Deep Learning of Human Visual Sensitivity in Image Quality Assessment Framework

Jongyoo Kim Sanghoon Lee*
Department of Electrical and Electronic Engineering, Yonsei Universiy, Seoul, Korea
{jongky, slee}@yonsei.ac.kr

Abstract

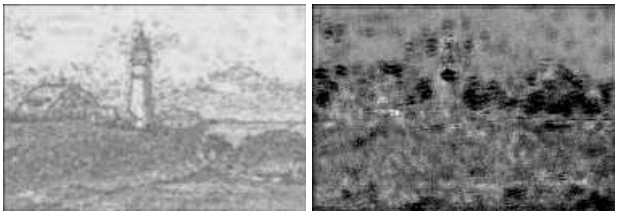
Since human observers are the ultimate receivers of digital images, image quality metrics should be designed from a human-oriented perspective. Conventionally, a number of full-reference image quality assessment (FR-IQA) methods adopted various computational models of the human visual system (HVS) from psychological vision science research. In this paper, we propose a novel convolutional neural networks (CNN) based FR-IQA model, named Deep Image Quality Assessment (DeepQA), where the behavior of the HVS is learned from the underlying data distribution of IQA databases. Different from previous studies, our model seeks the optimal visual weight based on understanding of database information itself without any prior knowledge of the HVS. Through the experiments, we show that the predicted visual sensitivity maps agree with the human subjective opinions. In addition, DeepQA achieves the state-of-the-art prediction accuracy among FR-IQA models.

1. Introduction

Predicting perceptual quality is the main goal of image quality assessment (IQA), which is applied in the wide field of image processing such as process evaluation, image and video encoding, and monitoring. Since human observers are the ultimate receivers of digital images and videos, quality metrics should be designed from a human-oriented perspective. Therefore, a great deal of effort has been made to develop IQA methods based on the analysis of the properties and mechanism of the human visual system (HVS).

When a distorted image is perceived by HVS, some error signals are emphasized and some others are masked. Figs. 1(a) and (b) show a distorted image by JPEG and its objective error map. The distortions around the houses and on the sky regions are easily observable. However, those on textural regions (e.g. rocks) are less noticeable though there are a lot of pixel-wise distortions as shown in Fig. 1(b). Therefore, simple pixel-wise metrics such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE) do not correlate well with perceived quality. To conduct reliable IQA, it is necessary to understand the *human visual sensitivity* which explains the perceptual impact of artifacts according to a spatial characteristic of pixels.

(a)



(b) (c)

Figure 1. Examples of predicted sensitivity maps: (a) is a distorted image; (b) is an objective error map; (c) is a predicted perceptual error map. Darker regions in (b) indicate more pixel-wise distorted pixels, and those in (c) indicate perceptually more distorted ones.

Based on these observations, many full-reference image quality assessment (FR-IQA) methods have adopted various computational models of the HVS from psychological

*Corresponding author. (E-mail: slee@yonsei.ac.kr)

This work was supported by the ICT R&D program of MSIP/IITP. [2017-0-00289, Development of a method for regulating human-factor parameters for reducing VR-induced sickness]

vision science [4], and made assumptions of the HVS's behavior to predict perceptual quality [32, 34, 35]. However, since the majority of the HVS models are complex and were designed in a limited and refined condition, it is difficult to assure the best performance by generalizing the HVS models to the practical IQA problem.

Recently, convolutional neural networks (CNN) have been the widely used in computer vision [11]. Beyond the classification framework, CNNs have been successfully used to generate image maps such as semantic segmentation map [16] and depth map [6]. Inspired by these works, we use the CNN to generate a visual sensitivity map which refers to a weighting map of describing the degree of visual importance of each pixel to the HVS.

The CNN model in our approach is dedicated to learn the HVS properties. Based on the objective error map, the model seeks the visual weight of each pixel. The predicted visual sensitivity map allocates local weights to the pixels according to their local spatial characteristics of the distorted images. This approach is similar to weighted pooling strategies adopted in FR-IQA methods [32, 35]. However, different from the previous works, our model finds a visual weight without any prior knowledge of the HVS, but relying only on the dataset: a triplet of a distorted image, its objective error map, and its ground-truth subjective score. Fig. 1(c) shows an example result of the proposed model. The dark regions indicate perceptually distorted pixels. Compared to the objective error map in (b), it is obvious that (c) emphasizes the visible distortions such as coding artifacts above the sea and around the house.

We name the proposed method as Deep Image Quality Assessment (DeepQA). Our contributions can be summarized as follows.

1. DeepQA learns the visual sensitivity characteristics of the HVS without any prior knowledge. By using a deep CNN, the visual weight of each pixel is sought by using a triplet of a distorted image, its objective error map, and its ground-truth subjective score.

2. DeepQA can generate the perceptual error map as an intermediate result, which provides us an intuitive analysis of local artifacts for given distorted images.

3. A novel deep CNN based FR-IQA framework is proposed. Our model can be trained by end-to-end optimization, and achieves state-of-the-art correlation with human subjective scores.

