Blind Quality Estimation by Disentangling Perceptual and Noisy Features in High Dynamic Range Images

N.K. Kottayil, Giuseppe Valenzise, Frederic Dufaux, Irene Cheng

Blind Quality Estimation by Disentangling Perceptual and Noisy Features in High Dynamic Range Images

Navaneeth Kamballur Kottayil, Giuseppe Valenzise, *Senior Member, IEEE,* Frederic Dufaux, *Fellow, IEEE*, Irene Cheng, *Senior Member, IEEE*

***Abstract*—High Dynamic Range (HDR) image visual quality assessment in the absence of a reference image is challenging. This research topic has not been adequately studied largely due to the high cost of HDR display devices. Nevertheless, HDR imaging technology has attracted increasing attention because it provides more realistic content, consistent to what the Human Visual System perceives. We propose a new No-Reference Image Quality Assessment (NR-IQA) model for HDR data based on convolutional neural networks. The proposed model is able to detect visual artifacts, taking into consideration perceptual masking effects, in a distorted HDR image without any reference. The error and perceptual masking values are measured separately, yet sequentially, and then processed by a Mixing function to predict the perceived quality of the distorted image. Instead of using simple stimuli and psychovisual experiments, perceptual masking effects are computed from a set of annotated HDR images during our training process. Experimental results demonstrate that our proposed NR-IQA model can predict HDR image quality as accurately as state-of-the-art full-reference IQA methods.**

I. Introduction

High dynamic range (HDR) images can present a much larger range of luminance compared to conventional images. This larger range of luminance is achieved by using 16-32 bit floating point values, instead of a conventional 8 bits per pixel integer representation. Viewers are able to perceive more vivid colors and scene content compared to viewing on a conventional Standard Dynamic Range (SDR) or Low Dynamic Range (LDR) display. This generates a better quality of viewing experience [1]. The advancement of HDR imaging technology has changed the landscape of the entire multimedia communication pipeline from capturing, processing and transmission, to the visualization of HDR content [2]. This technology has become an important development in the consumer market, e.g., TV and photography, with the support of industrial investments.

In this work, we consider evaluating HDR image quality on a HDR compatible display. Image quality assessment (IQA) can be broadly categorized into Full-Reference (FR) and No-Reference (NR). In FR-IQA, the quality of a given image is evaluated by comparing a distorted image with an undistorted version of the same image. In NR-IQA, the quality is evaluated by judging the distorted image only. Since the target audience of the HDR content is the Human Visual System (HVS), one IQA solution is to conduct user subjective tests. However, subjective tests are tedious and time-consuming. Even with the help of massive crowdsourcing projects, e.g., mturk, HDR IQA is difficult due to the high cost of acquiring systems capable of displaying HDR content. A better solution is to develop an efficient NR-IQA model, which can automatically assess HDR content, matching the human perception.

Our proposed model is capable of predicting the perceived HDR image quality and localizing the distortions. We use a convolutional neural network (CNN) based architecture to achieve this goal. Our model addresses low-level image distortions such as artifacts caused by image compression. We do not consider changes in image quality due to high-level effects, e.g., artistic intent, where complex aesthetical considerations need to be taken into account.

Our contributions in HDR image quality assessment lie in:

1) Proposing an NR-IQA model based on a convolutional neural network architecture, which can separate pixelwise errors from their impact on perception in a distorted image. Our model outperforms other NR-IQA models and is competitive with state-of-the-art HDR full-reference IQA algorithms.
2) Providing an accurate error prediction in a distorted image without a reference image.
3) Predicting the visual masking effects without the need of explicit psychovisual subjective tests.

The rest of this paper is organized as follows: Section II discusses the previous work in related fields. Section III explains our motivation for the new approach. In Section IV we discuss the conceptual idea and implementation of the proposed method. Section V deals with performance comparisons and test of our method against other algorithms. Followed by this, in Section VI, we discuss the functionality of each of the subcomponents of our system and show how they work. Section VII concludes the paper.

II. Previous work

In pace with the rapid development of HDR imaging technology, Full-Reference Image Quality Assessment (FR-IQA) of HDR images has been gaining attention in recent years, and a number of studies have addressed the evaluation of quality for HDR images and video [3], [4], [5]. However, much less

N.K. Kottayil and I. Cheng are with Dept of Computing Science, University of Alberta, Edmonton Canada

G. Valenzise and F. Dufaux are with the Laboratoire des Signaux et Systèmes (L2S), CNRS, CentraleSupélec, Université Paris-Sud.

has been done on No-Reference Image Quality Assessment (NR-IQA) of HDR images. Relevant work in the NR-IQA literature can be classified into two general categories:

- LDR NR-IQA, where the quality of images is estimated when it is visualized on LDR displays.
- Tone-Mapped NR-IQA, in which color and contrast values of a HDR image are mapped onto a smaller range of color and contrast values, using Tone Mapping Operators (TMO), and the output is evaluated on LDR screens. TMOs are often used to compress the dynamic range of HDR images.

Note that in our work, we assume instead that HDR images are directly evaluated on HDR displays.

A. LDR NR-IQA

Machine learning approaches are often used in LDR NR-IQA. These approaches start by creating a feature image and fit a relevant distribution. The parameters of this distribution are used as the feature vector of the distorted image. An early method based on machine learning in LDR NR-IQA is the Blind Image Quality Index (BIQI) [6]. It is a *two-step process*: (1) from a set of features, a Support Vector Machine (SVM) predicts the type of distortion, and (2) a set of Support Vector Regressors (SVRs) predict the score for each distortion type. The final quality score is computed as

$$score = \sum_{i=1}^{m} p_i \cdot q_i, \tag{1}$$

where m is the distortion type, p_i represents the probability of each distortion obtained from the SVM and q_i represents the quality score given by each of the SVR's. BIQI [6] used Daubechies 9/7 wavelet as the feature image.

Many methods follow a similar approach and show good performances on assessing LDR content on LDR screens without a reference image. Examples are Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [7], Distortion Identification-based Image Verity and INtegrity Evalutation (DIIVINE) [8] and Spatial-Spectral Entropy based Quality metric (SSEQ) [9]. These algorithms can have complex features, but just need a single SVR to predict the final quality of an image.

BRISQUE [7] computes the Mean Subtracted Contrast Normalized (MSCN) image as feature using $MSCN(i,j) = \frac{I(i,j)-\mu_{I,N,i,j}}{\sigma_{I,N,i,j}+1}$, where $\mu_{I,N,i,j}$ and $\sigma_{I,N,i,j}$ represent the mean and variance computed over a local Gaussian window of size N around the point (i,j). $I(i,j)$ is the image intensity at (i,j). DIIVINE [8] uses divisive normalized steerable pyramid decomposition coefficients to create the feature image.

SSEQ [9] generates features using entropy. A scale space decomposition is used to generate three scales of an image, and entropy is calculated for each image block in the spatial and DCT domain. The entropies are then pooled by percentile pooling. The mean and variance of the spatial and frequency components are used in the feature vector.

An alternative approach in LDR NR-IQA is using Convolutional Neural Networks (CNN) as in [10]. We refer to this method as kCNN throughout the paper. This is an enhanced version of [11]. The basic idea is to learn discriminative features that can perform IQA rather than using a handcrafted method; [11] uses dictionary learning to form discriminative filters and [10] improves the process by redesigning it as a CNN. The CNN has four layers that act on MSCN image blocks of size 32×32. The first layer is a convolution layer with 50 filters (kernels), followed by a pooling layer that reduces the dimensionality of the data, and finally there are two fully connected layers. The network is trained with the Mean Opinion Scores (MOS). This method can produce a "perceptual distortion map", which shows location of the errors of the distorted image.

In general, these LDR NR-IQA algorithms rely largely on statistical characteristics of the distorted images, i.e., they assume image distortions alter the statistical properties exhibited by "natural" undistorted images. However, they do not take in to consideration the human visual resistance to errors, e.g., due to masking phenomena. The natural image statistics are to be captured in the internal representation of the SVM or CNN. Thus, these LDR NR-IQA algorithms need an explicit training stage in order to learn what "natural" is and how the noise change the "naturalness" property. A recent work that tries to alleviate this problem is [12], which assesses perceived image quality corrupted by uniform and high-frequency noise. The method uses a combination of features, with appropriate feature weights to scale the errors. However, it cannot estimate compression errors. In contrast, we formulate the naturalness property in our model and our technique can also assess compression errors.

