Dynamic Receptive Field Generation for Full-Reference Image Quality Assessment

Woojae Kim, Anh-Duc Nguyen, Sanghoon Lee, *Senior Member, IEEE*, and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—Most full-reference image quality assessment (FR-IQA) methods advanced to date have been holistically designed without regard to the type of distortion impairing the image. However, the perception of distortion depends nonlinearly on the distortion type. Here we propose a novel FR-IQA framework that dynamically generates receptive fields responsive to distortion type. Our proposed method-*dynamic receptive field generation based image quality assessor* (DRF-IQA)-separates the process of FR-IQA into two streams: 1) dynamic error representation and 2) visual sensitivity-based quality pooling. The first stream generates dynamic receptive fields on the input distorted image, implemented by a trained convolutional neural network (CNN), then the generated receptive field profiles are convolved with the distorted and reference images, and differenced to produce spatial error maps. In the second stream, a visual sensitivity map is generated. The visual sensitivity map is used to weight the spatial error map. The experimental results show that the proposed model achieves state-of-the-art prediction accuracy on various open IQA databases.**

***Index Terms*—Full-reference image quality assessment (FR-IQA), dynamic receptive fields (DRFs), convolutional neural networks (CNNs), dynamic filter networks (DFNs), human visual system (HVS).**

I. Introduction

IN RECENT years, the rapid industry rollout of globally pervasive social media platforms and compressed image transmission systems have had to contend with a plethora of image quality degradations arising during the processes of content acquisition, transmission and storage [1], [2]. Efforts to improve these systems would greatly benefit from the development of models that can predict perceived image quality as accurately as possible. Accordingly, over the past few decades, numerous full-reference image quality assessment (FR-IQA) methods have been developed based on perceptual models that seek to mimic the human visual system (HVS) responses to distortion. However, designing an adequately detailed and holistic HVS model is an extremely difficult problem that is far from being solved.

A. Limitations of Conventional IQA Methods

Most existing FR-IQA methods exploit specific, well-understood visual characteristics, by mathematically formulating them into image processing models and algorithms. The simplest example is the peak signal-to-noise ratio (PSNR) and mean square error (MSE), which are based on the observation that the HVS is sensitive to differences or error signals between a reference image and a distorted version of it. More sophisticated approaches model basic perceptual processes such as the structural similarity (SSIM) model [3]. Other perception-driven IQA models include [4]–[8], the Visual Information Fidelity (VIF) model [9], which uses a perceptually relevant natural scene statistics approach [10]–[13], and FSIM, which embodies phase coherency in a SSIM-like computation [14].

Existing FR-IQA models have generally been developed from a holistic point of view, meaning that they operate in the same manner regardless of the nature of the distortion. Since different types of distortion are perceived as different, it is reasonable to believe that an image quality model could exploit these differences in perception.

B. Dynamic Receptive Field Generation

Motivated by these observations, we developed an FR-IQA framework that we term *dynamic receptive field image quality assessor* (DRF-IQA). DRF-IQA exploits the advantages of convolutional neural network (CNN) based deep learning methods to model the highly non-linear responses of the visual system to picture distortions of diverse types, spatial characteristics, and severities. Our proposed deep learning method is inspired by the following observation.

Humans with normal vision are quite good at perceiving visual quality. Indeed, the sense of distortion is largely driven by front-end, low-level processing. It may be viewed as a pre-attentive response to degradations of the statistical structure of pictures induced by distortion [15]. This broadly drives current high-performing holistic FR-IQA models. Humans perceive distortion rapidly, and they also understand differences in distortion appearance rapidly.

In this direction, we have developed a way to learn a set of dynamic receptive fields responsive to different

Manuscript received May 21, 2019; revised November 18, 2019; accepted January 11, 2020. Date of publication January 27, 2020; date of current version February 7, 2020. This work was supported by the Samsung Research Funding Center of Samsung Electronics under Project SRFC-IT1702-08. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marta Mrak. *(Corresponding author: Sanghoon Lee.)*

Woojae Kim, Anh-Duc Nguyen, and Sanghoon Lee are with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, South Korea (e-mail: wooyoa@yonsei.ac.kr; adnguyen@yonsei.ac.kr; slee@yonsei.ac.kr).

Alan Conrad Bovik is with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA (e-mail: bovik@ece.utexas.edu).

Digital Object Identifier 10.1109/TIP.2020.2968283

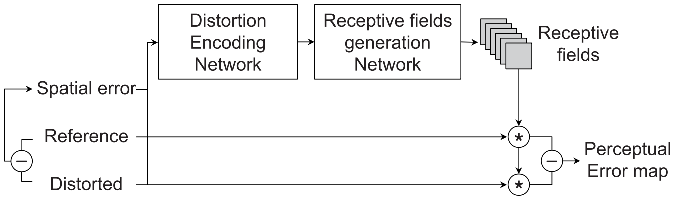


Fig. 1. Concept of dynamic receptive field generation for FR-IQA.

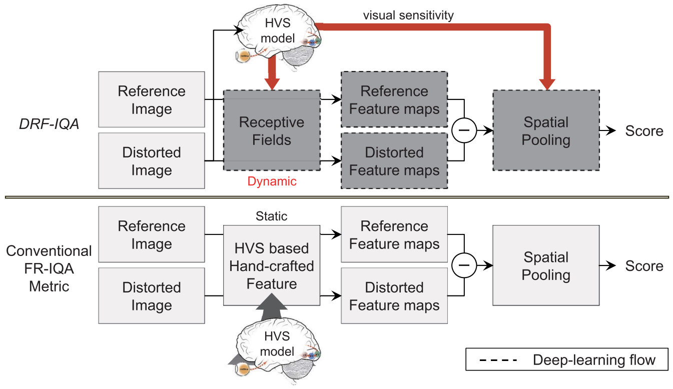


Fig. 2. Flow chart comparison of DRF-IQA and conventional FR-IQA. Gray boxes indicate the deep-learning flow.

distortion types. Fig. 1 shows the DRF-IQA framework. The dynamic receptive field generation module is motivated by the generation network in [16]. It consists of a *distortion encoding network*, which encodes the type of the distortion, and a *receptive field generation network*, which generates dynamic receptive fields having learned profiles using the distortion code. The generated DRFs process the reference and distorted images, then a perceptual error map is computed using a simple distance metric.

Fig. 2 shows a comparison between conventional FR-IQA models and DRF-IQA. Conventional methods compute local error maps using hand-crafted features by modeling physiological aspects of human visual perception. For example, the well-known Gabor-wavelet decompositions are widely used to model neuronal receptive fields in the primary visual cortex [9], [17], [18]. Nonetheless, these approaches ultimately yield limited quality prediction performance, because it is exceedingly difficult to design a holistic model that is capable of capturing the dynamics of each distortion. By contrast, DRF-IQA is designed to dynamically generate receptive fields responsive to the type of distortion.

To model the dynamic receptive fields, DRF-IQA employs a fully convolutional end-to-end network that learns the mappings between visual perception and the distortions embedded in IQA databases. We also introduce a new visual weighting process to spatially pool the quality abstractions produced by the network. The pooled scores predict the perceived degree of each distortion by type.

Our contributions are summarized as follows:

1) We create a first-of-kind DRF generation based FR-IQA framework implemented as an end-to-end CNN model.
2) The DRFs and visual pooling weights are learned without injecting any prior knowledge of the HVS.
3) Our learned IQA model achieves state-of-the-art image quality prediction performance.

The remainder of the paper is organized as follows. Section II introduces related work recent deep learning-based IQA models. Section III describes the architecture of the DRF-IQA framework, including the implementations of DRF generation, spatial pooling, and the training procedure. Section IV discloses and explains the experimental results of testing DRF-IQA under various ablation protocols. In addition, visualization and analysis of the trained deep model are presented. Lastly, concluding remarks are given in Section V.

II. Related Works

A. Image Quality Assessment

Image quality assessment models are usually classified into one of three categories: FR-IQA, which utilizes the distorted image and a reference to produce quality predictions, reduced-reference IQA (RR-IQA) which only uses incomplete reference information, and no-reference IQA (NR-IQA) where quality is predicted without using any reference information.

FR-IQA has been widely applied to gauge the performance of image/video transmission systems and lossy video compression [19], [20]. Popular perceptual FR-IQA methods include SSIM [3], feature similarity (FSIM) [14] and visual information fidelity (VIF) [9] which are based on the measurement of perceptual distances between possibly distorted images and their references [3], [9], [14]. While these are widely used, these and similar FR-IQA incorporate handcrafted models of distortion perception, which are inevitably incomplete given the vast range of possible distortion types, severities, and mixtures. This, of course, limits their performances in applications.