2. Related Work

2.1. Human Visual Sensitivity

A number of computational models of human perception have been developed in the literature. Luminance adaptation describes that the visibility threshold of the HVS is determined by background luminance, which is similar to Weber-Fechner law [3]. Contrast sensitivity function (CSF) refers to the varying sensitivity of the HVS according to spatial frequency of images [4]. The presence of texture decreases the visibility of image distortion, which is called contrast masking [13, 4]. In contrast, the coding artifact on homogeneous regions is easier to be observed. To imitate image representation in the visual cortex, sub-band decomposition models such as Gabor filters and steerable pyramids have been adopted [29].

Based on these observations, many FR-IQA metrics have been developed. There are two general strategies for FR-IQA: bottom-up and top-down frameworks. The former simulates the various processing stages in the HVS utilizing the computational models directly which is described earlier [4]. Meanwhile, the latter designs the overall function of IQA based on assumptions on the HVS. For example, structural similarity index (SSIM) assumed that contrast and structural distortions are critical to the HVS [32]. Feature similarity index (FSIM) assumed that phase congruency is the primary feature for the HVS perception [35].

2.2. Machine Learning Approach on IQA

Machine learning was mostly adopted in no-reference image quality assessment (NR-IQA). Since the reference images are not available in NR-IQA, researchers tried to design elaborate features which can discriminate distorted images from their pristine images. One of the popular features is a family of natural scene statistics (NSS) which assumes that natural scenes contain statistical regularity [26, 20]. Beyond the NSS, various kinds of features were developed for NR-IQA [31, 8].

On the other hand, machine learning has been adopted partially in FR-IQA. In [22], singular value decomposition features were extracted and regressed onto the quality score using support vector regression (SVR). Multi-method fusion (MMF) proposed to combine the multiple existing FR-IQA methods using machine learning to achieve a state-of-the-art accuracy [15]. In [23], multiple features were extracted from difference of Gaussian frequency bands, and regressed onto the quality score.

Relatively recently, there have been attempts to adopt deep learning for the NR-IQA problem. Hou *et al.* used a deep belief network where wavelet NSS features were extracted and fed into the deep model [7]. Kang *et al.* first applied a CNN to the NR-IQA problem without using any handcrafted features [8]. Kim and Lee described a two-stage CNN-based NR-IQA model, where local quality scores generated by a FR-IQA method were used as proxy patch labels [9]. Liang *et al.* [14] proposed a dual-path CNN-based FR-IQA model, where a non-aligned image of a similar scene as reference are also deal with.

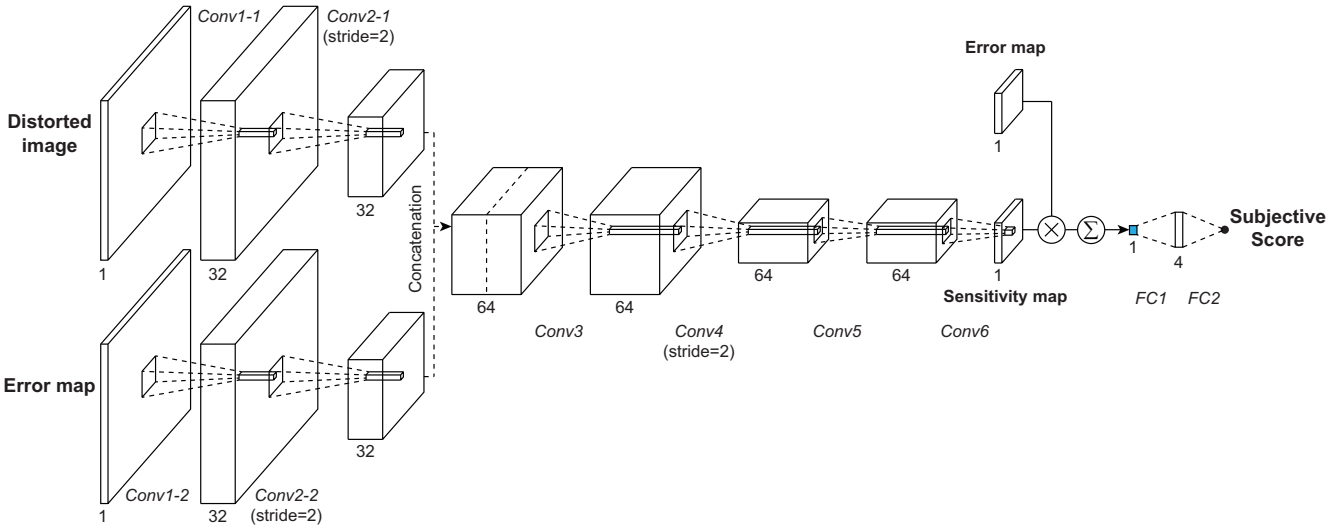


Figure 2. Architecture of DeepQA. The model takes as input a distorted image and an error map, and generates the sensitivity map. After multiplication with the error map, it is regressed onto the subjective score.

3. Sensitivity Map Prediction

3.1. Architecture

On deciding the human visual sensitivity, the most intuitive way is comparing the energies of an error signal and its background signal, where the error signal indicates the objective error map and the background signal indicates the reference image. However, in a real world situation, the HVS observes only a distorted image without knowing the error signal. Therefore, we test two scenarios in this paper. First, *DeepQA* takes both a distorted image and an error map as inputs. Second, *DeepQA–s* is a simpler version, where only a distorted image is used as an input. The architectures of DeepQA is shown in Fig. 2. DeepQA–s contains only branch from input distorted images (*Conv1-1* and *Conv2-1*), and there is no concatenation layer.

