APSIPA Transactions on Signal and Information Processing, 2024, 39, e25

Original Paper

Deep-Learning for Objective Quality Assessment of Tone Mapped Images

Ishtiaq Rasool Khan[1,2] and Romil Imtiaz[3*]

[1] *Abu Dhabi School of Management, United Arab Emirates*

[2] *College of Computer Science and Engineering, University of Jeddah, Saudi Arabia*

[3] *Pakistan Institute of Engineering and Technology, Multan, Pakistan*

ABSTRACT

High dynamic range (HDR) images capture real-world luminance values which cannot be directly displayed on the screen and require tone mapping to be shown on low dynamic range (LDR) hardware. During this transformation, tone mapping algorithms are expected to preserve the naturalness and structural details of the image. In this regard, the performance of a tone mapping algorithm can be evaluated through a subjective study where participants rank or score tone mapped images based on their preferences. However, such subjective evaluations can be time-consuming and cannot be repeated for every tone mapped image. To address this issue, numerous quantitative metrics have been proposed for objective evaluation. This paper presents a robust objective metric based on deep learning to quantify image quality. We assess the performance of our proposed metric by comparing it to 20 existing state-of-the-art metrics using two subjective datasets, including one benchmark dataset and a novel proposed dataset of 666 tone mapped images comprising a variety of scenes and labeled by 20 users. Our approach exhibits the highest correlation with subjective scores in both evaluations, confirming its effectiveness and potential to be a reliable alternative to laborious subjective studies.

*Corresponding author: Romil Imtiaz, romilimtiaz302@gmail.com.

Received 23 January 2024; revised 05 June 2024; accepted 22 July 2024

ISSN 2048-7703; DOI: 10.1561/116.20240007

Keywords: Image quality assessment, deep learning, image datasets, tone mapping, high dynamic range.

1 Introduction

High dynamic range (HDR) images capture real-world luminance values of the scene and have found numerous applications in immersive visualization, medical imaging, and computer graphics. With the increasing popularity of HDR content, tone mapping has become crucial in transforming HDR content to Low Dynamic Range (LDR) for display on existing screens. The tone mapping process involves compressing the dynamic range of content to match the target display, considering various factors such as the type of content, characteristics of the display, viewing conditions, and the perception mechanism of the average human viewer. Although several Tone Mapping Operators (TMOs) produce visually pleasing results, their performance is generally content-dependent, and no single TMO is optimal for all types of scenes.

To subjectively evaluate the performance of different TMOs, study participants are requested to rank/score several tone mapped images produced by different algorithms, and mean opinion scores are calculated for each TMO. However, this process is tedious and time-consuming, making it infeasible to be repeated for every new image and algorithm. Therefore, several algorithms have been produced which evaluate different image features to estimate the quality score. However, Traditional Image Quality Assessment (IQA) metrics such as the Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR), and Structural Similarity (SSIM) Index [45] cannot compare the tone mapped images against the reference HDR images due to vastly different pixel values of the two.

Several dedicated tone mapped image quality assessment (TM-IQA) metrics have been proposed, such as tone mapping quality index (TMQI) [46] and features fusion for natural tone mapped images quality evaluation (FFTMI) [25]. TM-IQA metrics are crucial in evaluating the performance of various tone mapping operators. Moreover, accurate TM-IQA metrics can produce better-quality tone mapped images, which are appealing visually and in applications like animation, computer graphics, and video games. These metrics have gained particular attention from researchers in recent years as they can facilitate building large training datasets for deep learning models. Such models have gained huge popularity and are being used for HDR tone mapping, inverse tone mapping, quality enhancement, quality evaluation, scene analysis, and several other applications.

The strength of a metric lies in how the features work together, not to let any distortion go undetected and unpenalized while scoring. The authors of [25] analyzed 60 features used by different FR metrics for their effectiveness

and shortlisted the best five to make a new FR metric called the FFTMI. The weights of these features were optimized for a widely used benchmark dataset, tone mapped image database (TMID) [46], and they indeed obtained better correlation with the subjective scores than all existing metrics. However, when tested on some images not included in TMID, the performance of FFTMI dropped significantly. This indicates that assessment of the quality of tone mapped images is a complex task, and even the best features picked from existing metrics cannot detect all distortions that can occur in tone mapped images. Therefore, there is a need for more effective and robust metrics that can function effectively across different types of images and scenes.