B. Dynamic Filter Network

Ordinarily, the filters learned by a traditional convolutional layer remain fixed after training. Our dynamic filter network generates filters that adapt to the input and that change as new samples are introduced [16]. The concept of dynamic filter generation has been used in a variety of recent computer vision applications [21], [22].

C. Deep Learning-Based IQA Models

A variety of ways to apply deep learning to the IQA problem have been recently proposed [23]. Hou et al. proposed a deep belief network (DBN) model using NSS-related features expressed in the wavelet domain [24]. Similarly, Li *et al.* derived deep-learning based DoG features which they regressed on quality scores [25]. Ghadiyaram and Bovik developed a large number of NSS features which they used to train a DBN to predict image quality [26]. However, these approaches utilized handcrafted features and small datasets, hence could not fully exploit the advantages of deep learning methods.

Another recent end-to-end CNN model was trained on many image patches, each labeled by the global mean opinion

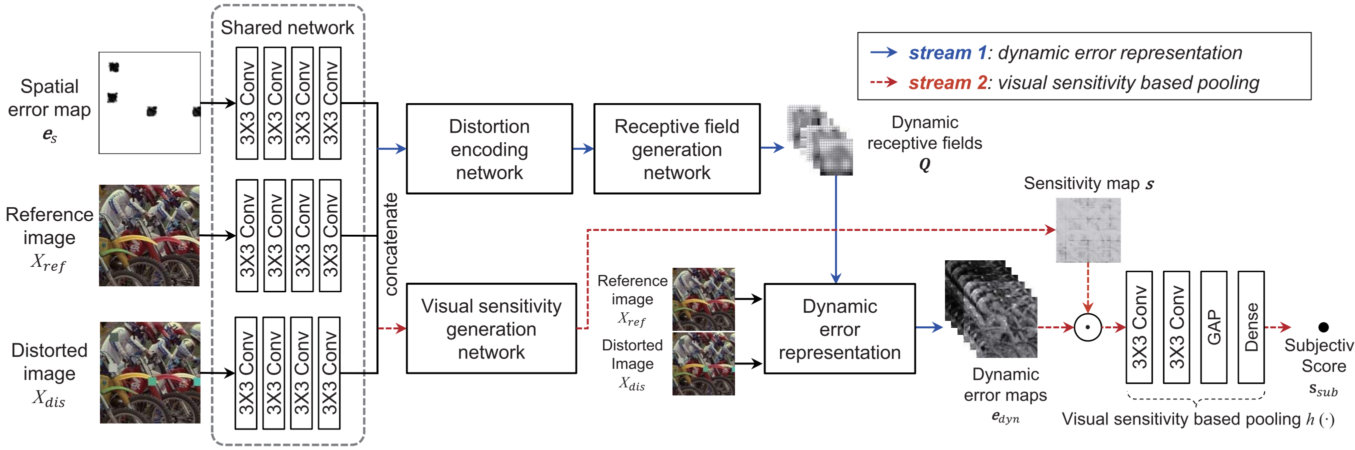

Shared network
Spatial error map e_s
Reference image X_{ref}
Distorted image X_{dis}
3X3 Conv
concatenate
stream 1: dynamic error representation
stream 2: visual sensitivity based pooling
Distortion encoding network
Receptive field generation network
Dynamic receptive fields Q
Sensitivity map s
Visual sensitivity generation network
Reference image X_{ref}
Distorted Image X_{dis}
Dynamic error representation
Dynamic error maps e_{dyn}
GAP
Dense
Subjective Score s_{sub}
Visual sensitivity based pooling $h(\cdot)$

Fig. 3. Flow diagram of the DRF-IQA framework.

score (MOS) of the entire image it is part of. Bosse *et al.* developed a model that spatially pools image quality over a set of image partitions using CNN model [27], again training on global image scores. Oh *et al.* applied patch based CNN model into stereoscopic 3D IQA by local to global feature aggregation [28]. Seo *et al.* incorporated measures of visual saliency and just noticeable difference (JND) into a CNN model to extract perceptually important features [29]. Other authors have used models of visual sensitivity to perceptually configure the network [23], [30], [31]. Another recent CNN-based FR video quality assessment (VQA) employs temporal HVS properties as intermediate training targets of the CNN [32]. An attentional temporal pooling model was also used in this end-to-end optimization. Our new DRF-IQA model, which generates DRFs and deploys a new visual sensitivity model, achieves state-of-the-art prediction accuracy, which we support with strong visualizations.

III. DRF-IQA Framework

The overall flow of the DRF-IQA framework is depicted in Fig. 3. The model is implemented as an end-to-end CNN, which processes two data streams: Stream 1 is a dynamic error representation, while Stream 2 is a visual sensitivity based pooling network. Each distorted image, corresponding reference image, and a spatial error map are fed into the shared CNN network. In Stream 1, the *distortion encoding network* encodes each input in accordance with characteristics of distortion. Next, the encoded distortion-specific information is used to generate DRFs using the *receptive field generation network*. Next, in the *dynamic error representation*, the input image pair is convolved with the generated DRFs, and channel-wise subtraction is conducted to create the dynamic error maps. Stream 2 takes the output of the shared network as an input, and generates a single-channel sensitivity map. The pixels in the dynamic error maps are then weighted by the values of the sensitivity map. Global average pooling (GAP) is then used to obtain the overall quality score [33]. Finally, the predicted score is regressed onto the subjective score in a supervised manner. The details of the dual-stream architecture are explained in Section III-A.

A. Model Architecture

1) Motivations: The design of the DRF generator is inspired by [16], while the visual sensitivity based pooling model is derived from [30], [32]. The most intuitive way to generate the DRFs is to process the input maps (the distorted image, reference image, and spatial error map), then infer distortion information, and then output the DRFs using the inferred information. In this way, the model is able to generate DRFs that reflect distortion information. In Stream 2, a sensitivity map is obtained along with the dynamic error maps. A U-Net architecture is used to generate the sensitivity map.

2) Model Design: DRF-IQA operates on multiple patches, drawn viz., both the distorted and reference images are partitioned into patches of the same sizes, which are fed into the system for inferencing in Section III-E. Each layer in Streams 1 and 2 uses a 3×3 convolutional filters and a ReLU activation function [34]. However, Streams 1 and 2 terminate with linear and sigmoid activations, respectively. In this way, each generated DRF is made to be zero-centered by the linear activation, while the sigmoid activation distributes the values of the sensitivity map between 0 to 1. As shown in on the left side of Fig. 3, the processed inputs of Streams 1 and 2 both enter the shared network. In the shared network, only convolutional layers are utilized to avoid losing spatial information.

B. Shared Network

The shared network serves a preprocessing stage. Three inputs are processed to produce feature maps: the distorted image, the reference image, and the spatial error map. Given an input distorted image X_{dis} and reference image X_{ref}, a spatial error map is obtained as an objective error signal, in the following manner.

Rather than using a simple squared-error distance metric between the reference and distorted samples, which can result in many zero values, and adversely affect both training convergence and application, we instead define a spatial error map $\boldsymbol{e}_S$ as a normalized and shifted log difference yielding a non-zero

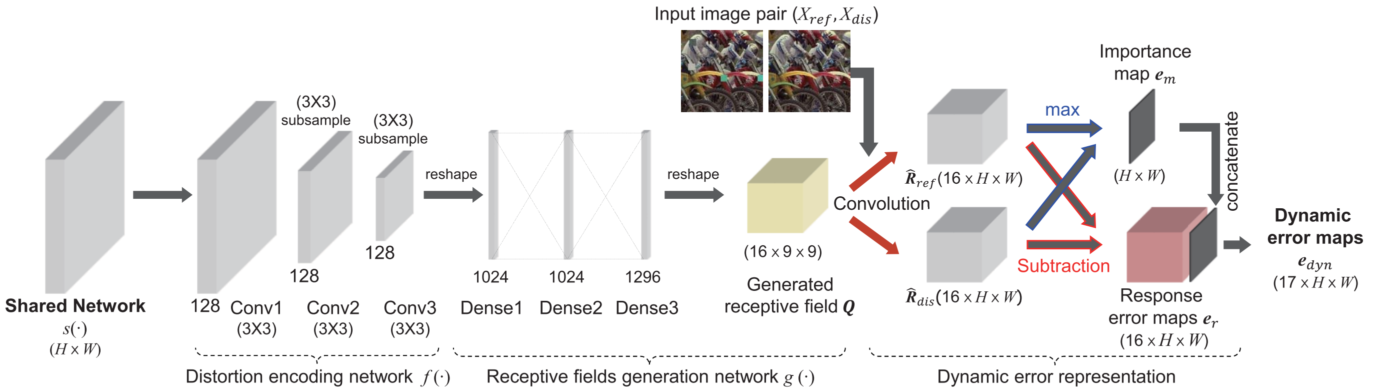


Fig. 4. Architecture of stream 1: the distortion encoding network and dynamic receptive field generating network. "Conv" represents the convolutional layers, while "dense" repres5ents fully connected layers. The text below "Conv" indicates the size of the filters. The red arrow represents the convolution operation.

centered distribution as in [30], [32]:

$$\boldsymbol{e}_S = \frac{\log(1/((X_{ref} - X_{dis})^2 + \epsilon/255^2)}{\log(255^2/\epsilon)}, \tag{1}$$

where we use $\epsilon = 1$ in the experiments. Each map is then individually fed into four convolutional layers (the number of filters is 32 in each layer), then their outputs are concatenated and used as the input the two streams.