Deep convolutional networks with 3×3 filters are used for both models inspired by the recent work [30]. To generate a sensitivity map without losing the pixel position information, the model consists of only convolutional layers. In DeepQA, the distorted image and the error map go through different convolutional layers at the beginning, and are concatenated after the second convolutional layer. To preserve the feature map size after convolution operations, zeros are padded around the border before each convolution. Two strided convolutions are used for subsampling, as shown in Fig. 2. Therefore the final output is $1/4$ of that of the original input image, and the ground-truth objective error maps are downscaled by $1/4$ correspondingly. In both DeepQA–s and DeepQA, each convolutional layer except *Conv6* employs a leaky rectified linear unit (LReLU) [18]. The *Conv6* layer adopts a rectified linear unit (ReLU) [21] since the weights are positive real numbers. In addition, the bias of *Conv6* is initialized to 1. At the end of the model, two fully connected layers are used to regress features onto the subjective scores. Here, LReLU and ReLU are used for hidden and output layers respectively.

3.2. Image Normalization

The distorted images are simply normalized before they are fed into the CNN. Let I_r be a reference image, and I_d be a distorted image. We first convert them into grayscale images, and rescale them to the range [0, 1]. Then the low pass filtered versions of them are subtracted. The normalized images are denoted by $\hat{I}_r$ and $\hat{I}_d$. This is because the HVS is not sensitive to the changes in low-frequency band. The contrast sensitivity function (CSF) shows a band-pass filter shape peaking at around 4 cycles per degree, and sensitivity drops rapidly at low frequency [4].

3.3. Sensitivity Map Prediction

To make the model learn to generate the sensitivity map, we utilize a triplet of a distorted image, its objective error map, and its corresponding ground-truth subjective score. Instead of just averaging the objective error map, the sensitivity map weights each pixel to reflect the HVS. Toward the end, we first define the objective error map using the normalized log difference function as

$$\mathbf{e} = \frac{\log\left(1/((\hat{I}_r - \hat{I}_d)^2 + \varepsilon/255^2)\right)}{\log\left(255^2/\varepsilon\right)} \tag{1}$$

where $\varepsilon = 1$ for the experiment.

The visual sensitivity map is obtained from the CNN

models

$$\mathbf{s}_1 = CNN_1(\hat{I}_d; \theta_1) \quad (2)$$

$$\mathbf{s}_2 = CNN_2(\hat{I}_d, \mathbf{e}; \theta_2) \quad (3)$$

where $CNN_1(\cdot)$ and $CNN_2(\cdot)$ indicate the CNN models of DeepQA–s and DeepQA with the parameters θ_1 and θ_2 respectively. Then the perceptual error map is defined by $\mathbf{p} = \mathbf{s} \odot \mathbf{e}$, where $\odot$ is the Hadamard product, and $\mathbf{s}$ is $\mathbf{s}_1$ or $\mathbf{s}_2$.

Since we padded zeros before each convolution, feature maps near the borders tend to be zeros. To alleviate this, we ignore the pixels near borders around the perceptual error map. Each four rows and columns for each border are excluded in the experiment, which can partially compensate the information loss. Therefore, the pooled score is derived by averaging the cropped perceptual error map as

$$\mu_{\mathbf{p}} = \frac{1}{(H-8)\cdot(W-8)} \sum_{(i,j)\in\omega} \mathbf{p} \quad (4)$$

where H and W are the height and width of $\mathbf{p}$, (i, j) indicates pixel index, and ω indicates the cropped region. Since it cannot be assured that the pooled score has a linear relationship with the subjective score, additional fully connected layers are used to conduct nonlinear regression. Then the final objective function is defined as

$$\mathcal{L}_s(\hat{I}_d; \theta) = \|(f(\mu_{\mathbf{p}}) - S)\|_F^2 \quad (5)$$

where $f(\cdot)$ is a nonlinear regression function, S is the ground-truth subjective score of the input distorted image, and θ is θ_1 for DeepQA–s or θ_2 for DeepQA.

3.4. Total Variation Regularization

When the model is optimized to minimize (5) without any constraints, it generates the sensitivity map which looks like high-frequency noise. To avoid this, a smoothing constraint is applied to the sensitivity map. We adopt a total variation (TV) L_2 norm, because it can penalize the high frequency component of the sensitivity map during the optimization of the CNN. Similar to [19], we define the TV regularization as

$$TV(\mathbf{s}) = \frac{1}{H \cdot W} \sum_{(i,j)} (sobel_h(\mathbf{s})^2 + sobel_v(\mathbf{s})^2)^{\beta/2} \quad (6)$$

where H and W are the height and width of the predicted sensitivity map, $sobel_h$ and $sobel_v$ are Sobel operation in horizontal and vertical directions respectively, and $\beta = 3$ in the experiment.