TM-IQA metrics can be classified as full reference (FR) methods that need the original HDR image to assess the quality of the tone mapped image and no reference (NR) or blind methods that do not need the reference image. Blind metrics measure image quality solely based on image characteristics without prior knowledge of the expected appearance. Reference images are not always available, and it is also difficult to compare them with their tone mapped versions due to a vastly different range of pixel values in the two. Therefore, blind metrics such as blind tone mapping quality index (BTMQI) [16] and global statistical features (GSF) [43] present a promising solution in TM-IQA.

This work proposes a novel blind image quality metric using a deep learning approach, aiming to conduct evaluations closely resembling the subjective evaluation of LDR images. The main contributions of this work are the following:

- We propose a deep neural network trained for evaluating the quality of tone mapped HDR images. The network works as a blind metric, taking only the tone mapped image as input and assigning a score to its quality. The assigned scores are well-correlated with the scores assigned by human subjects. This makes the proposed metrics an effective alternative to time-consuming subjective studies.

- We present a new custom dataset made up of 666 LDR images that were obtained by using various TMOs on various HDR scenes. Mean opinion scores obtained through a comprehensive subjective study involving 20 subjects are provided for each image. Reference images are also provided, which makes the dataset practical to advance research in designing and evaluating tone mapped image quality assessment metrics, both full-reference and blind.

It is noteworthy that the proposed method is blind and operates independently of the original image, making it applicable to general IQA rather than being specific to tone-mapped images. However, the datasets utilized for training consist of images generated using tone-mapping algorithms. This choice aims to ensure that any distortions unique to tone-mapping are detected by the

trained metric model. Therefore, we classify the proposed method as belonging to the category of TM-IQA metrics.

The paper is organized as follows. After this introductory section covering the problem statement and our contributions, we review the most relevant existing literature in Section 2. Sections 3 and 4 present the proposed dataset and the proposed deep network for tone mapped image quality assessment. Performance evaluation of the proposed network as a quality assessment metric is presented in Section 5. Concluding remarks are given in Section 6.

2 Literature Review

IQA can be broadly classified into subjective IQA and objective IQA based on the nature of the evaluation. Subjective IQA can be more accurate if conducted following standard protocols and guidelines, but it is time-consuming. Therefore, there has been an enormous interest in developing objective alternatives that can assign reliable scores to describe the quality of an image. In this section, we briefly describe some TM-IQA metrics. A vast range of methods has been developed for general IQA, not particular to TM-IQA. Those methods are not covered in this review.

Aydin *et al.* [3] proposed a method known as dynamic range independent image quality assessment (DRIM), which uses a model of the human visual system (HVS) to detect distortions in tone mapped images. The authors considered three types of distortions: loss of visible contrast due to detail compression, amplification of invisible contrast due to contouring and other artifacts, and reversal of contrast caused by strong distortions such as clipping or salient compression artifacts. Although DRIM generates visual distortion maps, it does not assign a quantitative score.

The TMQI [46] is a widely used method that includes two metrics – structure fidelity and statistical naturalness. Structural fidelity compares the structure in reference and test image patches, and if signal strengths of both HDR and LDR patches are above/below the visibility threshold, the structure is assumed to be intact. For naturalness, TMQI uses a blind approach utilizing the brightness and contrast attributes of the tone mapped image alone. Ma *et al.* [30] claimed that the structural fidelity measure of TMQI was overly sensitive to noise. To resolve the problem, they calculated the visibility threshold for each image patch separately instead of using a global value. For naturalness, they used the HDR image to determine desirable parameter values, thus changing it from blind to a full-reference (FR) measure. Their algorithm is generally referred to as TMQI2.

The feature similarity (FSIM) [48] index is an FR IQA method using contrast-invariant phase congruency and contrast-dependent gradient magnitude features. This method was extended by Nafchi *et al.* [35] for TM-IQA and was named the feature similarity index for tone mapped images (FSITM).

FSITM determines how well the local angle maps are preserved in each R, G, and B channel. However, FSITM does not consider the scene brightness, and therefore the authors recommended using FSITM with TMQI and not as a standalone metric.

Hadizadeh and Bajić [17] proposed a metric called tone mapped Image Quality (TIQ) that forms a "bag of features" extracted from both test and reference images representing structural fidelity, naturalness, and brightness. These features are used to train a support vector regression model to predict the quality of tone mapped images. Song *et al.* [41] used the local exposure function to divide the HDR picture into various pieces. A regression model was trained to forecast quality considering abnormal exposure ratio, leftover exposure energy, and a color-based feature.