C. Stream 1: Dynamic Error Representation

1) Distortion Encoding Network: The general architecture of stream 1 is shown schematically in Fig. 4. To generate DRFs responsive to input distortions, it is necessary to analyze the input distortion. Since there are no labels available regarding the type or degree of distortion, it is instead directly encoded by a convolutional encoding network.

To encode the input maps, three convolutional networks with 3×3 subsampling are used. Let us denote the procedure of network encoding by $f(\cdot)$. As a way of showing that the encoded feature maps $f(\cdot)$ are able to capture distortions by type, Fig. 5 visualizes the distribution of encoded feature maps using t-SNE [35] applied on the TID2008 image quality dataset, which contains 17 distinct distortion types, as depicted in Fig. 5 (a), while five of the more common distortion types are visualized in Fig. 5 (b): AGN, spatially correlated noise (SCN), Gaussian blur (GB), JPEG compression and mean shift (MS). Each distortion type is clearly encoded.

2) Receptive Field Generating Network: As shown in Fig. 4, the generating network $g(\cdot)$ is implemented using a few fully-connected layers, where the last layer contains 1296 neurons, corresponding to N receptive fields of resolution 9×9.

We denote the generated receptive fields as Q^n, where N is the number of receptive fields ($n = 1, \ldots, N$). We fixed $N = 16$ to achieve a broadly representative set of distortion-sensitive receptive fields. Fig. 7 shows a 3D visualization of the sixteen generated DRFs trained on the LIVE IQA database, along with cross-sections of them. As shown in the figure, the generated DRFs include various shapes, scales and orientations. In addition, it may be seen that the cross-sections of the DRFs exhibit similar similar appearance as classical receptive field models (*e.g.,* Gabor wavelets). Detailed examples are given in Section IV-F.

3) Dynamic Error Representation: Once the model is trained to generate DRFs, a 2D convolution operation is applied to each input image pair (reference and distorted images). As shown in Fig. 4, there are $N = 16$ responses corresponding to both the reference and distorted images. Each response is then subjected to separate *subtraction* and *max* operations. The perceptual error map is produced by the channel-wise subtraction between the two response maps, while the max operator is applied on both responses $\hat{R}_{ref}$ and $\hat{R}_{dis}$ to produce an importance map. Since not all of the spatial regions within an image are deemed to be distorted, those that are not included in the maps.

The dynamic error representation procedure is expressed precisely as follows. We use convolution with stride one and zero padding at the image borders. Then each response is denoted as $\hat{R}_{ref}$ and $\hat{R}_{dis}$ in $\mathbb{R}^{N \times H \times W}$, where H and W are the height and width of input image. The n^{th} response map is defined as

$$\hat{R}_k^n = Q^n * X_k \mid k \in \{ref\,,\, dis\} \tag{2}$$

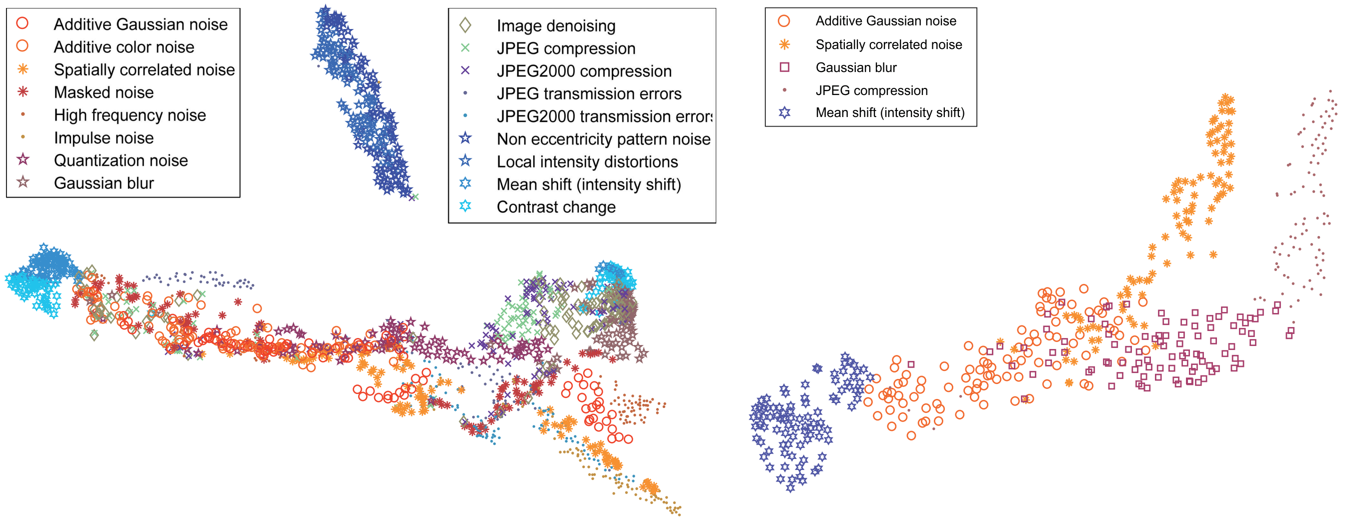
where $*$ is convolution, and Q^n is the n^{th} receptive field generated as explained in Section III-C. Then, the dynamic error maps are defined by concatenating the response error maps and the importance map as follows

$$\boldsymbol{e}_{dyn} = \mathrm{CONCAT}[\boldsymbol{e}_r,\, \boldsymbol{e}_m], \tag{3}$$

where $\boldsymbol{e}_r$ is the set of response error maps defined by $\boldsymbol{e}_r = |(\hat{R}_{ref} - \hat{R}_{dis})|$ and the importance map is given by $\boldsymbol{e}_m = \max(\hat{R}_{ref}, \hat{R}_{dis})$.

D. Stream 2: Visual Sensitivity Pooling

1) Visual Sensitivity Generation Network: Following prior work on the use of CNN-based sensitivity maps [23], [30]–[32], we designed an intuitive way to spatially weight the values of the spatial error map, using a generative approach (by generating a target from the encoded data) using the networks in [37], [38]. Fig. 6 shows the visual sensitivity generating network, denoted as $\boldsymbol{v}(\cdot)$. The network uses a U-Net structure [37] to preserve the dimension of the sensitivity map.



(a) (b)

Fig. 5. t-SNE scatter plot of the output of the distortion encoding network on the TID2008 database [36]. The database includes 17 distortion types. (a) t-SNE for over all distortion types. (b) t-SNE over five common distortion types. Notice that each encoded feature sample is coherently clustered according by distortion type.

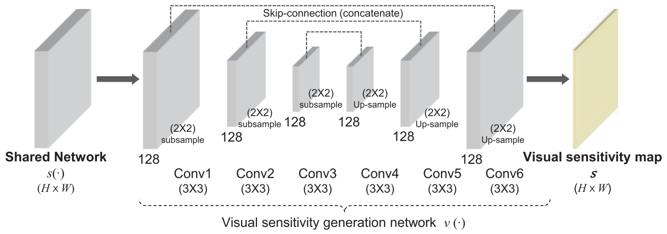


Fig. 6. Architecture of stream 2: the visual sensitivity generating network.

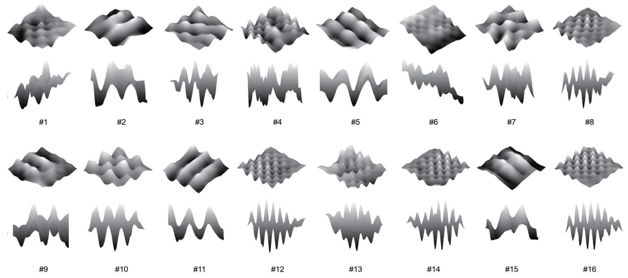


Fig. 7. 3D Visualization of the sixteen generated dynamic receptive field profiles and their horizontal cross sections. The DRFs are resized to 64×64 and the model was trained on the LIVE IQA database [39]. Note that each map has a different scale that has been normalized to [0, 1].

Moreover, the generated sensitivity map is weighted by the dynamic error maps which contain unique perceptual error signals produced by the DRFs. The network accepts the preprocessed input maps output by the shared network. Three convolutional layers with 2×2 subsample layers then encode the visual sensitivity information, and three dilated convolutional layers upsample the encoded feature maps. The skip connections between the encoding and upsampling paths are implemented via a concatenation operator.