3.5. Training Method

For better convergence of the optimization, the adaptive moment estimation optimizer (ADAM) [10] with Nesterov momentum [5] was employed to alter the regular stochastic gradient descent method. The learning rate was initially set to 5×10^{-4}. To balance between the regression loss (5) and the TV regularization (6), we multiplied 10^3 and 10^{-2} to them respectively. The effects of TV regularization is further discussed in Section 4.2. In addition, L_2 regularization was applied to all layers (L_2 penalty multiplied by 5×10^{-3}).

3.5.1 Patch-based Approach

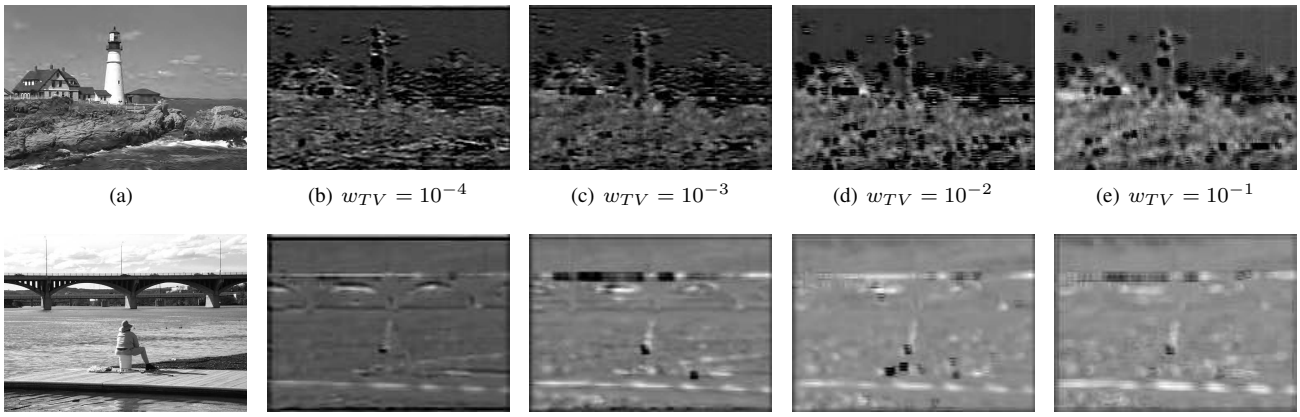
To train DeepQA on GPU, the sizes of input images need to be fixed. Therefore, to train the model using the LIVE IQA database [27], which contains images with various sizes, we divided the input images into patches with a fixed size. Here, it is necessary to avoid the overlapped regions when the perceptual error map is reconstructed. Therefore, the step of the sliding window is determined by $step_{patch} = size_{patch} - (N_{ign} \times 2 \times R)$ where N_{ign} is the number of ignored pixels, and R is the size ratio of the input and the perceptual error map. In the experiment with the LIVE IQA database, the ignored pixel was 4, the patch size was 112×112, and the sliding step was 80×80. In addition, during training stage, all the patches composing one image should be included in the same minibatch so that $\mu_{\mathbf{p}}$ in (4) can be derived from the reconstructed perceptual error map.

4. Experimental Results

4.1. Dataset

Four different IQA databases were used to evaluate the proposed algorithm: LIVE IQA [27], CSIQ [12], TID2008 [25], and TID2013 [24]. The LIVE IQA database contains 29 reference images and 982 distorted images with five distortion types: JPEG and JPEG2000 (JP2K), additive white Gaussian noise (WN), Gaussian blur (BLUR), and Rayleigh fast-fading channel distortion (FF). The CSIQ database includes 30 reference images and 866 distorted images of six distortion types: JPEG, JP2K, WN, GB, pink Gaussian noise (PGN), and global contrast decrements (CTD). TID2008 consists of 25 reference images and 1,700 distorted images with 17 different distortions at four levels of degradation, whereas TID2013 is expanded to contain 24 distortions types at five levels of degradation. In the experiment, the ground-truth subjective scores were rescaled to the range [0, 1]. For differential mean opinion score (DMOS) values (in LIVE IQA and CSIQ), the scale was reversed so that the larger values indicate perceptually better images.

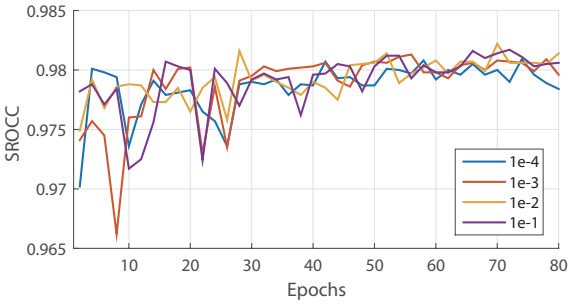
To evaluate the performances of the IQA algorithms, we used two standard measures, i.e., Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) by following [1].