Several blind IQA methods can be used for TM-IQA without using the reference HDR image. However, tone mapping induces specific distortions such as details and color loss [7], which may not be well-understood by general IQA methods, thus leaving room for the development of blind metrics specifically for tone mapped images. Gu *et al.* [16] proposed a blind version of TMQI, called BTMQI, in which the naturalness measure of TMQI remains unchanged while the structure is measured as the sum of pixels in the binarized gradient of tone mapped image. An additional feature, information content, is measured using local and global Shannon entropy in 9 nine intensity scaled versions of the tone mapped image. These 11 features are combined into a single score using a network trained on a subjective dataset.

Fang *et al.* [14] noted that BTMQI [16] does not consider microstructural level distortions, and therefore they used relative gradient magnitudes instead of absolute values in the measurement of structure. In addition, they used chromatic descriptors to penalize fading of color. The metric called visual quality evaluation using gradient and chromatic statistics (VQGC) trains a support vector regression model for scoring using these features.

Jiang *et al.* [24] used the same three attributes of information content, naturalness, and structure as in BTMQI [16] to define several features and used the extreme learning machine to obtain a quantitative score. Kundu *et al.* [27] proposed the HDR image gradient-based elevator (HIGRADE) metric based on a statistical model using log-derivative features and scene statistics in spatial and gradient domains. Yue *et al.* [47] extracted 38 features from the Tone mapped Image to evaluate image quality. These features include colorfulness, measured as saturation in opponent yellow-blue and red-green channels; exposure, measured using entropies of several intensity-scaled image variants; and structural variation, measured using gradient images. For naturalness measurement, illumination extracted using the Retinex theory was used. The average and range of lightness were used for contrast, while the halo effect was estimated using the gradient of the illumination channel. A support vector machine was trained using these features for scoring.

Chen *et al.* [9] segmented images and extracted global contrast and local entropy features from dark and bright regions and colorfulness from the normal region. All features were extracted at multiple image resolutions and mapped to an objective quality score by a random forest regression algorithm. Jiang *et al.* [21] extracted features similar to Chen *et al.* [9] from portioned luminance maps and additional features of microstructural distortions and halo measurements. These features were used to train a support vector regression model for scoring.

Wang *et al.* [44] used texture, structure, colorfulness, and naturalness attributes to extract several local and global features. These features were combined using regression into a single score. He *et al.* [18] proposed a multi-scale multi-layer convolutional neural network for image quality evaluation. Images were represented at different scales, and several local and global features were extracted, which were aggregated over several layers to predict a score for image quality. He *et al.* [19] considered the effect of color and details on the Human Visual System (HVS) and derived several local and global features to use in their Regional Sparse Response and Aesthetics (RSRA) metric. These features were combined using the random forest algorithm.

Cui *et al.* [10] used low and high-level perception characteristics to extract several features, some of them using a deep learning model, from tone mapped images. They used regression to obtain an overall score for image quality. These studies have made significant contributions to the field of objective image quality evaluation by proposing different features and models for scoring.

Alotaibi *et al.* [2] presented a metric that measures the loss of color, contrast, brightness, and structure using 16 features extracted from the test tone mapped image and the reference HDR image. The effect of these attributes on image quality is combined into a single score in the $[0, 1]$ range describing the quality of tone mapped image.

Jiang *et al.* [23] address the challenge of evaluating underwater image enhancement techniques by introducing a comprehensive benchmark dataset along with a tailored objective metric for quality assessment. Furthermore, Jiang *et al.* [22] present a real-world dataset for single image super-resolution, complemented by thorough subjective studies and the development of a dedicated objective quality metric. On other hand, Chen *et al.* [8] investigate the quality evaluation of style transfer algorithms, providing insights through subjective studies and proposing an objective metric to quantify the aesthetic success of arbitrarily stylized images.

3 Proposed Dataset

Datasets are crucial for adequate training of machine learning models. However, generating large datasets is a daunting task that takes excessive time and effort. Smaller datasets can lead to overfitting, which means that the model

becomes too specialized to the features of the small training data and fails to generalize well to new data. TMID [46] is quite popular in the existing literature because it provides reference and tone mapped images, making it useful for the training and testing of FR metrics. However, the dataset uses only 15 HDR images and 8 TMOS; therefore, it contains only 120 tone mapped images. The metrics trained on this small dataset are prone to overfitting [2].

Another dataset, ESPL LIVE [26], contains 1811 tone mapped images but does not provide reference images. The subjective scores for this large number of tone mapped images were collected through crowd-sourcing. This subjective evaluation method collects opinions and ratings from individuals on a particular subject or dataset [11]. Without reference images, ESPL LIVE cannot be used to train and test FR metrics. However, since we propose a blind metric in this work, ESPL LIVE is a good resource for us to train our network.