2) Visual Sensitivity Based Spatial Pooling: After passing through the visual sensitivity generating network, the dynamic error representation e_{dyn} is weighted by the sensitivity map, as shown in Fig. 3. Then, the weighted maps are fed to the two convolutional layers, and the final feature maps are regressed onto predicted subjective scores, followed by the GAP operation. Since DRF-IQA is patch based, the overall predicted score is averaged over all the patches in each image sample. Therefore, the final predicted quality score of the distorted image is

$$\mathbf{s}_{pred}(\mathbf{X}_d, \mathbf{X}_r; \theta_{f,g}, \theta_v) = \frac{1}{M} \sum_{i \in M} h(\boldsymbol{e}_w; \theta_{f,g}, \theta_v), \tag{4}$$

where M is the number of patches in the image sample, $\mathbf{X}_d, \mathbf{X}_r$ are the distorted and reference images, and $\theta_{f,g}, \theta_v$ are intrinsic CNN parameters, respectively. Let e_w be the sensitivity weighted dynamic error representation maps defined by

$$\boldsymbol{e}_w = \boldsymbol{e}_{dyn} \odot \boldsymbol{s} \tag{5}$$

where $\odot$ is the element-wise product. The loss function is the mean-squared error between the predicted and ground-truth subjective scores

$$\mathcal{L}_{mse} = ||\mathbf{s}_{pred} - \mathbf{s}_{sub}||_F^2,$$

where $\mathbf{s}_{sub}$ is the subjective score label (MOS) of the image sample. In addition, a total variation (TV) term was used to alleviate high-frequency noise in the sensitivity map. TV regularization is defined as

$$TV(s) = \frac{1}{H_s \cdot W_s} \sum_{(i,j)} (s_{horz}(i,j)^2 + s_{vert}(i,j)^2) \tag{6}$$

where H_s and W_s are the height and width of the sensitivity map, and s_{horz} and s_{vert} are Sobel-filtered (directional derivative) sensitivity maps taken in the horizontal and vertical directions, respectively. Moreover, L_2 regularization is applied to all the layers to avoid overfitting, as in [30]–[32]. Finally, the loss function that we utilized to perform the overall training process is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{mse} + \lambda_1 \cdot \mathcal{L}_{TV} + \lambda_2 \cdot \mathcal{L}_{l2} \tag{7}$$

where λ_1 and λ_2 are the relative weights of $\mathcal{L}_{TV}$ and $\mathcal{L}_{l2}$, respectively.

E. Training Strategy

Once the input $(\mathbf{X}_d, \mathbf{X}_r, \mathbf{e})$ is prepared, it is partitioned into patches of equal size. In our experiments, the patch size $size_{patch}$ was set to 112×112. After each patch is processed by their individual CNN models, they are averaged to obtain a predicted score, on each image sample. For example, if there are 12 patches in an image sample, then *M=12* in (4), and the final predicted score is obtained by averaging the 12 outputs of the CNN model. Note that the loss function is not obtained for each patch sample, but rather on all patches in the image unit.

To achieve better convergence, the adaptive moment estimation optimizer (ADAM) was used [40] rather than the usual form of stochastic gradient descent. We used the default hyperparameters suggested for ADAM in the literature [40], and the momentum parameter was set to 0.9. The learning rate was initially set to 0.001, then multiplied by 0.1 after every 10 epochs, to achieve stable optimization. Finally, the weight parameters were fixed at $\lambda_1 = 1 \times 10^{-5}$ and $\lambda_2 = 5 \times 10^{-4}$, respectively.

IV. Experimental Results

A. Dataset

We benchmarked our proposed deep IQA engine on five well-known public datasets: LIVE IQA [39], CSIQ [41], TID2008, TID2013 [36], and LIVE Multiply Distorted (LIVE MD) [42]. Since DRF-IQA is a full-reference model, the benchmarked datasets all include both reference and distorted images. The LIVE IQA database contains 29 reference images and 799 distorted images impaired by five distortion types: JPEG and JP2K compression, white noise (WN), Gaussian blur (GB), and Rayleigh fast-fading (FF) channel distortion. The CSIQ image database includes 30 reference images and 866 distorted images with six applied distortion types: JPEG, JP2K, WN, GB, pink Gaussian noise (PGN), and contrast distortion (CTD). TID2008 contains 25 reference images and 1,700 distorted images with 17 different distortions at four levels of degradation, whereas TID2013 expanded this to include 3000 distorted images with 24 distortion types at five levels of degradation. The LIVE MD database includes 15 reference images and 405 distorted images, each degraded by two types of distortion. Some are corrupted by GB followed by JPEG (GB+JPEG), while others are corrupted by GB followed by WN (GB+WN). All MOS or DMOS values in these databases were scaled to [0, 1], where 1 equates to the best quality.

B. Evaluation Metrics

To validate the performance of DRF-IQA, we employed two standard measures: Spearman's rank-order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) [43]. A value close to 1 for SROCC and PLCC refers to higher performance in correlation. We compared our model against several state-of-the-art IQA methods. We first randomly divided the reference images into two content-separated subsets (80% for training and 20% for testing) along with their corresponding distorted images. Since DRF-IQA was trained in a non-distortion-specific way, all of the distortion types in each database were considered simultaneously. The correlation coefficients of the testing model were averaged after the experiment was repeated 20 times, each time randomly dividing the training and testing sets to eliminate the performance bias (cross-validation). When measured on all the databases, the standard deviations of SROCC and PLCC after 20 repetitions were less than 0.01 except on the LIVE MD dataset, where a relatively higher standard deviation was obtained likely because of the increased difficulty of dealing with multiple distortions. To augment the number of training data, horizontally flipped version of the images, along with the same labels, were included. During training, an early stopping scheme was used to avoid overfitting. In the experiments, implementation was carried out using the Theano library on the Python 3.5 platform. The GPU and CPU were the RTX2080 and Intel Xeon Gold 6140, respectively. Using this setup, the training time was 6.58 min/epoch on the TID2008 database. We compared DRF-IQA against ten FR-IQA models: PSNR, SSIM [3], MS-SSIM [44], VIF [9], GMSD [45], FSIMc [14], DoG-SSIMc [25], IFC [46], DeepQA [30], WaDIQaM-FR [27] and also six NR-IQA BLIINDS II [11], BRISQUE [47], BIECON [48], DIQA [31], NIMA [49], and DIQaM-NR [27].

1) Performance on Individual Databases: Table I lists the performances of the compared FR/NR-IQA models on each of the IQA databases. The bold fonts indicate the three top-performing models on each database. Furthermore, the weighted averages of the SROCC and PLCC scores over the five databases (LIVE, CSIQ, TID2008, TID2013, and LIVE MD) are also recorded in the last column. The weight given to the scores from each database was proportional to the number of images contained in each, to maintain per-image parity. Among the FR/NR-IQA models, the deep learning-based methods generally delivered superior performance relative to previous automatic and "hand-crafted" methods. DRF-IQA attained the highest correlation with respect to human subjectivity on most of the databases, and also in terms of the across-database weighted average. It performed exceptionally well on both TID2008 and TID2013, which are difficult because of the large varieties of distortion types they contained. In addition, DRF-IQA generally performed very well in terms of SROCC. In the experiment, the convergence was generally very fast on all the datasets, and the best result was obtained after ∼30 epochs.

2) Performance on Individual Distortion Types: Table II reports the SROCC and PLCC of the compared FR/NR-IQA

TABLE I

SROCC AND PLCC COMPARISON OF IQA MODELS ON FIVE IQA DATABASES. *Italics* INDICATE DEEP LEARNING-BASED METHODS

Type	Method	LIVE		CSIQ		TID2008		TID2013		LIVE MD		Weighted	
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
NR	BLIINDS-II	0.912	0.916	0.780	0.832	0.548	0.652	0.536	0.628	0.887	0.902	0.636	0.711
	BRISQUE	0.939	0.942	0.775	0.817	0.602	0.654	0.572	0.652	0.897	0.921	0.668	0.737
	BIECON	0.958	0.962	0.825	0.838	0.748	0.773	0.721	0.765	0.912	0.928	0.780	0.822
	DIQA	**0.975**	0.976	0.884	0.915	0.847	0.876	0.825	0.850	**0.939**	**0.942**	0.863	0.885
	NIMA	0.637	0.698	-	-	-	-	0.750	0.827	-	-	-	-
	DIQaM-NR	0.960	0.972	-	-	-	-	0.835	0.855	-	-	-	-
FR	PSNR	0.876	0.872	0.806	0.800	0.553	0.573	0.636	0.706	0.666	0.704	0.667	0.704
	SSIM	0.948	0.945	0.876	0.861	0.775	0.773	0.637	0.691	0.745	0.767	0.745	0.768
	MS-SSIM	0.951	0.949	0.913	0.899	0.854	0.657	0.786	0.833	0.842	0.809	0.842	0.810
	VIF	0.963	0.960	0.920	0.928	0.749	0.808	0.677	0.772	0.765	0.826	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	0.868	0.890
	FSIMc	0.962	0.962	0.932	0.920	0.884	0.876	0.851	0.877	0.863	0.818	0.885	0.893
	DOG-SSIMc	0.963	0.966	0.954	0.943	**0.935**	**0.937**	0.926	0.934	0.937	**0.940**	**0.937**	**0.940**
	IFC	0.927	0.923	0.748	0.839	0.569	0.736	0.554	0.707	0.705	0.744	0.636	0.759
	DeepQA	**0.981**	**0.982**	**0.961**	**0.956**	**0.947**	**0.951**	**0.939**	**0.947**	**0.937**	**0.940**	**0.948**	**0.952**
	WaDIQaM-FR	0.970	**0.980**	-	-	-	-	**0.940**	**0.946**	-	-	-	-
	DRF-IQA (proposed)	**0.983**	**0.983**	**0.964**	**0.960**	**0.961**	**0.958**	**0.944**	**0.942**	**0.940**	0.934	**0.955**	**0.952**