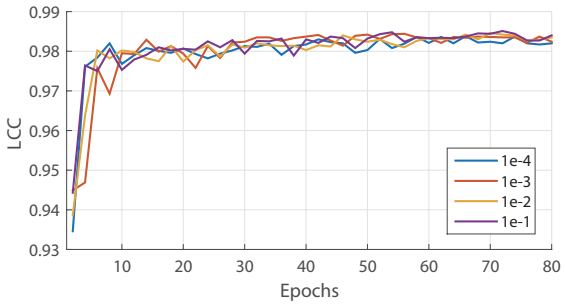
(a) (b) $w_{TV} = 10^{-4}$ (c) $w_{TV} = 10^{-3}$ (d) $w_{TV} = 10^{-2}$ (e) $w_{TV} = 10^{-1}$

(f) (g) $w_{TV} = 10^{-4}$ (h) $w_{TV} = 10^{-3}$ (i) $w_{TV} = 10^{-2}$ (j) $w_{TV} = 10^{-1}$

Figure 3. Examples of the predicted sensitivity maps with various TV regularization weights: (a) and (f) show distorted images; (b) - (e) and (g) - (j) are their predicted sensitivity maps with different TV regularization weights.



(a)



(b)

Figure 4. Comparison of SRCC and PLCC curves according to the degree of TV regularization weight.

4.2. Effects of TV Regularization

To analyze the effects of TV regularization, we tested with four different weights ($w_{TV} = 10^{-4}$, 10^{-3}, 10^{-2}, 10^{-1}) for the TV regularization during training. Fig. 3 shows the predicted sensitivity maps with different degrees of TV regularization weight. When the weight was very small, the sensitivity map was too detailed, which does not agree with the HVS well. As the weight w_{TV} increased (from (b) to (e), and from (g) to (j)), the sensitivity map tended to be smoother. In addition, it became clearly distinguishable between the perceptually less and more distorted regions as shown in (e). However, since the TV regularization promotes piecewise smoothing, black spots also increased as shown in (e).

To check if the TV regularization affects the prediction accuracy, SRCC and PLCC over 80 epochs with the four settings are drawn in Figs. 4(a) and (b). SRCC and PLCC were obtained from the testing set every 2 epochs. When $w_{TV} = 10^{-4}$, the SRCC and PLCC were slightly lower than the others, but there were no significant differences between the different degrees of TV regularization term.

4.3. Sensitivity Map Prediction

To validate if DeepQA agrees with the HVS, the predicted sensitivity maps and the perceptual error maps are shown in Fig. 5. Here, DeepQA was trained with $w_{TV} = 10^{-2}$. The distorted images with four different artifact types (JPEG2000, JPEG, WN, and GB) are shown in (a), (e), (i), and (m). Figs. (b), (f), (j), and (n) are the objective error maps obtained from (1), Figs. (c), (g), (k), and (o) are the predicted sensitivity maps, and Figs. (d), (h), (l), and (p) are the perceptual error maps. The darker regions indicate more distorted pixels. In (a), the distortion around the houses was more noticeable than that on the rocks, as shown in (d). For JPEG distortion, the banding artifact on the sky regions was emphasized in (h). In case of additive white noise, the objective error was uniformly distributed over the image, as shown in (j). In the perceptual error map, the distortion on the homogeneous regions was more noticeable than that on the textural regions, as shown in (l), which agrees with the contrast masking and CSF. When the image was distorted by Gaussian blur, strong edges were especially distorted as (n), and the perceptual error map also had similar tendency, as shown in (p).