For additional diversity of scenes, which is desirable in training, we take another set of 74 HDR images and tone map each of them using 9 different TMOs, thus producing 666 tone mapped images. Each tone mapped image was evaluated by a group of human subjects who ranked the tone mapped images between 1 and 9. The mean opinion scores for the images were normalized to a $[0, 100]$ scale, following the convention used in ESPL LIVE. The reference images, the tone mapped images, and the mean opinion scores are available in the public domain (will be provided with the published article). Note that ESPL LIVE does not provide reference images; hence, its use is limited to the design and evaluation of blind metrics, whereas the proposed dataset can be used for FR and blind metrics both. The only other noteworthy dataset of tone mapped images with reference images is TMID [46], which has only 120 images. Therefore, the proposed dataset is a valuable addition that can be useful in advancing research in the domain of TM-IQA.

Figure 1 employs Whisker plots to illustrate the variation in scores attributed to each TMO. The TMOs utilized to compile the dataset include Drago [13], Expo [37] , Kim and Kautz (KKT) [36], Larson [42], Logar [29], Normal [5], Reinhard [40], Tumblin [38], and Ward Global [12]. These algorithms are renowned, and their source codes are accessible in Banterle's HDR toolbox (https://github.com/banterle/HDR_Toolbox), ensuring error-free implementations and enabling result reproduction by readers. The 74 HDR images used to generate the dataset are sourced from the TMID [46], featuring diverse scenes under various lighting conditions. Figure 2 presents a set of nine tone-mapped images reproduced from a typical HDR image within this dataset.

4 The Proposed Deep-Learning Model

In this study, a deep learning model is proposed for predicting image quality scores using convolutional neural networks (CNNs) [1]. The proposed model

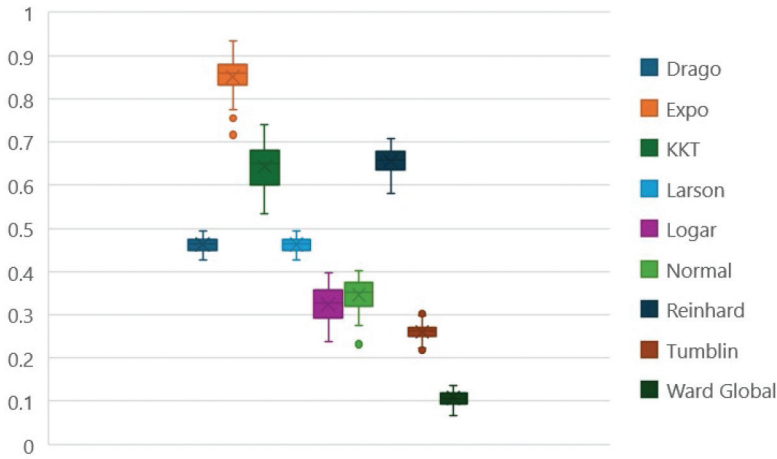

Figure 1: Visual representation of subjective scores with respect to TMOs.

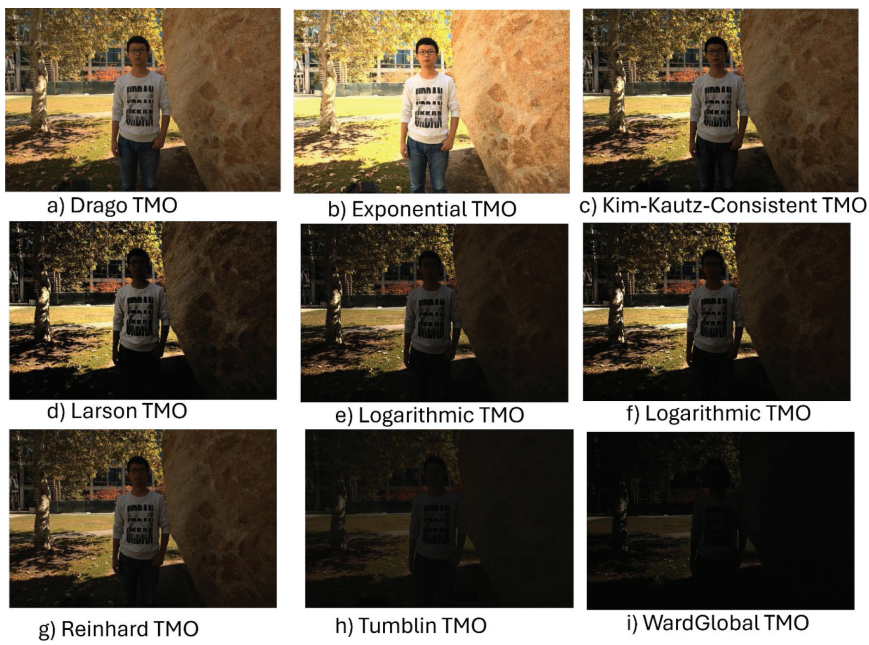

Figure 2: Representation of Each TMO result.