algorithms according to the type of distortion. The best three models for each distortion are again shown in bold. Since DRF-IQA is a holistic IQA model, it was trained on all the distortion types, then tested on each different distortion type. Nonetheless, DRF-IQA delivered the best performance on most distortion types. On the TID2013 database, it achieved top-3 prediction accuracy on most of the distortions.

However, since DRF-IQA is currently trained only on luminance, it could not accurately predict CCS, where color information is the major cue of distortion. On the LIVE IQA database, it outperformed most of the other FR/NR-IQA methods by a wide margin. On the CSIQ database, the overall performance of DRF-IQA was again generally superior. On the LIVE MD database, it delivered the best performance by a wide margin.

C. Ablation Study

We also conducted an ablation study. The training and testing sets were again split as the above training procedure. In each test, the SROCC and PLCC scores on the LIVE IQA and TID2008 were obtained, respectively.

We conducted four ablation tests with respect to the contributions of the dynamic error maps, response error maps, importance map and sensitivity map. First, we tested the model without including the dynamic error representation which calculates perceptual error signals from the generated DRFs. To test this model (DRF-IQA *w/o* dynamic error maps) in (3), the visual sensitivity based pooling $h(\cdot)$ function was directly connected to the generated DRFs. In other words, the model works as a simple CNN regression model on the three input maps (distorted, reference and spatial error maps). Second, we removed the response error maps for the two response maps ($\hat{R}_{dis}$ and $\hat{R}_{ref}$), which measure the perceptual distances between the distorted and reference images. To evaluate the ablation of the response error map (DRF-IQA *w/o* response error maps), the response maps of the distorted and reference images were simply concatenated. They were then directly input to the visual sensitivity based pooling $h(\cdot)$ function to regress the predicted score. We then studied the performance of DRF-IQA without the importance map (DRF-IQA *w/o* importance map), DRF-IQA without the sensitivity map (DRF-IQA *w/o* sensitivity map) and the full version of DRF-IQA (DRF-IQA *full*). To test the importance map, the dynamic error maps were calculated using only the response error maps without the importance map in (3). Similarly, to test the sensitivity map, DRF-IQA was trained with dynamic error maps but without weighting by the sensitivity map in (5).

Table IV shows the performance comparison for the four ablation test models against DRF-IQA *full*. As shown in the table, DRF-IQA *w/o* dynamic error map delivered worse performance than the other models, since the perceptual error plays a very important role in the proposed model. Nevertheless, it still operates as a simple CNN-based regression engine with reliable performance. The performance of DRF-IQA *w/o* response error maps was slightly higher than that of DRF-IQA *w/o* dynamic error maps, but it was still less effective than using the distance between the two responses. Moreover, it may be observed that removing either importance map or sensitivity map lowered the performance relative to DRF-IQA *full* on both databases, with the effect being more pronounced on the larger TID2008 dataset.

D. Effect on the Parameters of DRFs

To study the attained performance against the parameters of DRF-IQA, we tabulated the metric scores while varying the number and sizes of the DRFs. Table III shows the performance comparison for three different sizes of DRFs and varying numbers of DRF channels. As tabulated in

TABLE II

SROCC OF COMPARED IQA MODELS ON INDIVIDUAL DISTORTION TYPES. ITALICS INDICATE DEEP LEARNING-BASED METHODS

	Dist.type	BLIINDS-II	BRISQUE	*BIECON*	*DIQA*	PSNR	SSIM	VIF	GSMD	FSIMc	IFC	*DeepQA*	*DRF-IQA*
TID2013	AGN	0.714	0.706	0.913	0.915	**0.934**	0.867	0.880	0.911	0.910	0.773	**0.920**	**0.967**
	ANC	0.728	0.523	0.835	0.755	0.867	0.773	**0.876**	**0.878**	0.854	0.779	0.816	**0.940**
	SCN	0.825	0.776	0.903	0.878	**0.916**	0.852	0.870	**0.913**	0.890	0.767	0.884	**0.941**
	MN	0.358	0.295	0.835	0.734	0.836	0.777	0.868	0.709	0.801	0.730	0.839	**0.869**
	HFN	0.852	0.836	0.931	**0.939**	0.913	0.863	0.907	0.919	0.904	0.881	**0.935**	**0.934**
	IN	0.664	0.802	**0.913**	0.843	**0.900**	0.750	0.833	0.661	0.825	0.647	0.835	**0.861**
	QN	0.780	0.682	0.893	0.858	**0.875**	0.866	0.797	**0.887**	0.880	0.827	0.865	**0.941**
	GB	0.852	0.861	0.953	0.920	0.910	**0.967**	0.954	0.897	0.955	0.933	**0.969**	**0.965**
	DEN	0.754	0.500	0.917	0.788	**0.953**	0.925	0.916	**0.975**	0.933	0.929	0.899	**0.950**
	JPEG	0.808	0.790	**0.943**	0.892	0.922	0.920	0.917	**0.952**	0.934	0.917	0.896	**0.962**
	JP2K	0.862	0.779	**0.956**	0.912	0.886	0.947	0.971	**0.980**	**0.959**	0.952	0.919	0.946
	JGTE	0.251	0.254	**0.862**	0.861	0.806	0.845	**0.859**	**0.862**	0.861	0.806	0.830	0.832
	J2TE	0.755	0.723	0.827	0.812	0.891	**0.883**	0.850	0.883	0.912	0.791	**0.908**	**0.928**
	NEPN	0.081	0.213	0.923	0.659	0.679	**0.782**	**0.762**	0.760	**0.794**	0.572	0.661	0.680
	Block	0.371	0.197	0.401	0.407	0.330	0.572	**0.832**	**0.897**	0.553	0.193	0.324	**0.649**
	MS	0.159	0.217	0.292	0.299	**0.757**	**0.775**	0.510	0.649	0.749	0.372	0.596	**0.926**
	CTC	-0.082	0.079	**0.686**	**0.687**	0.447	0.378	0.819	0.466	0.468	0.424	0.671	**0.915**
	CCS	0.109	0.113	-0.159	-0.151	**0.634**	0.414	0.310	0.358	**0.836**	**0.826**	0.351	0.400
	MGN	0.699	0.674	**0.898**	**0.904**	0.883	0.780	0.847	0.835	0.857	0.879	**0.939**	0.860
	CN	0.222	0.198	0.649	0.655	0.841	0.857	0.895	0.912	**0.914**	**0.912**	0.885	**0.947**
	LCNI	0.451	0.627	0.922	0.930	0.916	0.806	0.920	**0.956**	**0.949**	0.901	0.934	**0.946**
	ICQD	0.815	0.849	**0.935**	**0.936**	0.920	0.854	0.841	0.897	0.882	0.893	0.893	**0.929**
	CHA	0.568	0.724	0.750	0.756	0.880	0.878	**0.885**	0.882	**0.893**	**0.886**	0.859	0.750
	SSR	0.856	0.811	0.903	0.909	0.911	0.946	0.935	**0.967**	**0.958**	0.952	0.920	**0.959**
LIVE	JP2K	0.930	0.914	0.955	0.961	0.895	0.961	0.969	0.968	**0.972**	0.910	**0.972**	**0.978**
	JPEG	0.950	0.965	0.968	0.976	0.881	0.972	**0.984**	0.973	0.979	0.944	**0.980**	**0.982**
	WN	0.947	0.977	0.984	**0.988**	0.985	0.969	0.985	0.974	0.971	0.937	**0.986**	**0.988**
	GB	0.915	0.951	0.955	0.962	0.782	0.952	**0.972**	0.957	0.968	0.965	**0.982**	**0.982**
	FF	0.874	0.877	0.908	0.912	0.891	0.956	**0.965**	0.942	0.950	**0.964**	0.963	**0.967**
CSIQ	WN	0.702	0.682	0.826	0.835	**0.963**	0.897	**0.958**	**0.944**	0.936	0.846	0.904	0.910
	JPEG	0.846	0.846	0.923	0.931	0.888	0.956	**0.971**	**0.963**	**0.966**	0.940	0.948	**0.961**
	JP2K	0.850	0.817	0.920	0.927	0.936	0.961	0.967	0.965	**0.970**	0.926	**0.972**	**0.968**
	PGN	0.812	0.743	0.886	0.893	0.934	0.892	0.951	0.939	0.937	0.828	0.930	**0.942**
	GB	0.880	0.808	0.867	0.870	0.929	0.961	**0.975**	0.959	0.973	0.959	**0.970**	**0.972**
	CTD	0.336	0.396	0.711	0.718	0.862	0.792	0.935	0.935	**0.944**	0.542	**0.956**	**0.963**
LIVE MD	GB+JPEG	0.899	0.903	0.887	0.896	0.736	**0.898**	0.902	0.911	0.885	0.895	**0.956**	**0.936**
	GB+WN	0.892	0.902	**0.932**	**0.941**	0.743	0.912	0.918	0.915	0.899	0.878	**0.952**	0.928