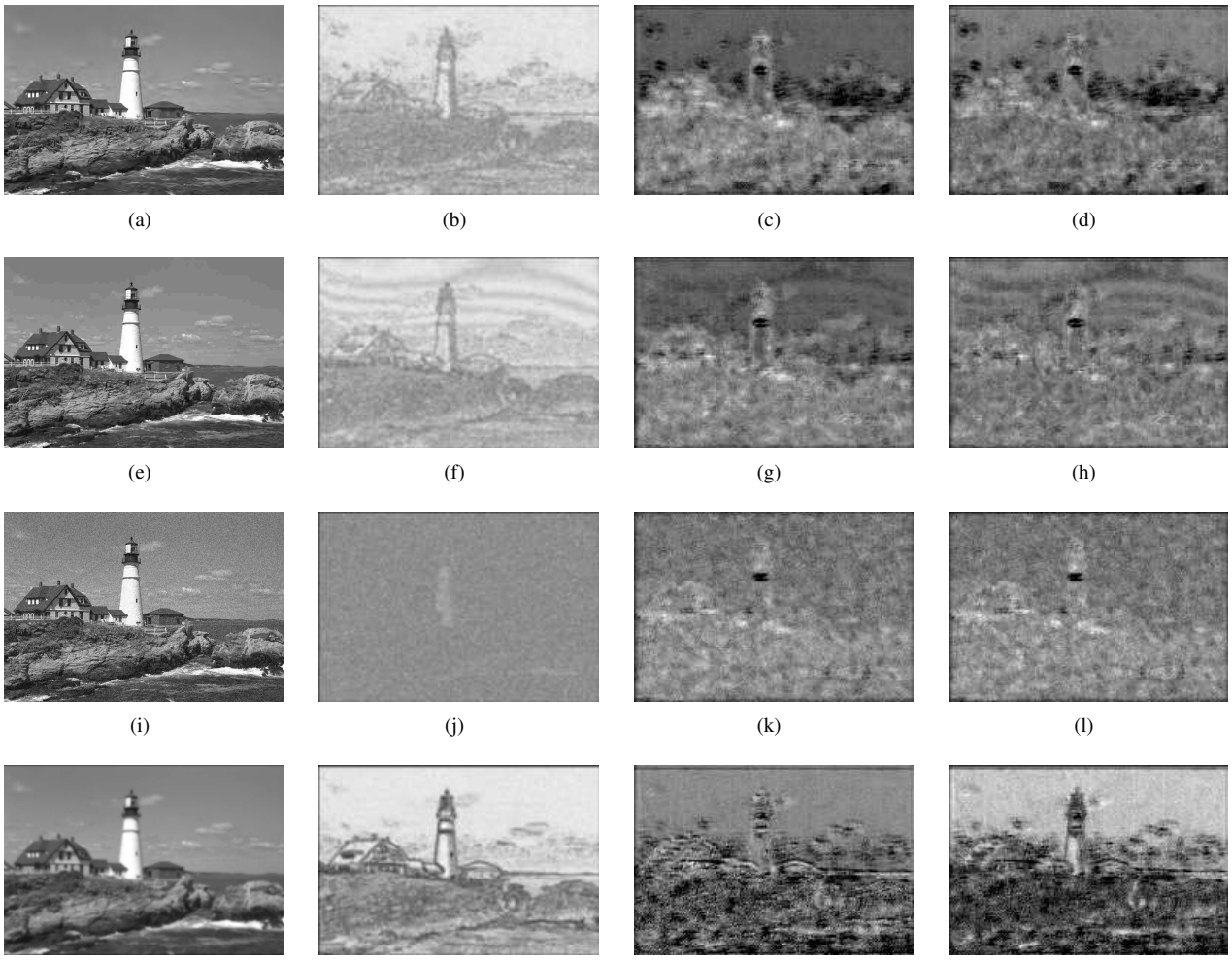


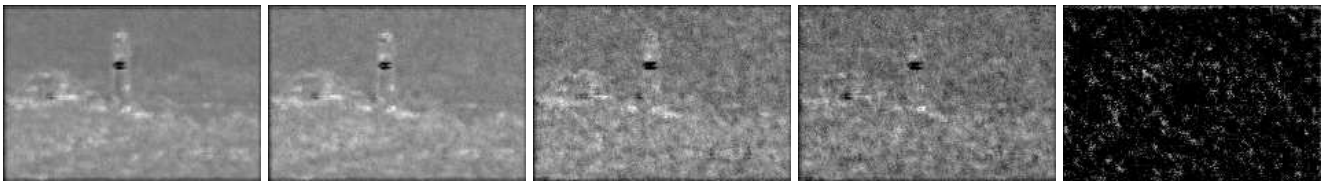
Figure 5. Examples of the predicted sensitivity maps; (a), (e), (i), and (m) are distorted images with JEPG2000, JPEG, white noise, and Gaussian blur; (b), (f), (j), and (n) are the objective error maps; (c), (g), (k), and (o) are the predicted sensitivity maps; (d), (h), (l), and (p) are the perceptual error maps.

In Fig. 6, the perceptual error maps of white noise and Gaussian blur with different distortion levels are shown. The fist row indicates the white noise, and the second row indicates the Gaussian blur. When there is strong white noise, the perceptual error map loses the structural details as shown in (e). In contrast, when the Gaussian blur increases, the distortion on the strong edges were more emphasized. Generally, as the degree of distortion increased, the averaged value of the perceptual error map decreased, which indicates that DeepQA makes rational quality prediction according to degree of distortion.

4.4. Performances Comparison

To evaluate the performance of DeepQA, we randomly divided the reference images into two subsets (80% for training and 20% for testing) and their corresponding distorted images were divided in the same way so that there was no overlap between the two sets. To increase the number of training samples, horizontally flipped images were additionally supplemented. DeepQA was trained in a non-distortion-specific way, that is, all the distortion types were used simultaneously. The training iterated 80 epochs, then the model with the lowest validation error was chosen over the epochs. The prediction accuracy mostly saturated after 50 epochs as shown in Fig. 4. The correlation coefficients of DeepQA were averaged after the procedure was repeated 20 times while dividing the training and testing sets randomly in order to eliminate the performance bias.

DeepQA was compared to eight FR-IQA metrics: PSNR, SSIM [32], MS-SSIM [33], VIF [28], GMSD [34], FSIMc [35], DOG-SSIMc [17], and FR-DCNN [14]. In addition, four deep learning-based NR-IQA methods were bench-



(a) 0.8037 (b) 0.7125 (c) 0.4260 (d) 0.3657 (e) 0.0125

(f) 0.7361 (g) 0.5696 (h) 0.5299 (i) 0.4816 (j) 0.3630

Figure 6. Examples of the perceptual error maps with various distortion levels of white noise, and Gaussian blur: (a) - (e) are distorted by white noise; (f) - (j) are distorted by Gaussian blur. The values indicate the averages of the perceptual error maps ($\mu_{\mathbf{p}}$).

Table 1. SRCC and PLCC comparison on the four IQA databases. FR (NR) indicates full-reference (no-reference) models, and italics indicate deep learning-based methods.