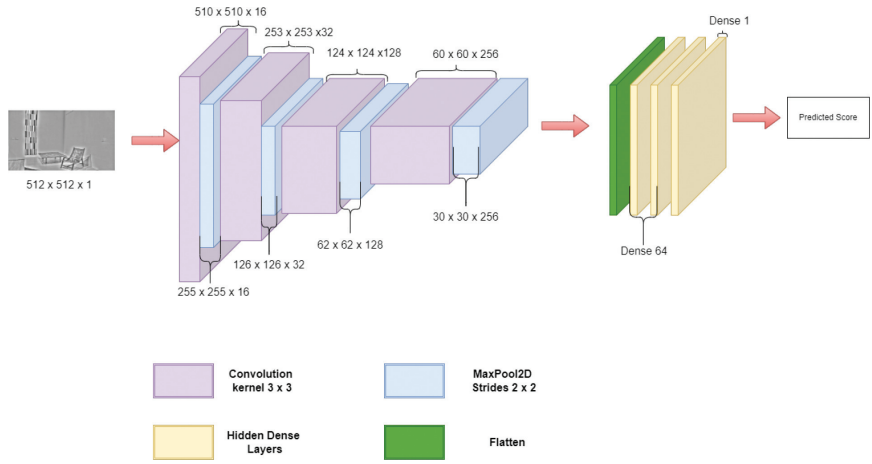

Figure 3: Proposed Model Architecture, containing several layers including convolutional, max-pooling, flatten and dense layers.

comprises multiple CNN and two dense layers, with a linear activation function in the output layer, as shown in Figure 3. The input to the model is an image of size $512 \times 512 \times 1$, represented as a single grayscale channel.

The CNN layers are designed to extract increasingly complex features from the input image using a combination of convolutional filters, activation functions, and max-pooling layers. The output of each CNN layer is passed through a max-pooling layer, which reduces the spatial dimensions of the output while retaining the most significant features. This process is repeated for multiple CNN layers, with increasing filter sizes and numbers of filters, to extract high-level features from the image.

After the final CNN layer, the output is flattened and passed through two dense layers with ReLU [15] activation functions. The purpose of the dense layers is to combine the high-level features extracted by the CNN layers and to generate a single output value that represents the predicted image quality score. Activation functions introduce non-linearity to the network and help in modeling complex relationships. Popular choices include ReLU (Rectified Linear Unit), sigmoid, and tanh. The selection of activation functions depends on the problem domain and the characteristics of the data. The final activation function in our design is linear [28] because the output of the model is continuous rather than a categorical value.

To finetune the performance, we experimented with different network configurations, including changing the number of layers and the number of nodes in each layer. We used the same training and validation data for all configurations. Through this experimentation, we determined the optimal

Table 1: Parameters of the proposed model for quality assessment of tone mapped low dynamic range images.

Layer	Type	Shape	# Parameters
conv2d	Conv2D	(None , 510, 510, 16)	160
max_pooling2d	Max_Pooling2D	(None , 225, 225, 16)	0
conv2d_1	Conv2D	(None , 253, 253, 32)	4640
max_pooling2d_1	Max_Pooling2D	(None , 126, 126, 32)	0
conv2d_2	Conv2D	(None , 124, 124, 128)	36992
max_pooling2d_2	Max_Pooling2D	(None , 62, 62, 128)	0
conv2d_3	Conv2D	(None , 60, 60, 256)	295168
max_pooling2d_3	Max_Pooling2D	(None , 30, 30, 256)	0
flatten	Flatten	(None , 230400)	0
dense	Dense	(None , 64)	14745664
dense_1	Dense	(None , 64)	4160
dense_2	Dense	(None , 1)	65

Note: Total parameters: 15,086,849
Trainable parameters: 15,086,849
Non-trainable parameters: 0

hyperparameters for the model, which were used to train the final model. These parameters are described in Table 1. Overall there are 15 million trainable parameters in the proposed model.