TABLE III

SROCC AND PLCC COMPARISONS AS A FUNCTION OF DRF CHANNEL SIZE AND THE NUMBER OF DRFS, ON THE LIVE IQA AND TID2008 DATABASES

size of DRFs	# of DRFs	LIVE IQA		TID2008	
		SROCC	PLCC	SROCC	PLCC
3x3	8	0.971	0.972	0.935	0.938
	16	0.975	0.972	0.938	0.941
5x5	8	0.977	0.980	0.945	0.944
	16	0.977	0.979	0.947	0.951
9x9	8	0.978	0.975	0.946	0.946
	16	0.983	0.983	0.961	0.958
	32	0.981	0.982	0.960	0.954
	64	0.972	0.970	0.947	0.945

TABLE IV

SROCC AND PLCC COMPARISONS ON THE LIVE IQA AND TID2008 DATABASES

Configuration	LIVE IQA		TID2008	
	SROCC	PLCC	SROCC	PLCC
DRF-IQA *w/o* dynamic error map	0.974	0.960	0.946	0.946
DRF-IQA *w/o* response error map	0.976	0.974	0.947	0.948
DRF-IQA *w/o* importance map	0.977	0.979	0.951	0.952
DRF-IQA *w/o* sensitivity map	0.983	0.980	0.958	0.953
DRF-IQA *full*	0.983	0.983	0.961	0.958

the Table, the overall performance was unarguably high for each single test, but when learning larger DRFs (9×9), as might be expected. when the number of DRFs was varied (8 and 16 DRFs were used for sizes 3×3 and 5×5, while 32 and 64 were also used for 9×9 DRFs). For the 3×3 and 5×5 models, using a larger number of channels yielded higher

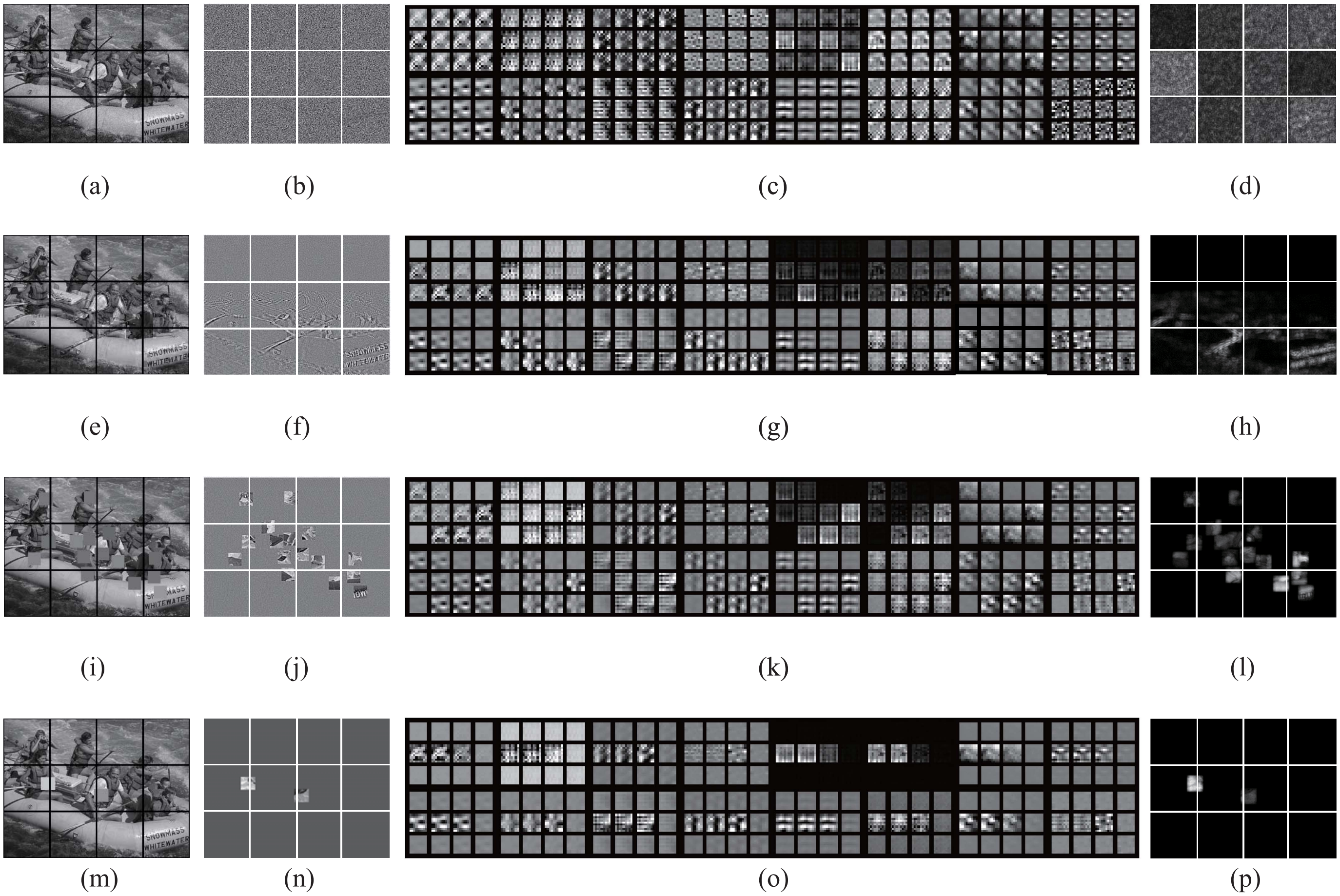


Fig. 8. Examples of predicted error maps using generated DRFs. (a), (e), (i), and (m) are images with three distortion types (AGN, JPTE, and two Block). (b), (f), (j) and (n) are spatial error maps for each distorted image. (c), (g), (k) and (o) are the DRFs generated on each distorted image and (d), (h), (l) and (p) are the error maps.

performance. However, for the 9×9 DRFs, the performance did not improve beyond 16 DRFs, which may be a limit imposed by the network size.

E. Cross Dataset Test

To test the generalization ability of DRF-IQA, we conducted a cross-dataset test. In this experiment, we tested both DRF-IQA *w/o* the sensitivity map and DRF-IQA *full*, as was done on the previous ablation set. The models were trained on a subset of the TID2008 database, then tested on the LIVE IQA database. Since TID2008 contains broader kinds of distortion, we only used five distortion types (JPEG, JP2K, Additive Gaussian noise (AGN), GB, and JPEG transmission errors (JPTE)) to match the LIVE IQA database more closely. The SROCC results are shown in Table V. As may be seen, both DRF-IQA *w/o* sensitivity map and DRF-IQA *full* yielded excellent performances on the LIVE dataset. More importantly, the performance was not biased towards or away from any specific distortion type. On the contrary, when testing on all the distortion types of the TID2008 dataset, after training on the LIVE dataset, DRF-IQA was less able to predict human judgments of distortions it had not been exposed to. This means that the generated DRF did not play an effective role in determining distortion types which are not in the training data set.