Type		LIVE IQA		CSIQ		TID2008		TID2013		Weighted Avg.	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
FR	PSNR	0.876	0.872	0.806	0.800	0.553	0.573	0.636	0.706	0.666	0.704
	SSIM	0.948	0.945	0.876	0.861	0.775	0.773	0.637	0.691	0.745	0.767
	MS-SSIM	0.951	0.949	0.913	0.899	0.854	0.657	0.786	0.833	0.842	0.809
	VIF	0.963	0.960	0.920	0.928	0.749	0.808	0.677	0.772	0.765	0.826
	GMSD	0.960	0.960	**0.957**	**0.954**	**0.891**	0.879	0.804	0.859	0.867	0.890
	FSIMc	0.960	0.961	0.931	0.919	0.884	0.876	**0.851**	**0.877**	**0.884**	**0.893**
	DOG-SSIMc	0.963	0.966	0.954	0.943	**0.935**	**0.937**	**0.926**	**0.934**	**0.937**	**0.940**
	FR-DCNN	**0.975**	**0.977**	-	-	-	-	-	-	-	-
	DeepQA–s	**0.977**	**0.975**	**0.957**	**0.956**	0.878	**0.892**	0.766	0.818	0.848	0.876
	DeepQA	**0.981**	**0.982**	**0.961**	**0.965**	**0.947**	**0.951**	**0.939**	**0.947**	**0.949**	**0.955**
NR	*SESANIA*	0.934	0.948	-	-	-	-	-	-	-	-
	CNN	0.956	0.953	-	-	-	-	-	-	-	-
	Patchwise	0.960	0.972	-	-	-	-	0.835	0.855	-	-
	BIECON	0.958	0.960	-	-	-	-	-	-	-	-

marked: SESANIA [7], CNN [8], a 'Patchwise' method in [2], and BIECON [9].

In Table 1, SRCC and PLCC of the FR-IQA algorithms on LIVE IQA and TID2008 are compared. In the last column, the weighted average of SRCC and PLCC over the four databases are reported, where each weight is proportional to the number of distorted images of each database. The top three models for each evaluation criterion are shown in boldface. The reported SRCC and PLCC scores of the deep learning-based models were taken from the original papers. When all the distortion types were considered together, the highest SRCC and PLCC were achieved by DeepQA, followed by DOG-SSIMc. DeepQA outperformed the other metrics on all the databases consistently, meanwhile DeepQA–s only achieved competitive performances on the LIVE IQA and CSIQ databases. From this observation, it is obvious that taking the error map as an input helps the CNN model extract more useful features to achieve a higher accuracy.

Table 2 shows the SRCC comparison according to individual distortion type on the LIVE IQA and TID2008 databases. Even when each distortion type was tested separately, DeepQA generally achieved the competitive accuracies on most distortion types. Since grayscale images were used in DeepQA, it achieved low performance in local block-wise distortions (Block), where color change is an essential cue of distortion. In addition, since the normalization process discards low-frequency components, DeepQA had low correlation on mean shift (MS), where the global brightness was changed consistently. Overall, DeepQA achieved the competitive and consistent accuracies across all the databases.

Table 2. SRCC comparison on individual distortion types on the LIVE IQA and TID2008 databases. Italics indicate deep learning-based methods.

	Dist.type	PSNR	SSIM	MS-SSIM	VIF	GMSD	FSIMc	*DeepQA–s*	*DeepQA*
LIVE IQA	JP2K	0.895	0.961	0.963	0.969	0.968	**0.972**	**0.973**	**0.970**
	JPEG	0.881	0.972	**0.981**	**0.984**	0.973	**0.979**	0.976	0.978
	WN	**0.985**	0.969	0.973	**0.985**	0.974	0.971	**0.991**	**0.988**
	BLUR	0.782	0.952	0.954	**0.972**	0.957	0.968	**0.979**	**0.971**
	FF	0.891	0.955	0.947	**0.965**	0.942	0.950	**0.956**	**0.968**
TID2008	AGN	0.907	0.811	**0.953**	0.880	**0.918**	0.875	0.913	**0.980**
	ANMC	**0.899**	0.803	**0.913**	0.876	**0.898**	0.893	0.745	0.863
	SCN	**0.917**	0.815	0.809	0.870	**0.913**	0.871	0.902	**0.970**
	MN	**0.852**	0.779	0.805	**0.868**	0.709	**0.825**	0.717	0.795
	HFN	**0.927**	0.873	0.821	0.907	0.919	0.913	**0.958**	**0.974**
	IMN	**0.873**	0.673	**0.811**	**0.833**	0.661	0.771	0.710	0.725
	QN	0.870	0.853	0.869	0.797	**0.887**	**0.873**	0.842	**0.901**
	GB	0.870	**0.954**	0.691	**0.954**	0.897	0.947	0.939	**0.950**
	DEN	0.942	**0.953**	0.859	0.916	**0.975**	**0.961**	0.859	0.929
	JPEG	0.872	0.925	**0.956**	0.917	**0.952**	0.929	0.848	**0.940**
	JP2K	0.813	0.962	0.958	**0.971**	**0.980**	**0.978**	0.922	0.958
	JGTE	0.752	0.868	**0.932**	0.859	0.862	**0.876**	0.730	**0.880**
	J2TE	0.831	0.858	**0.970**	0.850	**0.883**	0.856	0.871	**0.928**
	NEPN	0.581	0.711	**0.868**	**0.762**	**0.760**	0.751	0.244	**0.760**
	Block	0.619	0.846	**0.861**	0.832	**0.897**	**0.847**	0.095	0.517
	MS	0.696	**0.723**	**0.738**	0.510	0.649	0.655	**0.703**	0.652
	CTC	0.586	0.525	**0.755**	**0.819**	0.466	0.651	0.602	**0.838**

Table 3. Cross dataset test on the subset of the TID2008 database (SRCC).