B. Tone-Mapped IQA

As explained earlier, Tone-Mapped IQA is a related research area, where HDR images are tone-mapped to LDR images, which are then evaluated on LDR displays. The Tone-Mapped Quality Index (TMQI) metric [13] follows the structural fidelity criterion [14], to compare an HDR image with its tone-mapped version, by embedding the knowledge of the Contrast Sensitivity Function (CSF) at different values of luminance [15], [16]. In addition, similar to NR-IQA, a naturalness measure is also included to compare the statistics of the tone-mapped image to those of natural images. This idea is further explored in [17], where the performance is improved with better error pooling and naturalness measure. Phase congruency is added as a feature in [18] for the same purpose to compare two images.

A NR-IQA approach for Tone-Mapped HDR [19] employ MSCN images as spatial domain features. It also obtains gradient computations on different neighborhoods of every pixel. This is followed by Gaussian parameter extraction and a Support Vector Regressor (SVR) process like the other techniques described in Section II-A. The idea is to generate a HDR image by fusing images captured using multiple exposures. The HDR image is then tone-mapped to a LDR image. The groundtruth Mean Opinion Score (MOS) is obtained based on subjective quality evaluation using LDR displays. The method is statistics-based and does not incorporate perceptual modeling. In contrast, our evaluations are based on displaying

HDR images on compatible HDR displays and the human perception component is an integral part of our model. It should be understood that a tone-mapped HDR image has a reduced gamut of colors and luminance compared to the original HDR image. Experimental evidence of psychophysical differences in viewing HDR and LDR image content is provided in [20]. Their study demonstrates how HDR and LDR (or reduced dynamic range) contents are perceived differently when displayed on a HDR screen. The authors collected opinions, including users' ratings of naturalness, visual appeal, spaciousness, and visibility. Here, visibility refers to the details in the image. The study found statistically significant difference in how users rated visibility for HDR and LDR images when these are displayed on a HDR screen.

III. Motivation

Conventional LDR displays have a maximum luminance of about 300 cd/m^2. High Dynamic Range (HDR) displays have a luminance of 4,000 cd/m^2 and above, which delivers more realistic scenes and vivid content to the Human Visual System (HVS). In addition to advanced HDR acquisition devices, HDR images can be generated using multi-exposure fusion algorithms [21] and tone-mapped onto lower dynamic range images [22], [23] for evaluation on LDR displays. Current NR-IQA methods focus on quality assessment on LDR displays. If these methods are applied to predict the quality of HDR images displayed on a HDR display, the result is not accurate, as shown in Section V. This is because they rely on statistical modeling of noise and fail to take into account how the Human Visual System (HVS) responses to HDR displays [24]. Therefore, to design a robust HDR NR-IQA system, which is consistent with how the HVS perceives real-world content, we need a new NR-IQA model that incorporates perceptual factors to predict the visibility of error on HDR screens. To date, we are not aware of any related work that is designed for HDR viewing conditions.

A. HVS response to HDR displays

How the HVS responses differently to HDR and LDR displays has been extensively studied by Aydin et al. [24]. In their experiments, the authors used a LDR display with luminance range 1-100 cd/m^2 and another HDR display with luminance range 10-1,000 cd/m^2 to evaluate the difference in perceived image qualities. They found that a distorted image, when viewed on the brighter display, was perceived as worse compared to the same image displayed on the display with lower brightness. This shows that viewing content on LDR displays and HDR displays has different perceptual effects.

Hence, we argue that, in addition to statistical comparison, HDR NR-IQA should incorporate the psychophysical phenomena that can determine the perception of distortion in HDR conditions. Although such perceptual approach is not common in NR-IQA, it is often used in FR-IQA models. HDR-VDP-2.2 [3] and HDR-VQM [4] are two examples. HDR-VDP 2.2, simulates the early processing stages of the HVS, based on psychophysical measurements. HDR-VQM uses sub-band decomposition and spatial-temporal error pooling to simulate visual recognition. Both are FR-IQA methods, which generate a local error visibility map. An alternative approach is to apply SSIM [25], which is a FR-IQA method for PU-encoded HDR data. These are the best performance algorithms in FR-IQA for HDR data based on the survey presented in [5]. We will compare the performance of our proposed NR-IQA system with these FR-IQA.

IV. Proposed system

In order to design a perceptual model consistent with how the HVS perceives HDR real-world content, it is necessary to understand the contrast sensitivity associated with a complex image. The traditional approach considers various visual features like contrast, frequency and background luminance. Very often, a handcrafted function is used to compute a quality score based on a combination of these features. The quality score function is obtained by fitting opinions collected from subjects in psychophysical experiments using sample datasets. The research question is how to generalize user study results for all real-world images.

We approach this problem of designing a perceptual HDR NR-IQA model by dividing the visual quality analytic process into sub-components. We represent visual quality perception as the result of two functional units. The first unit takes a distorted image and detects error, and the second unit performs a perceptual scaling of this error to compute a quality score. By using a supervised learning approach, the mathematical behavior of these two units can be modeled. The data required for this training is obtained from an IQA dataset, which contains images and the corresponding quality scores.

A. Model Overview

To formulate the above idea, we design a Convolutional Neural Network (CNN) (Fig. 1) that processes HDR image blocks composed of linear luminance values. We use a block size of 32x32 pixels. This is the same block size that was suggested in [26]. Our CNN model has three major parts: E-net, P-net and a Mixing function. E-net estimates the *Error* $\delta(i,j)$ of an image block centered at (i,j). P-net computes the *Perceptual Resistance* $T(i,j)$ of the block. The output of these two systems are then input to a *Mixing function*, to produce the local block quality. We obtain Differential Mean Opinion Scores (DMOS) for each image block. The block scores are then combined to generate the final image quality score. In our model, DMOS is a number directly proportional to the level of distortion in a HDR image.

B. E-net Error Estimation

The Error $\delta(i,j)$ quantifies the change in statistics in a distorted image block. For an image block centered at (i,j), we define the error as,

$$\delta(i,j) = mean(\left|Y_R(i,j) - Y_D(i,j)\right|) \tag{2}$$

where Y_R and Y_D are, respectively, the original and distorted linear HDR luminance values of the image block centered at

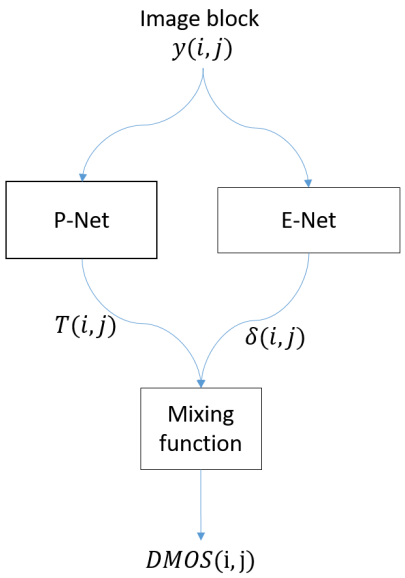

Fig. 1: Proposed strategy for a robust HDR NR-IQA. E-net detects the error, P-net detects the perceptual resistance, and Mixing function consolidates the results and computes a DMOS.

(i, j). This does not indicate a Full Reference computation, as the original version is only used during training (a pair of HDR image and its distorted version is used here). The objective is to train E-net with the distortion characteristics, like blocky artifacts, blurring effects, jagged edges, etc.

We use ℓ_1 norm for error computation (Eq. 2) instead of alternative measures such as ℓ_2 norm, to avoid over-emphasizing large errors. This is particularly important in HDR images where the histogram of Y is generally very skewed and some pixels take very high luminance values. We use our own CNN approach to design E-net to obtain and estimation $\hat{\delta}(i, j)$ of the error in Eq. (2).

C. P-net Perceptual Resistance

For each image block centered at (i, j), we compute the *Perceptual Resistance* $T(i, j)$. This value represents the difficulty for a viewer to perceive the error $\delta(i, j)$ of the block. A high T value implies that it is less likely to see the error, and hence the quality of the block is less affected (high perceptual resistance). Conversely, a low value implies that the image block will be perceptually degraded by error.

Perceptual Resistance $T(i, j)$ aims to represent a combination of all perceptual effects exhibited by an image block centered at (i, j). Though it is functionally similar to the pixel-wise just noticeable error measure used in conventional IQA systems like [12], [27] and [3], we introduce Perceptual Resistance as a new term because our model generates local quality scores (DMOS), as opposed to a local probability of error detection.

Instead of following the traditional perceptual modeling method of deriving perceptual thresholds from psychophysical experiments, we solve this problem by a data driven method. We use a convolutional neural network (CNN) based architecture, P-net, to derive the Perceptual Resistance of a block. The *CNN automatically computes the features required to do this task by a training process using real-world images*. A detailed analysis of how a generic CNN generates its features is explained in [28]. In our system, P-net approximates a function that maps the image block values onto a perceptual resistance value. Differing from a conventional psychophysical perceptual function, the 'function' that is captured by P-net is represented as weights in the CNN.