5 Performance Validation

In the previous section, we discussed our strategy to design a reliable metric for TM-IQA involving designing and training a neural network using CNNs. We also presented a finetuned set of hyperparameters for accurate score prediction. The results of the proposed deep learning model are very promising on both datasets discussed in Section 3. In this section, we present a detailed comparison of the proposed metric with the existing state of that art algorithms.

To train the model, we used the combined dataset of 2477 images, containing 1811 images from the ESPL LIVE dataset and 666 images of our own, each labeled with a quality score ranging from 0 to 100. To prepare the data for training, we performed some preprocessing steps, which included resizing the images to 512×512 pixels, computing the luminance to transform 3-channel color images to single-channel grayscale images, and normalizing pixel values in $[0, 1]$ range by dividing them by 255.

The images were split randomly into three sets, with 70% of the data used for training, 20% used for testing, and 10% used for validation. For training, we used the mean squared error (MSE) loss function to measure the difference

between predicted and actual scores. Stochastic gradient descent (SGD) was used as the optimization algorithm. Learning rate determines the step size at each iteration during gradient descent optimization. A larger learning rate can lead to faster convergence but may cause instability, while a smaller learning rate can result in slower convergence. We determined a suitable learning rate of 0.001 through experimentation which lead to convergence at a reasonably fast speed.

The network was trained for 30 epochs with a batch size of 16. During training, the data is divided into batches, and the weights are updated after processing each batch. The batch size determines the number of samples seen before each update. Smaller batch sizes provide faster convergence, but larger batch sizes may yield a more accurate gradient estimate.

During the learning process, we fine-tuned various hyperparameters, including the learning rate, regularization strength, and batch size. The error graph was monitored during the training process, as shown in Figure 4, which indicated that the model was learning well without overfitting or underfitting. It can be seen that the error reduces drastically after a few iterations, indicating that the model well-learned the relevant features for quality evaluation from the data, thereby validating that the design configurations and the hyperparameters were appropriate. Here we would like to address the temporary spike observed in the validation error around epoch 20. It is not uncommon to see such fluctuations in the learning process. These are mainly caused by data shuffling, which introduces new patterns to the model, leading to brief adjustments in performance. This effect normalizes in subsequent epochs, as shown in the figure.

The most common way to evaluate the performance of an objective metric is to study how aligned its scores are with the mean opinion scores assigned by the human subjects. Pearson's correlation coefficient [4], Spearman's rank-order correlation coefficient [34], and Kendall's rank-order correlation coefficient [31] are widely used algorithms to measure this correlation between the two scores. Pearson's correlation coefficient calculates the covariance of the two variables divided by the product of their standard deviations. For a sample size n of two variables x and y, the Pearson's coefficient can be written as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2 \cdot \sum_{i=1}^{n}(y-\bar{y})^2}} \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of $x$and y. Spearman's rank-order correlation coefficient measures the monotonicity of the relationship between two sets, and it is a variant of the Pearson Correlation Coefficient for the data that is ranked (and not scored on a continuous scale). If ranks of values in the sets

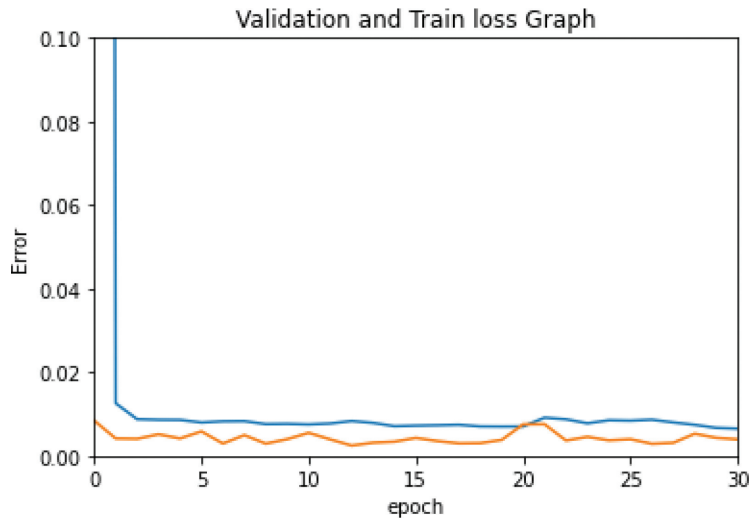

Figure 4: Training (blue curve) and Validation (orange curve) loss Graph. The error drops after a few epochs indicating that the design and the selected hyperparameters are adequate.