Metrics	JP2K	JPEG	WN	BLUR	ALL
SSIM	0.963	0.935	0.817	0.960	0.902
DeepQA–s	0.932	0.912	0.896	0.947	0.916
DeepQA	0.945	0.928	0.890	0.957	0.940

4.5. Cross Dataset Test

To evaluate the generalization of DeepQA models, they were trained using the LIVE IQA database, and tested on the TID2008 database. Since the TID2008 database contains broader kinds of distortion types, we chose four distortion types (JPEG, JPEG2000 compression, WN, and BLUR) which are common in the two databases. The results of the cross dataset test are shown in Table 3. It can be concluded that both DeepQA–s and DeepQA perform well and that the performances of them do not depend on the database.

5. Conclusion

In this paper, we have described a new FR-IQA framework where the CNN model learns the human visual sensitivity. By using a triplet of a distorted image, its objective error map, and its subjective score, the proposed model could learn the behavior of the HVS. Moreover, a TV regularization was proposed to penalize the high frequency components in the predicted sensitivity map. With the proper TV regularization, the sensitivity map becomes more visually plausible without loss of performance. Through the rigorous experiments, we checked that the predicted perceptual error maps agree with the HVS. The sensitivity maps were predicted well with various distortion types and degrees of distortion. In addition, proposed DeepQA achieved the state-of-the-art correlation score on every database. In future study, we plan to advance the proposed framework to NR-IQA to predict the subjective score without the reference images, which is the most challenging problem in IQA.

References

[1] Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II. *VQEG*, 2003.

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

[3] C.-H. Chou and Y.-C. Li. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. *IEEE Trans. Circuits Syst. Video Technol.*, 5(6):467–476, 1995.

[4] S. J. Daly. The visible differences predictor: An algorithm for the assessment of image fidelity. In *Proc. SPIE, Human Vision, Visual Processing, and Digital Display III*, volume 1666, pages 2–15, 1992.

[5] T. Dozat. Incorporating Nesterov momentum into Adam. Technical report, Stanford University, Tech. Rep., 2015., 2015.

[6] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2366–2374, 2014.

[7] W. Hou, X. Gao, D. Tao, and X. Li. Blind Image Quality Assessment via Deep Learning. *IEEE Trans. Neural Netw. Learn. Syst.*, 26(6):1275–1286, 2015.

[8] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1740, 2014.

[9] J. Kim and S. Lee. Fully deep blind image quality predictor. *IEEE J. Sel. Top. Signal Process.*, 11(1):206–220, 2017.

[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[12] E. C. Larson and D. M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imaging*, 19(1):011006–011006–21, 2010.

[13] G. E. Legge and J. M. Foley. Contrast masking in human vision. *J. Opt. Soc. Am.*, 70(12):1458–1471, 1980.

[14] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng. Image quality assessment using similar scene as reference. In *European Conference on Computer Vision (ECCV)*, pages 3–18. Springer, Cham, 2016.

[15] T. J. Liu, W. Lin, and C. C. J. Kuo. Image quality assessment using multi-method fusion. *IEEE Trans. Image Process.*, 22(5):1793–1807, 2013.

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *arXiv:1411.4038 [cs]*, 2014.

[17] Y. Lv, G. Jiang, M. Yu, H. Xu, F. Shao, and S. Liu. Difference of Gaussian statistical features based blind image quality assessment: A deep learning approach. In *IEEE International Conference on Image Processing (ICIP)*, pages 2344–2348, 2015.

[18] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*, 2013.

[19] A. Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vis.*, 2016.

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

[21] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

[22] M. Narwaria and W. Lin. SVD-based quality metric for image and video using machine learning. *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, 42(2):347–364, 2012.

[23] S. C. Pei and L. H. Chen. Image quality assessment using human visual DOG model fused with random forest. *IEEE Trans. Image Process.*, 24(11):3282–3292, 2015.

[24] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.

[25] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - A database for evaluation of full-reference visual quality assessment metrics. *Adv. Mod. Radioelectron.*, 10(4):30–45, 2009.

[26] 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.*, 21(8):3339–3352, 2012.

[27] H. Sheikh, M. Sabir, and A. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3440–3451, 2006.

[28] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Trans. Image Process.*, 15(2):430–444, 2006.

[29] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *International Conference on Image Processing*, volume 3, pages 444–447, 1995.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556 [cs]*, 2014.

[31] H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 305–312, 2011.

[32] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[33] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *IEEE Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1398–1402, 2003.

[34] 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.*, 23(2):684–695, 2014.

[35] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20(8):2378–2386, 2011.