D. Mixing function

We use a *Mixing function* $f(\hat{\delta}, T)$, which combines the estimated error and Perceptual Resistance to generate a quality score. This is a critical part of the system because it is this value that is optimized by the training process to match human quality scores. The output of P-net would change based on how the Mixing function is designed.

The Mixing function is designed as follows, with error expressed in multiples of Perceptual Resistance:

$$DMOS = f(\hat{\delta}, T) = G\left(\frac{\hat{\delta}}{T}\right), \tag{3}$$

where G is a monotonically increasing function. By using this, we express error in *JND like* measure ($\frac{error}{JND}$), so that the error quantity is mapped onto a more perceptually relevant scale. Such interpretation is common in IQA literature, e.g., [12], [24] and [3].

Mapping $\frac{\hat{\delta}}{T}$ to quality scores is achieved by the function G. Since increase in visible error always leads to decrease in quality and increase in DMOS, the latter must monotonically increase with $\frac{error}{JND}$, implying that G also has to be monotonically increasing with $\frac{\hat{\delta}}{T}$. Thus, any monotonically increasing function is sufficient for G. However, choosing a G that is too complex can lead to optimization problems because of unstable data points along the function, or low values for gradients, leading to slow or zero learning. We do not go into the mathematics of CNN convergence and optimization functions as it is beyond the scope of this work.

Based on the above considerations, we use $G(x) = 1 - \exp(-|kx|)$ and DMOS is computed as:

$$DMOS(i, j) = 1 - \exp\left(-\left|\frac{k * \hat{\delta}(i, j)}{T(i, j)}\right|\right) \tag{4}$$

This choice is inspired by the error model proposed in [12], but we introduce a scaling factor k. Here the added parameter (weight) k can be tuned during the training process, so that the predicted values of DMOS are as close to the ground truth DMOS as possible.

As seen from the plot of Eq. 4 in Fig. 2, G can characterize different rates of increase for DMOS, depending on the values of $\hat{\delta}(i, j)$, $T(i, j)$ and k. Note how k serves to control how slowly DMOS increases with $\frac{\hat{\delta}}{T}$ (Fig. 2 B).

The block-wise DMOS scores obtained from Eq. (4) is converted to a global score by averaging. A weighted scheme is not required here since the perceptual scaling of errors based on content is already handled by Eq. (4) (the T computed by the CNN changes based on image content and handles content-dependent scaling).

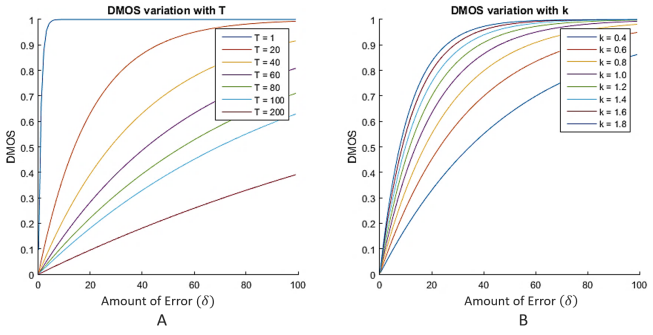

Fig. 2: Behavior of the Mixing function. (A) Varying T with k fixed at 1 illustrates the impacts of changing δ and T on DMOS. (B) Varying k with T fixed at 20 illustrates the impacts of changing k and δ values on DMOS. An optimal k value is determined during the training process.

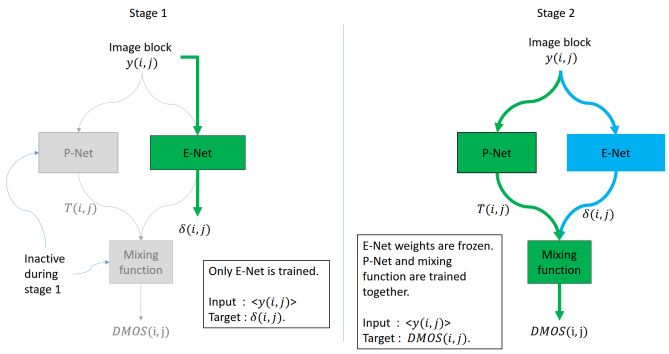

Fig. 3: Two-stage training process.

E. Two-Stage Training process

An important element in a CNN-based system is the selection of right labels for training. To force the desired behavior of the sub-components, we need to provide the right examples to each of the CNN's.

E-net detects blockwise errors. It is trained with linear luminance values of the distorted image as input, and per pixel errors (Eq. (2)) as output, which are available in the training stage.

For P-Net, the ideal training data is a numeric quantity, encapsulating all perceptual effects on the HVS, generated from an image block. Although we cannot get such a final value directly, our system can produce a quality score after the Mixing function process. We use this score for training. This two-stage training forces the P-net to extract a set of perceptual features from the image blocks and to derive a single final Perceptual Resistance value. The optimal value of k in the mixing function is also computed.

We therefore define our two-stage training process as follows:

Stage 1: E-net is trained with distorted image blocks as input and errors δ as target. The error is computed with Eq. (2).

Stage 2: all the training weights of E-net are frozen by setting their learning rate to zero. The whole network is then trained with image blocks as input and ground truth image quality of the whole image, $DMOS_{gt}$, as target. We use J as the cost function for any image block centered at (i, j), where

$$J(i,j) = |DMOS(i,j) - DMOS_{gt}|. \quad (5)$$

$DMOS(i, j)$ is the output of the Mixing function. The P-net and the mixing function (optimal value of k) get trained during this stage.

The overall process is illustrated in Fig 3.

Notice that in Eq. (5) we assume that the local quality of an image block is the same as the global image quality score, similarly to the setting in [26]. While this assumption is somehow inaccurate (as distortion can be unevenly spread across a picture), it has been proven to be accurate enough to predict image quality without reference [26]. We leave to future work how to produce more accurate groundtruth quality labels for training, e.g., as in the very recent method [29].

F. Network Architecture

Our network architecture has two major advantages. First, a separation of the perceptual component from the physical error gives more intuitive results that can be used in applied fields of IQA like image and video compression. With adequate calibration, Perceptual Resistance values can be used to optimize compression or transmission. Secondly, it simplifies the learning process leading to *improved results* and *better generalization of those results* to real-world cases.

E-net is a typical CNN architecture consisting of five layers. The E-net structure is shown in Fig. 4. Spatial pooling is used after each filtering stage. The final layer consists of one node corresponding to the output. Spatial dropout layers [30] are added to prevent over-fitting of data.

For P-net, we define a customized CNN layer, namely Augmented Input Layer. In this layer, in addition to the original luminance values of the block, we compute the mean, variance and MSCN images. For the MSCN image, we use the formulation in [7], i.e., $MSCN(y_N(i,j)) = \frac{y_N(i,j) - \mu_{y_N(i,j)}}{\sigma_{y_N(i,j)} + 0.01}$, where $\mu_{y_N(i,j)}$ is the mean and $\sigma_{y_N(i,j)}$ is the variance. They are computed by replacing every pixel (i, j) with the mean and variance, respectively, over a local Gaussian window of size N around (i, j). We use a smaller value 0.01 as the stabilizing constant in order not to significantly impact the MSCN values. Since neural network training requires the input values in a specific range, we scale the input, variance map and the MSCN map with trainable scalar weights, w_i, $i = 1, \ldots 4$, whose values are determined as part of the overall optimization process. Hence the output of the Augmented Input Layer are four scaled feature maps: $w_1 \cdot y_N(i,j)$, $w_2 \cdot \mu_{y_N(i,j)}$, $w_3 \cdot \sigma_{y_N(i,j)}$, $w_4 \cdot MSCN(y_N(i,j))$. The subsequent layers are convolutional and fully connected layers. The final layer has one node corresponding to the output. The P-net structure is shown in Fig. 5.

The results of E-net and P-net are integrated and analyzed by the Mixing function, whose behavior is modeled by Eq. 4, with trainable parameter k tuned during the training process.

V. Experiments and Results

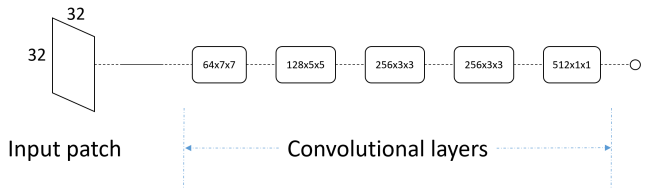

Fig. 4: E-net structure for error estimation.

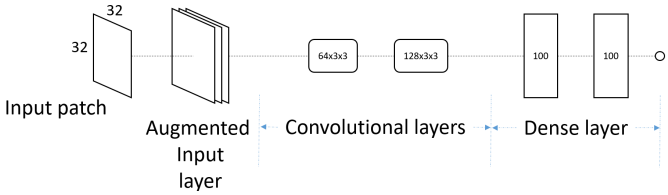

Fig. 5: P-net structure for perceptual resistance computation.