are distinct integers, then Spearman coefficient can be calculated as:

$$SRCC = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2 - 1\right)} \tag{2}$$

where d is the difference between the objective and subjective ranks, and N is the number of tone mapped images in one set. Kendall's rank-order correlation coefficient also finds the correlation between the ranks of scores and is defined as

$$KRCC = \frac{2(N_c - N_d)}{N\left(N - 1\right)} \tag{3}$$

where N_c and N_d are the number of concordant and discordant pairs. The term concordant is used for the pairs with the same relative order of magnitudes in the objective and subjective scores, and those which do not meet this condition are called discordant. The correlation coefficients lie in the range of $[0, 1]$, where a higher value indicates a better correlation. A high correlation value indicates that the metric can well-replicate the laborious and time-consuming subjective studies for assessment of the quality of the image.

Table 2 compares our method with 20 advanced state-of-the-art techniques currently available in the existing literature using our proposed dataset of 666 images. Our dataset provides reference images; therefore, both FR and blind metrics are included in this study. Scores are calculated for the images in the dataset using each metric, and their correlation with the subjective scores

Table 2: Correlation of subjective and predicted scores on the proposed dataset. The dataset provides reference and tone mapped images; therefore, both full-reference and blind metrics can be compared. The winner and the runner up scores are shown in bold faces. The proposed method outperforms all other metrics.

	Pearson		**Spearman**		**Kendall**	
Methods	**Correlation Coefficient**	***p*-value**	**Correlation Coefficient**	***p*-value**	**Correlation Coefficient**	***p*-value**
SSIM [45]	0.7758	0.0165	0.8451	0.0094	0.7027	0.0139
FSIM [48]	0.6267	0.1047	0.4489	0.3132	0.3296	0.3504
BRISQUE [33]	0.8503	0.0060	0.8320	0.0182	0.7134	0.0246
TMQI [46]	0.8798	0.0033	0.8620	**0.0068**	0.7408	0.0091
TMQI_II [30]	0.6128	0.1707	0.7910	0.0468	0.6662	0.0583
FSITM_RED [35]	0.0028	0.5450	0.1916	0.5428	0.1987	0.4766
FSITM_GREEN [35]	0.1521	0.4720	0.2613	0.4781	0.2383	0.4486
FSITM_BLUE [35]	0.2931	0.3794	0.3623	0.3806	0.3114	0.3690
TMQI+FSITM_RED	0.7732	0.0274	0.7525	0.0378	0.6273	0.0373
TMQI+FSITM_GREEN	0.8066	0.0160	0.7883	0.0253	0.6578	0.0288
TMQI+FSITM_BLUE	0.8185	0.0127	0.8066	0.0192	0.6791	0.0216
HIGRADE_II [27]	**0.8822**	**0.0032**	**0.8677**	0.0077	**0.7560**	**0.0090**
NLPD	0.7192	0.0399	0.7013	0.0509	0.5306	0.0841
FFTMI [25]	0.5354	0.1949	0.5922	0.1286	0.5093	0.0980
GSF [43]	0.0448	0.2954	0.1373	0.3119	0.1195	0.3260
VQGC [14]	0.5498	0.1725	0.3071	0.3840	0.2154	0.3793
BMPRI [32]	0.8066	0.0136	0.8326	0.0121	0.6989	0.0171
BLINDS-II [39]	0.3752	0.3351	0.3889	0.2951	0.3758	0.3354
BTMQI [9]	0.0845	0.3125	0.0829	0.3521	0.0819	0.4156
LSHS [2]	0.8125	0.0109	0.7960	0.0175	0.6509	0.0227
Proposed	**0.9232**	**0.0016**	**0.9281**	**0.0013**	**0.8301**	**0.0060**

is calculated using Pearson [4], Spearman [34], and Kendall [31] correlation algorithms. Our proposed metric demonstrates outstanding performance surpassing all existing methods. It obtained the highest correlation coefficients among all existing techniques.

The table also shows the confidence level (p-values) in computing the correlation. The p-value indicates the probability that the null hypothesis (zero correlation between metric and subjective scores) is true. The p-values obtained for the correlation coefficients in Table 2 are generally small, except for the cases when the correlation is very low. For the proposed metric, the p-values are close to zero, indicating a nearly 100% confidence level.

In the second experiment, we tested our algorithm on the ESPL LIVE HDR dataset [26]. Again, our method showed a high correlation to the subjective scores compared to other state-of-the-art methods, as demonstrated in Table 3. In this experiment, only blind metrics are included since ESPL LIVE dataset does not provide reference images. The proposed metric performed well and easily surpassed the performance of all other metrics.