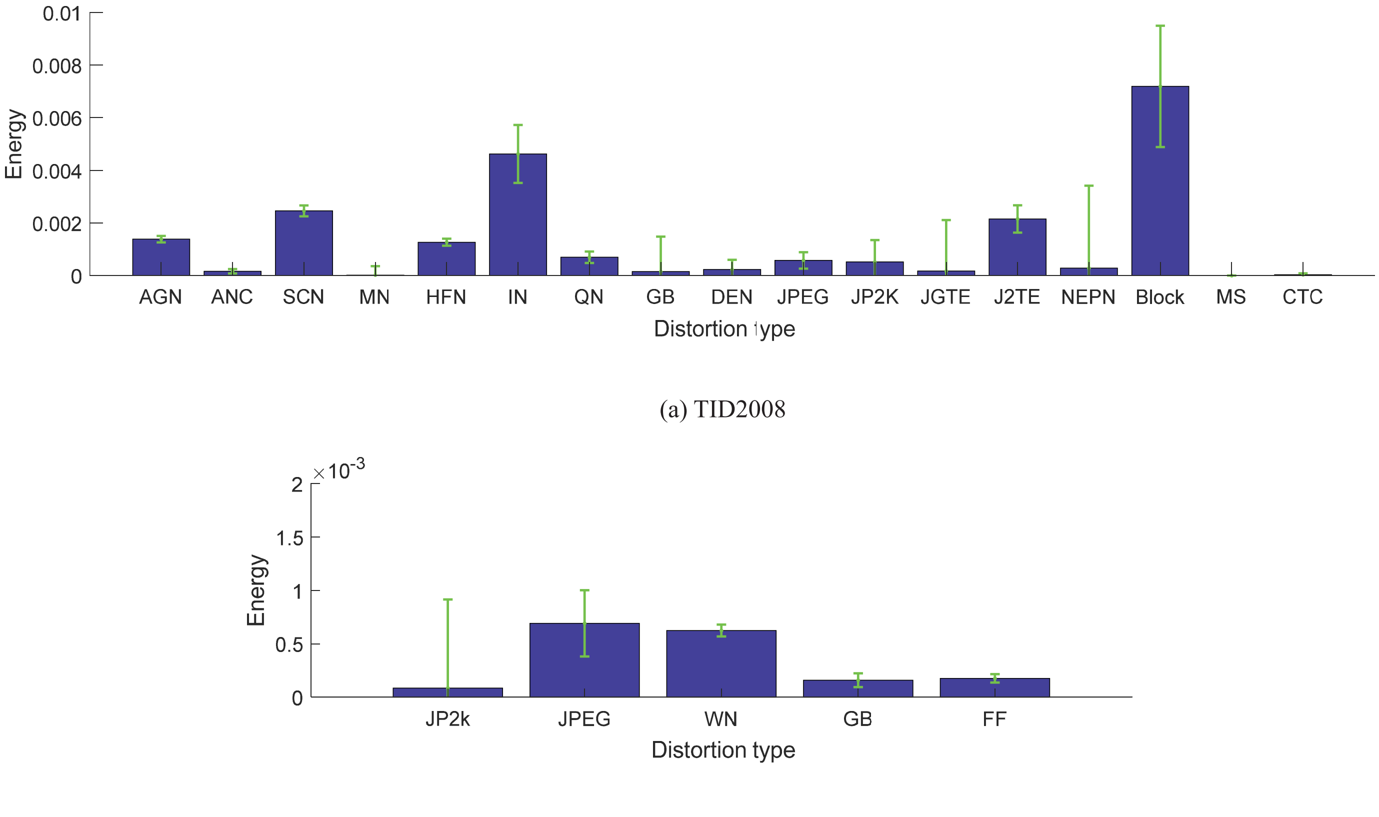
TABLE V
CROSS DATASET TEST ON THE LIVE IQA DATABASE (SROCC)

Models	JP2k	JPEG	WN	GB	FF	**ALL**
DRF-IQA *w/o* sensitivity map	0.8149	0.8028	0.8736	0.8438	0.8552	0.8327
DRF-IQA *full*	0.8213	0.8217	0.8788	0.8224	0.8477	0.8343

F. Visualization Results

Here we visualize the designed DRFs and the error representations generated by Stream 1, while varying the type of distortion. We also analyze the sensitivity map in Stream 2.

1) DRFs: In Fig. 8, the generated DRFs, their response error maps and their spatial error maps are compared for four types of distorted images. The "I14" distorted images from the TID2008 database were used to show the results. For accurate visualization, this image was not included in the training set. The Figure shows the patch partition and the corresponding results of processing via Stream 1. The four images in the first column are images distorted with



(a) TID2008

(b) LIVE IQA

Fig. 9. Energies of the generated DRFs for each distortion type in the (a) TID2008 and (b) LIVE IQA databases, respectively.

AGN, JPTE, and two block distortions, respectively. The second column shows the spatial error maps, where darker regions indicate more distorted pixels. The 16 DRFs generated on each distorted patch images are shown in the third column, and the error maps are in the fourth column. For AGN (a)-(d), the distortion is distributed uniformly over the image, and the generated DRFs also have supported over the entire image. However, for the other distortion types, the generated DRFs are only supported over local areas where errors occur. Interestingly, some of the generated DRFs have similar structures across all the distortion types, while the structures of other DRFs are different, depending on the type of distortion.

We also investigated the coefficient distributions of the DRFs to better understand how distortion-specific DRFs are generated. Toward this, define the energy of the generated DRFs as follows: $E_{DRF} = \sum_n \sum_{x \in \mathbb{N}^2} Q^n(x)^2$, where $Q^n(x)$ are the generated DRFs ($n = 1, \ldots, N$). The energy of each DRF influences the impact of the error maps since humans make non-linear visual quality judgments depending on the type of distortion. In this regard, DRF-IQA generates different DRF energies for each type of distortion. For example, the error representations of the four distorted images in Fig. 8 clearly have different energy distributions. Fig. 9 statistically plots the mean and standard deviation of the energy of the each of the generated distortion-specific DRFs, on TID2008 (Fig. 9 (a)) and on LIVE IQA (Fig. 9 (b)). As may be seen, the DRF energies vary by distortion. In particular, the DRFs generated by extremely unrealistic distortions, such as block distortion have stronger energies than other DRF types. In other words, when the perceptual error signals associated with some distortion type(s) become very strong, the energy of the corresponding generated DRFs significantly increases to better map predictions to non-linear human judgments.

2) Visual Sensitivity: To determine whether the visual sensitivity map agrees with perception, four reference and distorted image pairs, the spatial error maps generated on them, and the corresponding visual sensitivity maps are shown in Fig. 10. The images are impaired by four different distortion types (AGN, MN, QN, and JP2K). Figs. 10 (c), (g), (k), and (o) are the spatial error maps obtained using (1), while Figs. 10 (d), (h), (l), and (p) are the corresponding predicted sensitivity maps. Darker regions in the spatial error maps indicate more distorted pixels. Darker regions in the sensitivity maps indicate decreased sensitivity to distortion. In Figs. 10 (a)-(b), the AGN around the sky region is more noticeable than on the sea regions, as indicated by Fig. 10 (d).

On the image with masked noise in Figs. 10 (e)-(f), regions having high spatial frequencies are assigned lower weights, as shown in Fig. 10 (h). In the case of QN, the objective error is strongly distributed over the entire image (Fig. 10 (k)). However, edge regions are given lower weighting, as shown in Fig. 10 (l). For the image distorted by JPEG 2000 (Fig. 10 (m)), the textured forest region is given reduced sensitivity. Broadly, the visual sensitivity map agrees with the effects of visual contrast masking.