In this section, we compare the performance of our algorithm with existing methods and show a clear improvement in performance. We conduct two tests: 1) a test of overall performance of the proposed method on a large dataset of subjectively annotated HDR images; 2) a cross-dataset test to assess the generalization capabilities of the proposed approach.

A. Datasets

To obtain a large enough number of HDR images, we combine subjective scores from five different datasets [31], [32], [33], [34] and [5]. The authors of the respective datasets conducted subjective evaluations using different HDR displays, of different maximum luminance and viewing distances. The characteristics of the datasets are described in Table I. In order to align the subjective scores of the different datasets, we follow the same setup as the evaluation of HDR FR-IQA in [5], i.e., we employ the iterated nested least square algorithm (INLSA) proposed by Pinson and Wolf [35]. Details about the application of this method for the considered datasets are given in [5]. All the experiments in the following are done using the aligned scores obtained after INLSA.

Dataset Number	Number of Reference Images	Number of Distorted Images	Distortion type	Maximum Luminance (Cd/m^2)
#1 [31]	27	140	JPEG	1000
#2 [32]	29	210	JPEG 2000	1000
#3 [33]	24	240	JPEG-XT	4000
#4 [34]	15	50	JPEG JPEG2000 JPEGXT	4000
#5 [5]	15	50	JPEG JPEG2000 JPEGXT	4000

TABLE I: Datasets characteristics

The datasets provide only MOS values of the images. Since our system requires the difference of mean opinion scores (DMOS), we convert MOS to DMOS as follows:

$$DMOS_{gt}(i) = \frac{MOS_{MAX} - MOS(i)}{MOS_{MAX}}, \qquad (6)$$

where $DMOS_{gt}(i)$ is the ground truth DMOS score for image i, MOS_{MAX} represents the maximum MOS in the IQA training dataset and $MOS(i)$ is the MOS of the i^{th} image of combined database after INLSA alignment.

B. Experimental setup

The proposed system was implemented on a computer with an Intel core i7 processor, 16GB RAM, and a Nvidia GTX1070 graphics processor. Python was used as the programming language with Keras on Theano backend, Imageio and Open CV as supporting libraries. We used Adam optimizer ([36]) to optimize the weights of the CNN. The parameter values of Adam was learning rate $= 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and decay $= 0.0$. The batch size used was 200. The training was done for 10 epochs.

Performance comparison measures used were Spearman Rank order Correlation Coefficient (SRCC), Kendall Rank-order Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE). A good NR-IQA is characterized by a higher value for SRCC, KRCC, and PLCC, and a lower value for RMSE. The metric values and subjective DMOS were scaled to [0,100] for evaluation of RMSE.

C. IQA Reference Schemes

Since the research on HDR NR-IQA is still at its preliminary stage and there is no generally accepted benchmark metric, we compared our approach with a number of state-of-the-art LDR NR-IQA methods: BRISQUE [7], SSEQ [9], BIQI [6], DIIVINE [8], and kCNN [10], with and without pre-processing operators. The results were obtained after retraining the algorithms on the respective processed HDR datasets. PU encoding was used as a pre-processing operator because it has been shown to perform well in a similar context for HDR FR-IQA [34]. We also used a number of tone-mapping operators, which include [23], [37], [38] and [39], in pre-processing. Features were extracted using the implementation provided by the authors. In the case of SSEQ [9], we normalized the images by the maximum intensity under the respective schemes (4000 for linear HDR and 455 for PU encoded data). For training the SVM, a grid search on the cost and kernel parameters of the SVM was conducted for a range of 10^{-15} to 10^{15} before training. The algorithm in [40] was re-implemented using Python.

D. Overall performance

In order to test the overall performance of the proposed method, we (re)train each algorithm on our combined image dataset, by splitting it in training/testing subsets (80% for training and 20% for testing). We repeat this procedure 100 times, similar to [40]. We assure that the training and testing sets do *not* contain the same contents. Note that the results can vary slightly since the CNN weights initialization is random. The computed median SRCC, KLCC, PLCC and RMSE are shown in Table II.

Considering the NR-IQA originally designed for LDR content, we see an acceptable performance in SRCC (about 0.7

Scheme	Processing	SRCC	KRCC	PLCC	RMSE
BRISQUE	Lin	0.7274	0.5430	0.7231	18.1797
	PU	0.8047	0.6116	0.7825	17.3576
	TMO - Drago	0.7374	0.5415	0.7203	19.1261
	TMO - Reinhard 02	0.7782	0.5853	0.7699	18.1523
	TMO - Reinhard 05	0.6903	0.5061	0.6643	20.3307
	TMO - Mantiuk	0.6172	0.4559	0.6148	22.1868
SSEQ	Lin	0.6022	0.4378	0.6008	23.3017
	PU	0.7342	0.5451	0.7175	19.4117
	TMO - Drago	0.6853	0.5011	0.6954	20.8766
	TMO - Reinhard 02	0.6866	0.5183	0.6688	21.0673
	TMO - Reinhard 05	0.6568	0.4845	0.6467	20.5737
	TMO - Mantiuk	0.4185	0.2926	0.4651	25.7570
BIQI	Lin	0.1817	0.1391	0.1466	38.7513
	PU	0.3387	0.2406	0.3445	30.5220
	TMO - Drago	0.2803	0.1923	0.2960	41.0579
	TMO - Reinhard 02	0.3756	0.2778	0.3766	33.2005
	TMO - Reinhard 05	0.3097	0.2213	0.2874	27.7294
	TMO - Mantiuk	0.2822	0.1957	0.2408	39.0999
DIIVINE	Lin	0.6677	0.4853	0.6759	21.8020
	PU	0.7156	0.5290	0.7193	18.7586
	TMO - Drago	0.7418	0.5562	0.7400	18.9959
	TMO - Reinhard 02	0.7149	0.5266	0.7024	20.7177
	TMO - Reinhard 05	0.7900	0.5932	0.7809	17.2134
	TMO - Mantiuk	0.4946	0.3549	0.4936	27.4918
kCNN	Lin	0.8363	0.6530	0.8134	19.1753
	PU	**0.8638**	**0.6852**	**0.8497**	**16.8937**
	TMO - Drago	0.7700	0.5853	0.7485	18.2759
	TMO - Mantiuk	0.8075	0.6188	0.8053	17.7948
	TMO - Reinhard 02	0.8613	0.6668	0.8179	17.7157
	TMO - Reinhard 05	0.6438	0.4631	0.6074	22.3484
Proposed	PU	**0.8860**	**0.7170**	**0.8871**	**16.4171**
Proposed	Lin	**0.8920**	**0.7184**	**0.8860**	**14.1464**

TABLE II: Overall Performance comparison. Highlighted in bold are the high performing metric.

after retraining) for many of the algorithms. Best performances are obtained by using BRISQUE [7] and kCNN [26]. The high performances of BRISQUE and kCNN can be attributed to the features they use, i.e., the MSCN coefficients. It is likely that the normalization by variance cancels the effects of the increased dynamic range and yields a similar distortion pattern as LDR images. Practically, kCNN is more useful because it produces a perceptual distortion map in addition to the quality score. The perceptual distortion map indicates what errors a human viewer can notice on the noisy image. Furthermore, we observe a clear performance improvement in LDR NR-IQA algorithms if the data is pre-processed and the dynamic range of the data is reduced to LDR levels. PU encoding improves the performance in most of the cases. The best performance among LDR NR-IQA is obtained while using PU encoding in conjunction with kCNN.

The performance of the proposed system is significantly better than the other algorithms in all cases both with or without preprocessing using PU encoding.

E. Cross-dataset testing

In order to demonstrate the generalization capabilities of the proposed NR-IQA technique to different conditions and contents, we train the algorithms using datasets #1, #2 and #3, and test them on datasets #4 and #5. This represents a real-world scenario, where the validating conditions are different from that of the training data. In addition, this testing method allows us to compare performance with FR-IQA algorithms. From a machine learning point of view, we have a sufficient number of examples of each type of distortion in datasets #1, #2 and #3. There is also a combination of all distortion types in dataset #4 and #5. The test set contains DMOS scores uniformly distributed in the range [20, 90].

Since the CNN is initialized with a set of randomly generated weights, the results of training can vary. We report the median score after 10 train-test cycles. Our results are given in Table III. By itself, BRISQUE, BIQI, SSEQ and DIIVIINE seem to be unable to adapt to the different image sizes and luminance ranges in the testing set, when these features are different from the training set. This can be explained by the fact that the features used by these algorithms are computed over a global histogram from the entire image.

