbers of used features on the performance of RR IQM [33], commonly NR and FR approaches are not studied in an unified framework. The proposed neural network-based methods allow for such an unified evaluation. A straight forward approach to do so would be to systematically change the network architecture from FR, see Fig. 2, to NR, see Fig. 1, by reducing dimensionality of the number of channels in the branch of the CNN extracting the features from the reference image. However, this approach requires models specifically trained for each number of reference 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). This would not require retraining, but allow for a systematic dimensionality reduction that would lead to NR IQA in the extreme case when only the first principal component would be considered. Following this idea, the PCA is estimated based on the feature vectors of 4000 reference patches sampled from the training set and used for dimensionality reduction of f_r during testing. Fig. 6 shows the performance of this RR IQM on one TID2013 test split for increasing dimensionality of the reference patch feature vectors. The unweighted pooling method is still able to make useful predictions even without any reference information. This is not the case for the weighted pooling method, where more information about the reference image is needed. This supports the previous conjecture that weighted average patch aggregation, i.e. reliable estimation of the weights, is depending on information from the reference image, but not from the distorted image. For both pooling methods 3 principal components (dimensions) are already enough to recover the performance obtained with the 512-dimensional original feature vector.

V. Conclusion

The paper presented a neural network-based approach to FR and NR IQA. For this, novel network architectures were presented. A weighted average patch aggregation method was proposed for improving the pooling from local patch quality 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 data-driven IQA methods the generalization offer considerable room for improvement.

As the generalization performance shows, 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. In future work, unsupervised pre-training might also be a way to tackle this problem. Even though a relative generic neural network is able to achieve high prediction performance, incorporating IQA specific adaptations to the architecture potentially lead to further improvements. Our results show that there is room for optimization in terms of feature dimensionality.

In conjunction with visualization techniques such as layer-wise relevance propagation (LRP) [34], neural network-based IQA methods could be used to learn about local features and their spatial relation driving image quality perception. In this context it is important to gain a better understanding about the weighting maps output from the network and evaluate conceptual similarities with attention and saliency models.

From application-oriented perspective the proposed method should be adapted and evaluated for quality assessment of 3D images and 2D and 3D videos. The performance of the presented approach and, given the fact that no domain knowledge is necessary, its relative simplicity suggests that neural network-based approaches to IQA will be relevant for future research.

REFERENCES

[1] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnstrom, "Psychophysiology-Based QoE Assessment: A Survey," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2016.

[2] S. Bosse, K.-R. Müller, T. Wiegand, and W. Samek, "Brain-Computer Interfacing for Multimedia Quality Assessment," in *Proceedings of the Workshop in Brain-Machine Interface Systems at IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2016.

[3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ImageNet Challenge*, pp. 1–10, 2014.

[4] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a Siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.

[5] S. Chopra, R. Hadsell, and L. Y., "Learning a similiarty metric discriminatively, with application to face verification," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 349–356, 2005.

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

[7] 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.

[8] R. Dosselmann and X. D. Yang, "A comprehensive assessment of the structural similarity index," *Signal, Image and Video Processing*, vol. 5, no. 1, pp. 81–91, nov 2009.

[9] 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.

[10] 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.

[11] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A Haar Wavelet-Based Perceptual Similarity Index for Image Quality Assessment," *CoRR*, vol. abs/1607.0, 2016.

[12] V. V. Lukin, N. N. Ponomarenko, O. I. O. Ieremeiev, K. O. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," in *SPIE/IS&T Electronic Imaging*, vol. 9394. International Society for Optics and Photonics, 2015, pp. 93 940K—-93 940K.

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

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

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

[16] 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.

[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. 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, 2014.

[19] 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.

[20] 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.

[21] R. Girshick, "Fast R-CNN," in *The IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[23] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proceedings of the 27th International Conference on Machine Learning*, no. 3, pp. 807–814, 2010.

[24] 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.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

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

[27] L. Prechelt, "Early stoppingbut when?" in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 53–67.

[28] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms." *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 15, no. 11, pp. 3440–51, nov 2006.

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

[30] E. C. Larson and D. M. Chandler, "Consumer subjective image quality database," 2009.

[31] L. L. Zhang, L. L. Zhang, S. Member, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[32] S.-c. Pei, L. Fellow, and L.-h. Chen, "Image Quality Assessment Using Human Visual DOG Model Fused With Random Forest," *Image Processing, IEEE Transactions on*, vol. 24, no. 11, pp. 3282–3292, 2015.

[33] R. Soundararajan and A. C. Bovik, "RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 517–526, 2012.

[34] 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.