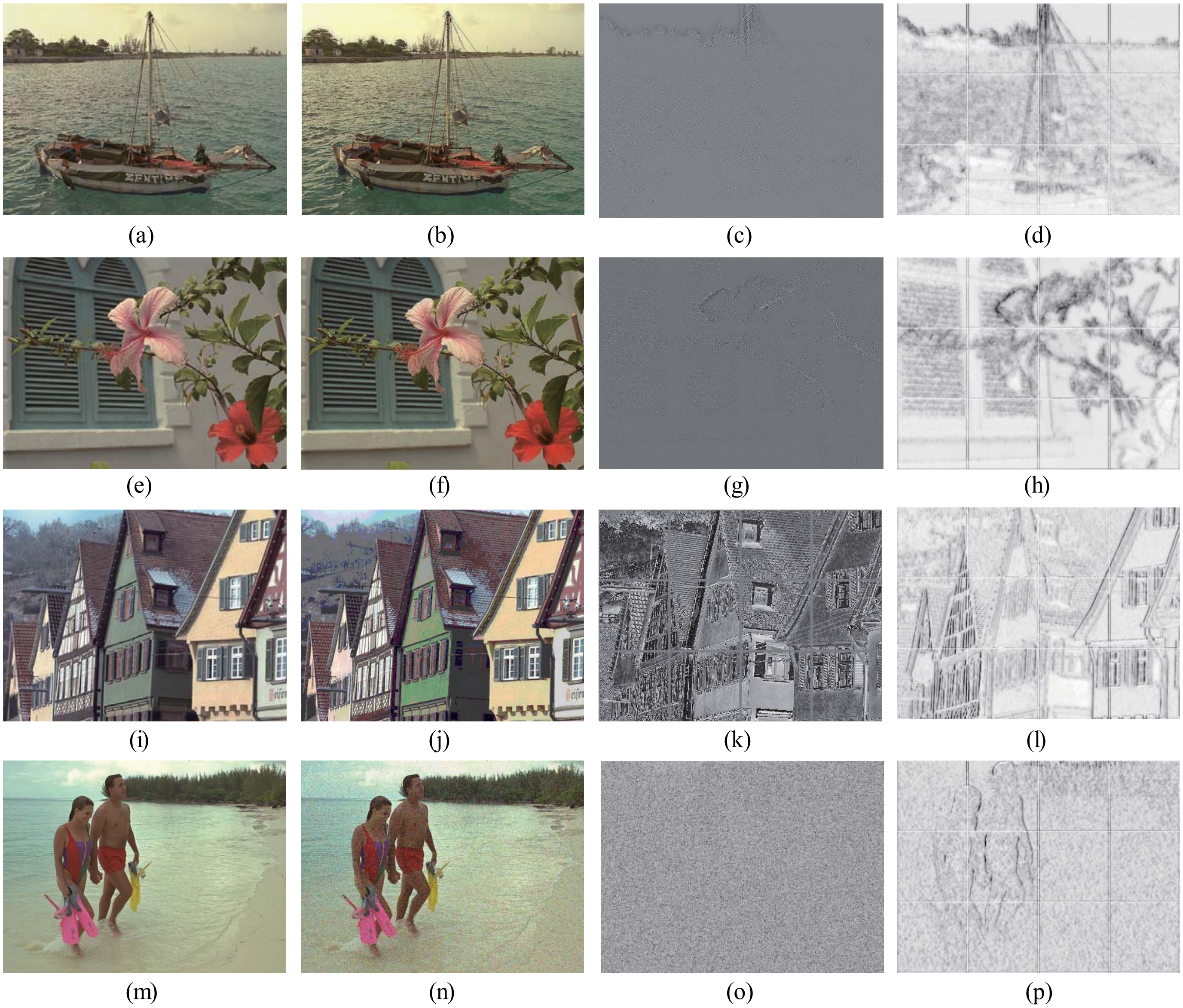


Fig. 10. Examples of generated sensitivity maps. (a), (e), (i) and (m) are four different reference images, while (b), (f), (j) and (n) are distorted versions of them with AGN, MN, QN and JP2K, respectively. The spatial error maps of the distorted images are shown in (c), (g), (k) and (o), while the predicted sensitivity maps computed on the distorted images are shown in (d), (h), (l) and (p).

V. Conclusion

We have proposed a novel approach to the FR-IQA problem using a dual-stream CNN. By dynamically generating receptive fields, the proposed model is able to assess image quality in a manner that closely agrees with perception. Through rigorous simulations, we demonstrated that the predicted DRFs and sensitivity maps agree with perception. Indeed, DRF-IQA achieves state-of-the-art performance on various IQA databases. In the future, we plan to advance the proposed framework for the challenging NR-IQA problem.

References

[1] S. McCarthy, "Quantitative evaluation of human visual perception for multiple screens and multiple CODECs," in *Proc. SMPTE Annu. Tech. Conf. Exhib.*, 2012, pp. 1–8.

[2] W. Kim *et al.*, "Modern trends on quality of experience assessment and future work," *APSIPA Trans. Signal Inf. Process.*, vol. 8, pp. 1–19, Sep. 2019.

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

[4] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[5] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Proc. SPIE*, vol. 5666, pp. 149–159, Mar. 2005.

[6] S. Lee, M. Pattichis, and A. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 977–992, Jul. 2001.

[7] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1200–1203.

[8] L. Lu, Z. Wang, A. Bovik, and J. Kouloheris, "Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2003, pp. 61–64.

[9] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

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

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

[12] H. Sheikh, A. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[13] A. K. Moorthy and A. C. Bovik, "Statistics of natural image distortions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 962–965.

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

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

[16] X. Jia, B. De Brabandere, T. Tuytelaars, and L. Van Gool, "Dynamic filter networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 667–675.

[17] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1233–1258, Dec. 1987.

[18] K. Lee, A. K. Moorthy, S. Lee, and A. C. Bovik, "3D visual activity assessment based on natural scene statistics," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 450–465, Jan. 2014.

[19] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[20] S. Li, L. Ma, and K. N. Ngan, "Full-reference video quality assessment by decoupling detail losses and additive impairments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1100–1112, Jul. 2012.

[21] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[22] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3224–3232.

[23] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[24] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Jun. 2015.

[25] 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 *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2344–2348.

[26] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, Jan. 2017.

[27] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[28] H. Oh, S. Ahn, J. Kim, and S. Lee, "Blind deep S3D image quality evaluation via local to global feature aggregation," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4923–4936, Oct. 2017.

[29] S. Seo, S. Ki, and M. Kim, "JND-SalCAR: A novel JND-based saliency-channel attention residual network for image quality prediction," 2019, *arXiv:1902.05316*. [Online]. Available: https://arxiv.org/pdf/1902.05316

[30] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1676–1684.

[31] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.

[32] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 219–234.

[33] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400

[34] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[35] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[36] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[38] P. Esser and E. Sutter, "A variational U-net for conditional appearance and shape generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8857–8866.

[39] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[41] D. M. Chandler,"Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.

[42] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Conf. Rec. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012.

[43] *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II*, VQEG, Boulder, CO, USA, 2003.

[44] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Jul. 2004, pp. 1398–1402.

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

[46] H. Sheikh, A. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

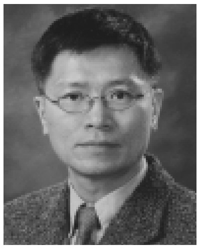
[47] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

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

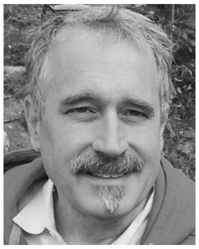
[49] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

Woojae Kim received the B.S. degree in electronic engineering from Soongsil University, Seoul, South Korea, in 2015. He is currently pursuing the joint M.S. and Ph.D. degrees with the Multidimensional Insight Laboratory, Yonsei University. He was a Research Assistant with the Laboratory for School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore, in 2018, under the guidance of Prof. W. Lin. His research interests include image and video processing based on the human visual system, image/video quality assessment, computer vision, and machine learning.

Anh-Duc Nguyen received the B.Eng. degree in automatic control from the Hanoi University of Science and Technology, Vietnam, in 2015. He is currently pursuing the joint M.Sc. and Ph.D. degrees with Yonsei University, South Korea. His research interests are image/video analysis, geometric computer vision, and deep learning.



Sanghoon Lee (Senior Member, IEEE) received the B.S. degree from Yonsei University, Seoul, South Korea, in 1989, the M.S. degree from the Korea Advanced Institute of Science and Technology, South Korea, in 1991, and the Ph.D. degree from The University of Texas at Austin, Austin, TX, USA, in 2000. From 1991 to 1996, he was with Korea Telecom, South Korea. From 1999 to 2002, he was with Lucent Technologies, Morris Plains, NJ, USA. In 2003, he joined the EE Department, Yonsei University, as a Faculty Member, where he is currently a Full Professor. His current research interests include image/video processing, computer vision and graphics. He is a Board of Governors Member of APSIPA in 2020. He received the 2015 Yonsei Academic Award from Yonsei University, the 2012 Special Service Award from the IEEE Broadcast Technology Society, and the 2013 Special Service Award from the IEEE Signal Processing Society. He was the General Chair of the 2013 IEEE IVMSP Workshop, and has been served as steering committees for the IEEE and APISPA conferences. He was the IVM Technical Committee Chair of APSIPA from 2018 to 2019. He has been serving as the Chair of the IEEE P3333.1 Quality Assessment Working Group since 2011. He was the IEEE IVMSP Technical Committee from 2014 to 2019. He has been the IEEE MMSP Technical Committee since 2016. He also served as an Editor of the *Journal of Communications and Networks* from 2009 to 2015 and a special issue Guest Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING in 2013. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2010 to 2014. He served as an Associate Editor from 2014 to 2018. He is currently a Senior Area Editor of the IEEE SIGNAL PROCESSING LETTERS.



Alan Conrad Bovik (Fellow, IEEE) is currently a Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. His research interests include image processing, digital television, digital streaming video, and visual perception. He was a recipient of the 2019 Progress Medal from The Royal Photographic Society, the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from the Optical Society of America, the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, and the Norbert Wiener Society Award and the Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. A perennial Web of Science Group Highly-Cited Researcher, he has also received about ten best journal paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. His recent books include The Essential Guides to Image and Video Processing. He co-founded and was the longest-serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING, and also created/chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, in 1994.