The kCNN method performs well and shows good adaptability to a different test image size. This can be attributed to the fact that an image block is used to train the kCNN and hence the overall image size becomes less important. The method, however, does not take into account perceptual factors and we can see an improvement if PU encoding is used in pre-processing.

Our proposed algorithm outperforms related methods in all test cases when considering generalization to a real-world scenario. The superior performance demonstrates the strength of our two-stage training design, which successfully adapts to different image and luminance range. Perceptual Resistance values are able to scale the errors based on the luminance and image content. Our result achieves performance close to FR-IQA methods without the need of a reference image. For the full reference methods, the metric values are scaled to the range of [0,100] as suggested in [5]. A scatter plot comparing the scores generated by the proposed method and the human judged DMOS is shown in Fig. 6. A high correlation is observed.

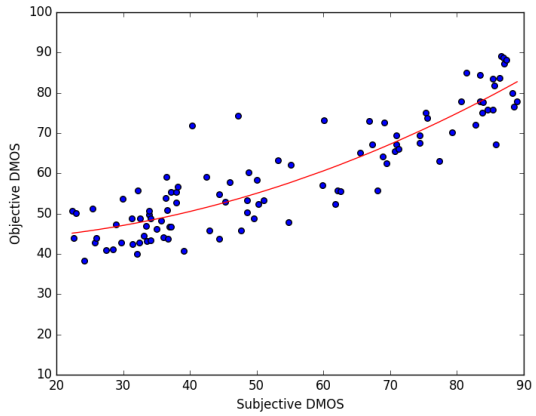

Fig. 6: Scatter plot between objective scores from the proposed method and subjective human judged DMOS.

VI. Analysis and Discussion

In this section, we analyze each sub-component of the proposed system and show the functions performed by them in detail. We highlight some interesting results provided by our system in addition to the quality score of an image. We also discuss the weakness and failure cases. In the following, we color-code the results with red points showing high values, green intermediate and blue low values.

Scheme	Processing	SRCC	KRCC	PLCC	RMSE
BRISQUE	Lin	0.5400	0.3732	0.4772	28.8475
	PU	0.7135	0.5121	0.6503	20.5534
	TMO - Drago	0.6337	0.4483	0.5903	21.7118
	TMO - Reinhard 02	0.6583	0.4670	0.6512	18.4500
	TMO - Reinhard 05	0.3524	0.2482	0.3946	30.6615
	TMO - Mantiuk	0.5887	0.4103	0.5493	22.7529
SSEQ	Lin	0.5287	0.3599	0.4714	25.2588
	PU	0.6492	0.4543	0.6111	19.6977
	TMO - Drago	0.5865	0.3956	0.5634	22.6987
	TMO - Reinhard 02	0.5810	0.4075	0.5644	22.9900
	TMO - Reinhard 05	0.4990	0.3401	0.5036	24.9193
	TMO - Mantiuk	0.4973	0.3398	0.4770	21.2044
BIQI	Lin	0.2845	0.1876	0.2831	31.0686
	PU	0.4386	0.3041	0.4399	21.2084
	TMO - Drago	0.5332	0.3780	0.4436	25.6200
	TMO - Reinhard 02	0.4632	0.3196	0.4358	22.0376
	TMO - Reinhard 05	0.5748	0.4048	0.5630	19.4825
	TMO - Mantiuk	0.4651	0.3204	0.4571	24.2268
DIIVINE	Lin	0.5041	0.3429	0.5209	20.6506
	PU	0.5318	0.3691	0.5442	19.6772
	TMO - Drago	0.4143	0.2852	0.4065	25.9697
	TMO - Reinhard 02	0.3634	0.2434	0.3953	26.1464
	TMO - Reinhard 05	0.5558	0.3849	0.5374	19.3122
	TMO - Mantiuk	0.4138	0.2838	0.4496	21.0499
kCNN	Lin	0.6991	0.5156	0.7008	19.3677
kCNN	PU	0.7694	0.5845	0.7544	18.5854
Proposed	Lin	**0.8672**	**0.6773**	**0.8780**	**18.626**
HDR-VDP-2.2	Full Reference	**0.9298**	**0.7691**	**0.8710**	**10.120**
HDR-VQM	Full Reference	**0.9193**	**0.7444**	**0.8940**	**10.725**
PU-MSSIM	Full Reference	**0.8969**	**0.7125**	**0.7589**	**12.775**
PU-SSIM	Full Reference	**0.9121**	**0.7339**	**0.7121**	**11.688**

TABLE III: Cross-dataset results for different IQA methods.

A. *Error Estimation*

Per block error estimation is performed by E-net (Figure 1). The output $\hat{\delta}$ of E-net on a few example images from Datasets #4 and #5 is reported in Fig. 7, which shows the distorted images, the actual error δ in the image (equation 2) and the error estimations generated by E-net. As a reference, we also include the corresponding Root Mean Square Error (RMSE) between δ and $\hat{\delta}$ in Fig 7 (e).

A more detailed look at the error estimations can be seen in Figure 8, where we show enlarged regions of the image containing actual error and the corresponding error estimations. It is apparent from the error maps that E-net is able to successfully identify major errors in the images.

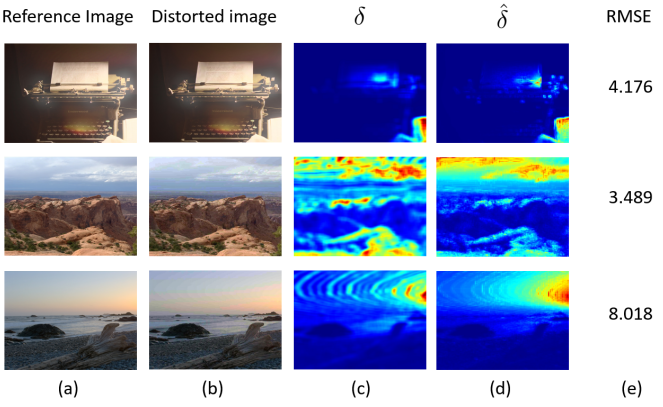

Fig. 7: *Error estimation.* (a) reference images; (b) distorted images; (c) ground truth error map provided by the database; (d) error maps estimated by E-net; (e) RMSE between the estimated and ground truth errors on the HDR luminance images.

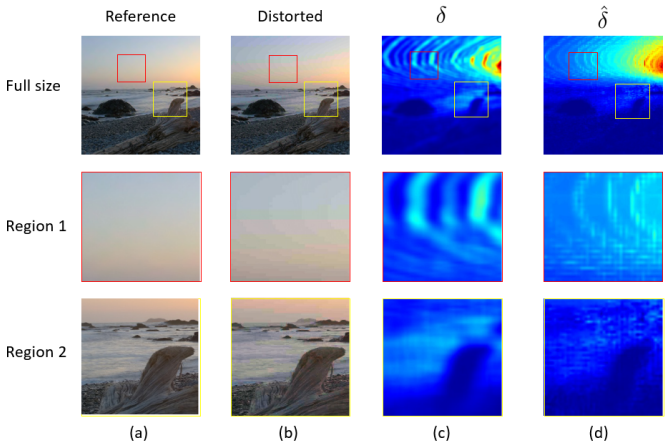

Fig. 8: *Error estimation.* Detailed comparison of actual and estimated error. Columns (a) to (d) have the same meaning as Figure 7.

B. *Perceptual Resistance*

The Perceptual Resistance value T produced by P-Net shows how unlikely it is that an error in a region can reduce local visual quality. We use the image in Fig. 9 to illustrate how P-net handles contrast sensitivity and masking effects. We consider a JPEG compressed image from the CSIQ dataset [41]. The image is scaled to a maximum luminance of 100 cd/m^2. This is done for ease of explanation as this is the luminance range commonly seen in everyday computer and television screens. The image is input to the proposed algorithm as blocks of size 32×32. The output from E-net, P-net and the Mixing function are shown in Fig. 9.

a) Analysis of spatial masking: Spatial masking effects can be observed in the output of P-net (T in Fig. 9). In the color map, the Perceptual Resistance T is lower for pixels in the sky, which is brighter and smoother, compared to other image blocks. Thus, errors in the sky region are easy to notice. On the contrary, in regions with high spatial frequency, e.g., bushes and architecture, T is higher indicating that errors are less noticeable. The error $\hat{\delta}$ estimated by E-net and the actual error δ is shown in Fig. 9. The perceptual distortion map obtained after the mixing function reflects high values in the sky and low values in the more densely textured regions.

b) Analysis of sensitivity to luminance: As explained in Section III.A, [24] performed a user study by displaying a distorted image on two screens with maximum luminance of 100 cd/m^2 and 1,000 cd/m^2, respectively. The study revealed that users rated the perceived quality on the high luminance monitor worse compared to the the same image displayed on the lower luminance monitor. We reproduced this finding by applying our algorithm on the LDR images in the CSIQ database [41]. We used all the JPEG distorted images in the database and linearly scaled the images to luminance ranges of 100 cd/m^2 and 1,000 cd/m^2, respectively. We then computed the quality of these images using the proposed method and compared the scores of the two luminance ranges. We found that the mean value of DMOS for the images of low brightness was 68.76 and those of high brightness was

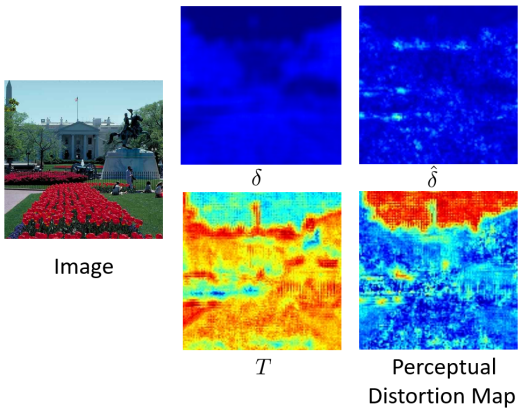

Fig. 9: *Perceptual resistance*. Example output of E-net and P-net. For the content of 'Image', Actual error δ, Output of E-net $\hat{\delta}$ - Estimated error, Output of P-net T - Perceptual Resistance and perceptual distortion map $DMOS(i,j)$ from the Mixing function is shown.

76.72. This implies that the brighter images appear more distorted than their corresponding low brightness counterparts. We then employed ANalysis Of VAriance (ANOVA) to check for statistical significance of the difference between DMOS for the low and high brightness images. The p value was found to be < 0.05 and $F < F_{crit}$ for the hypothesis that the difference in means was zero; proving that the difference is statistically significant. Hence, we can reproduce the results obtained in [24] using our data driven method without the need of conducting low level visibility user studies.

We performed further analysis on the image discussed in Fig. 9. Fig. 10 shows the output of our proposed method at different noise levels under 2 different luminance ranges. Notice that:

1) In both luminance ranges, the estimated $\hat{\delta}$ is the same, because the measure is based on content statistics and does not take into account the change in luminance.
2) The estimated T is instead affected by the luminance range, as discussed in Section III. For a given distortion level, the perceptual resistance is lower for higher luminance values, consistently with the the findings in [24]. However, the mapping of luminance to perceptual resistance is not a simple global linear scaling, but takes into account the complex content-dependent characteristics that determine contrast and luminance masking.

c) Comparison between perceptual resistance and contrast sensitivity: HDR-VDP 2.2 can provide a contrast threshold, i.e., the per pixel contrast (difference) such that an error is visible with a certain probability. This map is related to our Perceptual Resistance values. However, the contrast threshold is the result of psychophysical experiments to determine contrast sensitivity at different luminance levels, while Perceptual Resistance is indirectly obtained through quality scores.

A visual comparison, Fig. 13, can give and idea of the relative values of the two measures. The color maps shown are generated using the same scale and value range, i.e., log scale normalized to the range [0,1]. The comparison shows similar perceptual results of the two algorithms. The color maps show that the location and relative intensity of the visual errors are similar. While the contrast threshold map is perceptually more accurate, it requires the knowledge of the pristine image. The proposed two-stage network architecture enables to approximate it without any reference.

	Proposed	**Linear**	**tanh()**	**Logistic**
SRCC	0.8672	0.8616	0.8560	0.8476
KRCC	0.6773	0.6630	0.6719	0.6474
PLCC	0.8780	0.8597	0.8688	0.8535
RMSE	18.6268	18.8270	16.2990	26.7700

TABLE IV: Performance with various mixing functions

C. Comparison using alternative Mixing functions

As explained in Section 4.D, an alternative Mixing function can be used as long as it is monotonically increasing and the network can converge.

We tested the following cases to study the effects of different mixing function formulations:

1) Proposed: $DMOS(i,j) = 1 - \exp(-\left|\frac{k*\hat{\delta}(i,j)}{T(i,j)}\right|)$.
2) Linear mixing: $DMOS = \frac{\hat{\delta}}{T}$.
3) Hyperbolic tangent: $DMOS = \tanh(\frac{\hat{\delta}}{T})$.
4) Logistic function:

$$DMOS(i,j) = \frac{1}{1+\exp(-k(x-x_0)}$$

where $x = \frac{\hat{\delta}}{T}$.

We found that during the train-test cycles in cases 2 and 4 above, the network failed to converge and the training error kept increasing. This happens when the random weight initialization causes these functions go out of bound, interfering the gradient propagation. Cases 1 and 3 do not have convergence problem. The results are shown in Table IV. We observed a similar performance whenever the network converged, implying the possibility of having multiple choices for G.

To investigate further, we selected a distorted image and generate the results using different mixing functions and P-net as shown in Fig. 11. As explained earlier, a mixing function outputs the DMOS and P-net outputs the Perceptual Resistance of an image block. A comparison of HDR-VDP 2.2 error probability and contrast threshold is shown for reference.

The perceptual distortion maps do not change significantly as the mixing function changes. This is because the CNN optimization process tries to minimize the difference between the mixing function output and the actual human judged DMOS. Upon convergence, the results will be similar.

Considering P-net, we see that the results show a similar trend in terms of relative values. For example, visibility values in the sky are generally higher compared to the texture-rich forest area. However, the level of change in Perceptual Resistance varies depending on the type of mixing function used. The proposed Mixing function defined in Eq. 4 is used

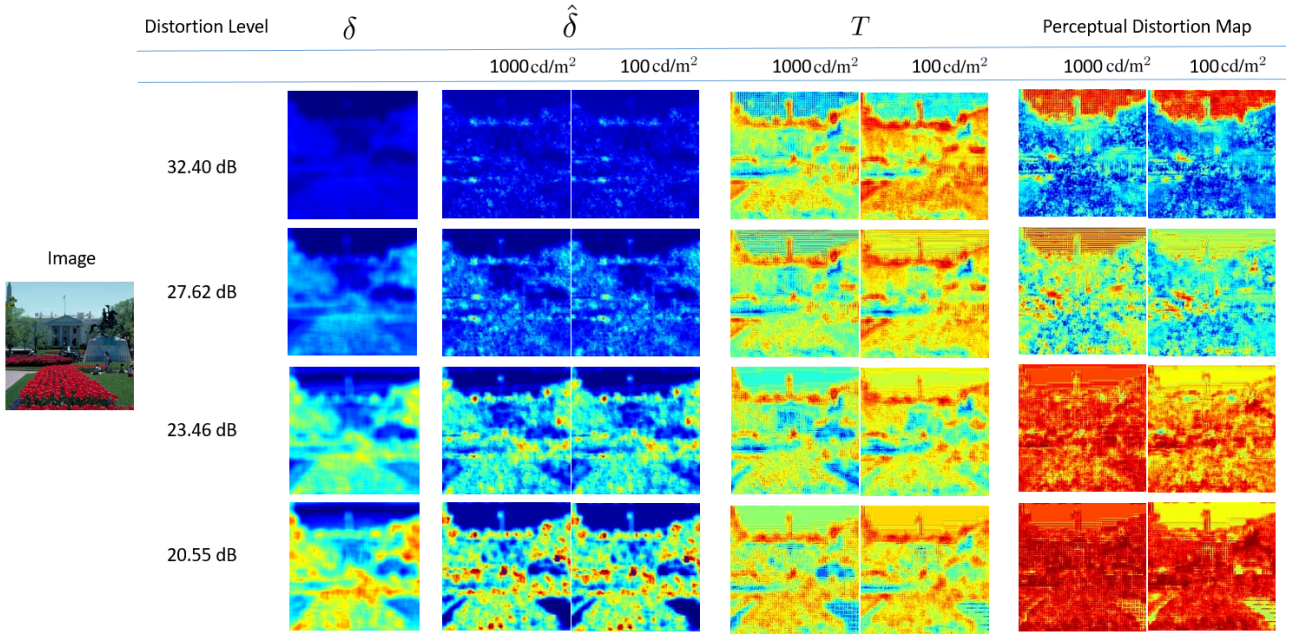

Fig. 10: *Perceptual resistance*: Behavior of proposed system at different luminance levels. Distortion Level of JPEG is shown in dB, δ is the Acutal error, $\hat{\delta}$ is the Estimated error by E-net, T is the Perceptual Resistance by P-net. Within each column, the left and right images are output when the input image is linearly scaled to 1,000 and 100 cd/m^2 respectively.

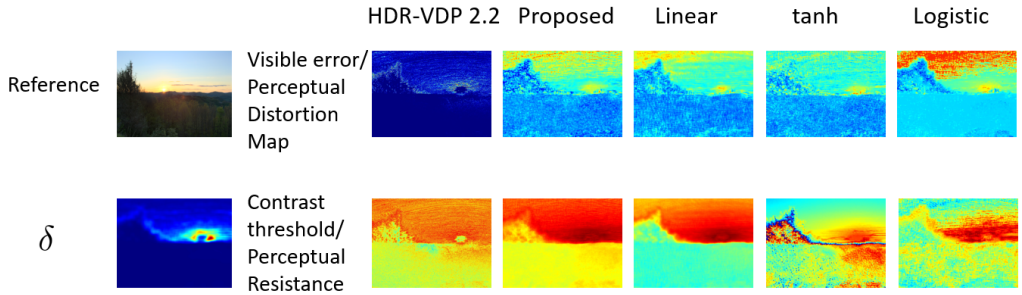

Fig. 11: *Effect of different mixing functions*. Comparison of perceptual distortion maps and perceptual resistance generated from different mixing functions.HDR-VDP error probability and contrast threshold are shown for reference.

in our system because it generates the closest results to that of the FR-IQA metric HDR-VDP2.2.

Finally, the Mixing function could automatically be learned from data using another CNN. However, this process would involve more weights and difficulties in optimization, with no guarantee that the overall model converge to a good solution. We confirmed this experimentally by using a Deep Belief Network in place of the Mixing function. Even if the system did converge with a DBN, the gain in performance with respect to the proposed mixing function is not substantial, and this function would be rather a "black box" with no intuitive interpretation.

D. Perceptual distortion map

One of the advantages of the proposed CNN based NR-IQA scheme is that it gives the approximate location of the perceived errors in a *perceptual distortion map*. A comparison of the perceptual distortion maps produced by FR-IQA (HDR-VDP 2.2, PU-SSIM) and NR-IQA (PU-kCNN) algorithms and the proposed method for some example images is shown in Fig. 12. In order to highlight the errors, the probability of error detection is shown for HDR-VDP 2.2, and an inverted PU-SSIM map is shown for PU-SSIM. Notice that the map obtained using HDR-VDP 2.2 is conceptually different from those of PU-SSIM and PU-kCNN: the former represents per pixel *probability* of detecting distortion, while the latter two convey information about the *magnitude* of local distortion. Instead, our map summarizes the two types of information into a local perceptual distortion, which can be seen as the inverse of local perceptual quality.

Although they express different properties of distortion, the perceptual distortion maps produced by the proposed scheme, PU-kCNN and HDR-VDP2.2 are consistent in general. To show the performance improvement with respect to the current state-of-the-art PU-kCNN to produce a perceptual distortion map, we compute the MSE between perceptual distortion maps of proposed and PU-kCNN with respect to HDR-VDP2.2, reported in Fig. 12. From both MSE and visual inspection,

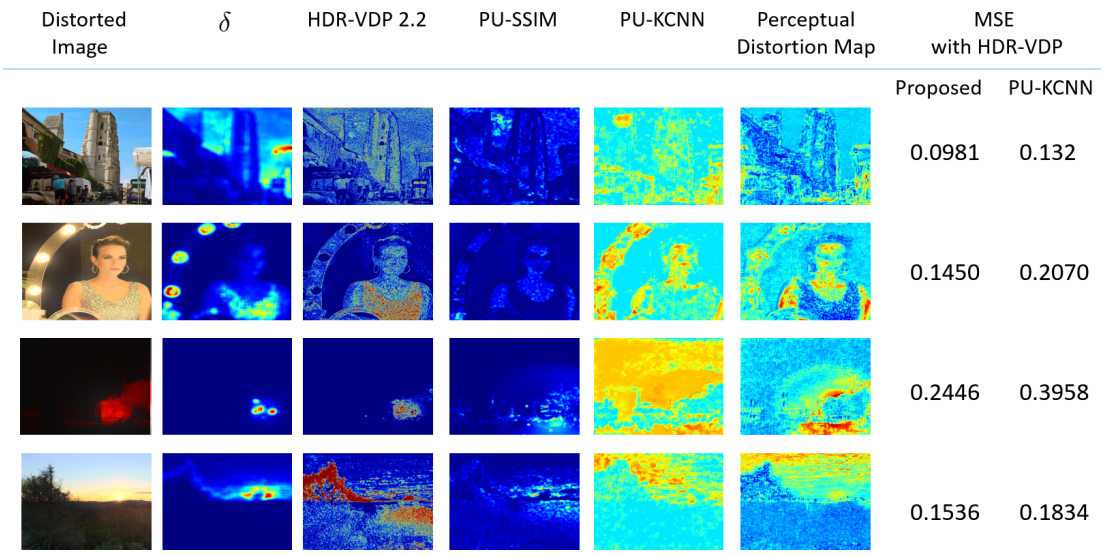

Fig. 12: *Perceptual distortion maps*. Comparison of perceptual distortion maps estimated by various IQA schemes. MSE shown is computed between the distortion maps of proposed method and PU-kCNN with HDRVDP2.2.

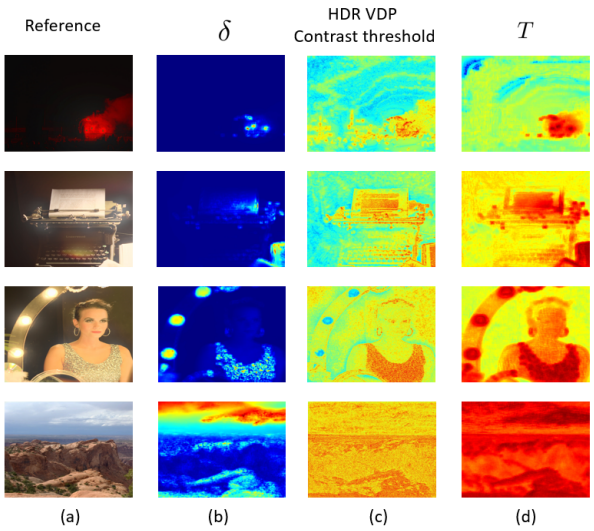

Fig. 13: *Perceptual distortion map*. Comparison between P-net Perceptual Resistance and HDR-VDP2.2 contrast threshold. Image pixel values are in log scale and normalized to [0,1].

it is clear that our perceptual distortion map is closer to HDR-VDP visibility map, compared to PU-KCNN, i.e., it discounts errors that are less likely to be perceived due to masking and reduced sensitivity to contrast. This explains the overall better performance of the proposed approach over kCNN in Table III.

E. Effect of preprocessing

In HDR FR-IQA it is common to preprocess images through a perceptual transfer function [24], [5], [42] in order to compute perceptually meaningful differences between pixels. However, in our approach we decouple the error computation from its perceptual scaling, and this preprocessing turns out to be unnecessary, as it is implicitly carried out by the P-net.

To support this claim, we test the performance of our system by preprocessing images with the PU encoding [24] before training the network. The performance with this setting is reported at the bottom of Table II. Although PU encoding improves the performance of all other NR-IQA (since it provides perceptual scaling), it does not improve the performance of the proposed system – in fact, correlations are slightly lower if pixels are PU encoded. This could be attributed to the loss of information due to perceptual encoding, where information is somehow quantized and some slight variations in the data are compressed with a fixed scaling function.

F. Failure cases

We isolated some cases where the predicted DMOS quality has a large error compared to the groundtruth. In some of those cases, the perceptual distortion maps produced by our method are poorly correlated with the visibility maps produced by HDR-VDP or with the local PU-SSIM estimated distortion. Some examples are reported in Fig. 14. We observe that the perceptual distortion maps produced by our method are not consistent with the ones produced by HDR-VDP and PU-SSIM (e.g., in the sky for the first image, or in the lake for the second or the dress of the woman in Fig. 12): in one case, the FR algorithms estimate minimal distortion, while our method predicts higher errors, while in the other two cases, our method underestimates perceptual distortion. MSE between the estimated distortion map and HDR-VDP error visibility, as well as groundtruth DMOS values, are given for reference.

To analyze this further, we note that the discrepancy appears mainly in large smooth areas, e.g., sky and the water of the lake, where HDR-VDP shows high error detection, and in cluttered ground area where HDR-VDP shows low error detection in smooth region. PU-SSIM estimates a similar spatial distribution of the error into both regions. Another instance of this is in Fig. 12, second row, where the highly

textured dress of the woman is shown to have high errors in HDR-VDP and PU-SSIM, but the proposed method suggests a low error.

We can see from intermediate results of the proposed method inside blue box of Fig. 14 that the errors are caused by different components of the system. Over-estimation of errors in the smooth sky is the effect of over-estimation of errors by E-net in smooth areas, whereas under-estimation of errors in highly textured regions is the result of perceptual resistance being too high. These errors are likely due to some bias in our dataset, where a large fraction of examples contain smooth or textured regions have low and high quality respectively.

We notice, however, that those results could indeed be meaningful from a perception point of view, and that the visibility map produced by HDR-VDP alone is not a good indicator of quality [34]. In order to further assess the prediction of local perceptual error and its impact on overall quality, we would need per block groundtruth of quality scores, which not only is unavailable nowadays, but is very difficult to produce as assessing quality is by definition a *global* task. Furthermore, the proposed method estimates quality per block *independently*, which is a major simplification in the model. Extensions to how to embed higher order dependencies between regions of the image, and possibly semantic considerations (which become realistic using deep CNN architectures), are left to future work.

G. Computational complexity

Assuming a fully trained system, the computational complexity of our method is high because of the large amount of convolutions involved. The asymptoptic time complexity is mainly due to E-net because of its convolutional layers. Each convolutional layer uses the results of the previous layer. This makes our method slower compared to kCNN where there is only one layer of convolutional filters. We found that in our GPU implementation, computation of score for an image with 1920×1080 resolution takes 0.33 seconds (assuming the theano graph is in memory). For KCNN, a similar coding style gave us execution time of 0.15 seconds per image. Note that the execution speed can be improved with better implementation. This is left as a future work; our focus in this work is to improve NR-IQA performance.

VII. Conclusion

We propose a HDR NR-IQA scheme that uses a CNN based architecture, composed of E-net, P-net and Mixing function, to generate values corresponding to Error Estimation and Perceptual Resistance in an image, which are then combined to generate a perceptual distortion map and DMOS. Taking into account perceptual phenomena is directly derived from real-world data driven optimization, and does not involve psychophysical user studies. The derived Perceptual Resistance shows similar characteristics compared with other perceptual models. Experimental results demonstrate that our algorithm accurately predicts visual distortions such as compression artifacts. Our algorithm scores correlate well with human scores. It outperforms other NR-IQA methods and the performance is competitive compared to HDR FR-IQA methods, with the advantage that no reference image is needed.

VIII. Acknowledgments

We thank Dr Anup Basu for his insightful comments and valuable feedback and Emin Zerman for his help with the HDR datasets.

References

[1] F. Dufaux, P. L. Callet, R. Mantiuk, and M. Mrak, *High dynamic range video : from acquisition to display and applications*.

[2] A. Chalmers and K. Debattista, "{HDR} video past, present and future: A perspective," *Signal Processing: Image Communication*, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092359651730019X

[3] M. Narwaria, R. K. Mantiuk, M. P. Da Silva, and P. Le Callet, "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," vol. 24, no. 1. International Society for Optics and Photonics, 2015, pp. 010 501–010 501.

[4] M. Narwaria, M. Perreira Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46–60, 2015.

[5] E. Zerman, G. Valenzise, and F. Dufaux, "An extensive performance evaluation of full-reference HDR image quality metrics." in *Springer: Quality and User Experience*, vol. 2, no. 1, 2017.

[6] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *Signal Processing Letters, IEEE*, vol. 17, no. 5, pp. 513–516, 2010.

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

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

[9] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.

[10] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1733–1740.

[11] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 987–994.

[12] T. Zhu and L. Karam, "A no-reference objective image quality metric based on perceptually weighted local noise," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–8, 2014.

[13] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, 2013.

[14] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. Ieee, 2003, pp. 1398–1402.

[15] M. Reddy, "Perceptually optimized 3d graphics," *IEEE computer Graphics and Applications*, vol. 21, no. 5, pp. 68–75, 2001.

[16] D. Luebke and B. Hallen, "Perceptually driven simplification for interactive rendering," in *Rendering Techniques 2001*. Springer, 2001, pp. 223–234.

[17] D. Kundu and B. L. Evans, "Visual attention guided quality assessment of tone-mapped images using scene statistics," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 96–100.

[18] H. Z. Nafchi, A. Shahkolaei, R. F. Moghaddam, and M. Cheriet, "FSITM: A feature similarity index for tone-mapped images," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1026–1029, 2015.

[19] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference image quality assessment for high dynamic range images," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, 2016.

[20] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, "Do HDR displays support LDR content?: a psychophysical evaluation," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 38, 2007.

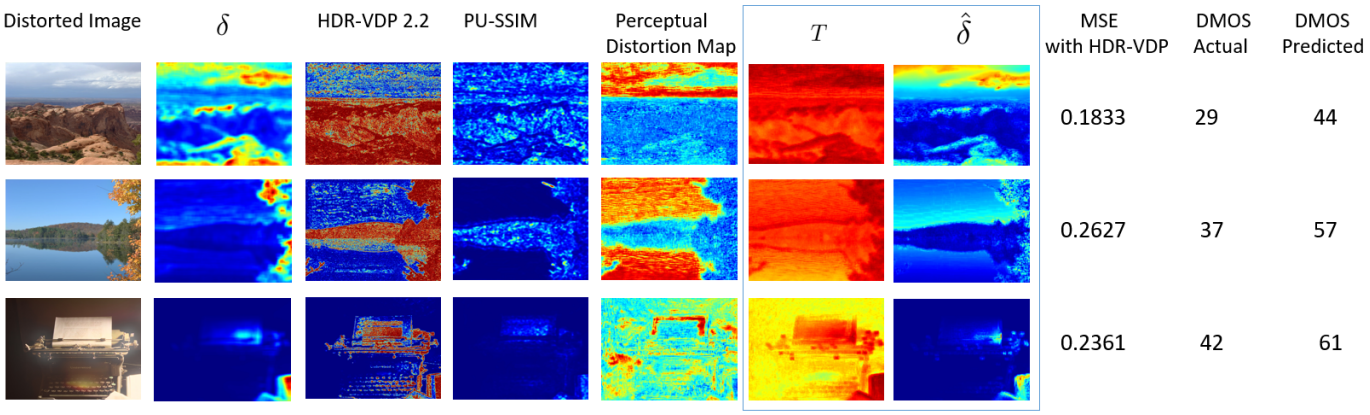

Fig. 14: *Failure cases*. There are cases where the perceptual distortion maps from different perceptual IQA show inconsistency, but there is no existing benchmark metric to evaluate distortion maps. Images inside the blue box are intermediate results.

[21] R. Shen, I. Cheng, J. Shi, and A. Basu, "Generalized random walks for fusion of multi-exposure images," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3634–3646, 2011.

[22] G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "Lightness perception in tone reproduction for high dynamic range images," in *Computer Graphics Forum*, vol. 24, no. 3. Wiley Online Library, 2005, pp. 635–645.

[23] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3, pp. 267–276, 2002.

[24] T. O. Aydın, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," in *Electronic Imaging 2008*. International Society for Optics and Photonics, 2008, pp. 68 060B–68 060B.

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

[26] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2791–2795.

[27] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1992, pp. 2–15.

[28] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer vision–ECCV 2014*. Springer, 2014, pp. 818–833.

[29] W. Heng and T. Jiang, "From Image Quality to Patch Quality: An Image-Patch Model for No-Reference Image Quality," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 1238–1242.

[30] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[31] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pepion, "Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality," *Optical Engineering*, vol. 52, no. 10, pp. 102 008–102 008, 2013.

[32] M. Narwaria, M. P. Da Silva, P. Le Callet, and R. Pépion, "Impact of tone mapping in high dynamic range image compression," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2014, pp. pp–1.

[33] P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of HDR images compressed with JPEG XT," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[34] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for HDR image compression," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2014, pp. 92 170C–92 170C.

[35] M. H. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets." in *Visual Communications and Image Processing*, 2003, pp. 583–592.

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

[37] E. Reinhard and K. Devlin, "Dynamic range reduction inspired by photoreceptor physiology," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11, no. 1, pp. 13–24, Jan 2005.

[38] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," in *Computer Graphics Forum*, vol. 22, no. 3. Wiley Online Library, 2003, pp. 419–426.

[39] R. Mantiuk, S. Daly, and L. Kerofsky, "Display adaptive tone mapping," in *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3. ACM, 2008, p. 68.

[40] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 26, no. 6, pp. 1275–1286, 2015.

[41] E. C. Larson and D. Chandler, "Categorical image quality (CSIQ) database," *Online, http://vision.okstate.edu/csiq*, 2010.

[42] S. Miller, M. Nezamabadi, and S. Daly, "Perceptual signal coding for more efficient usage of bit codes," *SMPTE Motion Imaging Journal*, vol. 122, no. 4, pp. 52–59, 2013.