The metrics used for comparison with the proposed method in Tables 2 and 3 are generally feature-based methods. These methods have certain advantages, such as shorter inference time, smaller model sizes, and lower computation requirements. However, deep learning models can generally surpass them in terms of accuracy. The disadvantages of deep models are their large size, i.e., the number of layers, which increases the computational complexity and training requirements. In Table 4, we have shown a comparison of the proposed method with three state of the art deep learning based blind IQA metrics proposed by Zhang *et al.* [49], Bosse *et al.* [6], and Jia *et al.* [20]. These methods employ deep learning models featuring 16, 14, and 10 layers, respectively; whereas our approach stands out by utilizing only 4 layers, tailored to accommodate smaller datasets, and achieving computational efficiency. In terms of training data, Jia *et al.* [20] and Zhang *et al.* [49] work with 2055 and 852,891 images respectively, while Bosse *et al.* [6] leverage a substantial training dataset comprising 294 million image patches. In contrast, our model is trained on a very small number of images, 500 and 1358, respectively when tested on the proposed custom and the ESPL LIVE datasets. Despite its significantly smaller scale in both computational complexity and training data, our proposed method exhibits slightly inferior yet comparable accuracy to these deeper models on the ESPL LIVE dataset. On the proposed dataset, our model ranks second, outperformed only by Bosse *et al.* [6].

Overall, the proposed method represents a significant advancement in the field of image quality assessment. By combining machine learning and image processing techniques, it is able to accurately and efficiently assess image quality, even in the presence of different types of distortions. Its impressive results and high correlation with human scores make it a promising tool for a

Table 3: Correlation of subjective and predicted scores on the ESPL LIVE dataset [26]. The dataset contains tone mapped images but does not provide reference images; therefore, only blind metrics are included. The winner and the runner up scores are shown in bold faces. The proposed method outperforms all other metrics.

	Pearson		**Spearman**		**Kendall**	
Methods	**Correlation Coefficient**	***p*-value**	**Correlation Coefficient**	***p*-value**	**Correlation Coefficient**	***p*-value**
BRISQUE [33]	0.6528	0.1251	0.6722	0.1162	0.6125	0.1925
BTMQI [9]	0.3903	0.3294	0.4014	0.3126	0.3864	0.3562
VQGC [14]	0.7234	0.0220	0.7346	0.0192	0.7054	0.0243
GSF [43]	**0.7859**	**0.0151**	**0.7918**	**0.0112**	**0.7524**	**0.0182**
BLINDS-II [39]	0.0374	0.5143	0.0918	0.4618	0.0346	0.5952
Proposed	**0.8025**	**0.0134**	**0.8212**	**0.0123**	**0.7953**	**0.0192**

Table 4: Comparison with large deep learning based blind IQA metrics. Despite the significantly smaller scale of model size and training data, the proposed model showed reasonable performance. The proposed method outperforms all other metrics.

				Accuracy (Correlation Coefficient)		
Methods	Test Dataset	Number of Layers	Training dataset Size	Pearson	Spearman	**Kendall**
Zhang *et al.* [49]	ESPL LIVE	16	852,891	**0.9350**	**0.9680**	**0.9225**
Bosse *et al.* [6]		14	294 million	**0.9725**	**0.9625**	**0.9415**
Jia *et al.* [20]		10	2055	0.8887	0.9025	0.7925
Proposed		**4**	**500**	0.8025	0.8212	0.7953
Zhang *et al.* [49]	Custom Dataset	16	852,891	0.8815	0.8955	0.8025
Bosse *et al.* [6]		14	294 million	**0.9725**	**0.9625**	**0.9415**
Jia *et al.* [20]		10	2055	0.9102	0.9135	0.8124
Proposed		**4**	**1358**	**0.9232**	**0.9281**	**0.8301**

wide range of applications, including digital image processing, computer vision, and multimedia technology.

6 Conclusion

This paper introduced a novel no-reference metric for evaluating the quality of tone mapped images. The metric constitutes a deep neural network trained on a large number of images labeled for their quality by human subjects. The metric scores are compared with the subjective scores by computing three widely-used algorithms, and a very high correlation between the two is observed. Therefore, the proposed algorithm can effectively replicate human subjects' observation and thus eliminate the need for time-consuming subjective studies for image quality assessment. This performance makes it potentially a valuable tool for various applications, including digital image processing, computer vision, and multimedia technology.

The paper also proposed a new dataset of 666 tone mapped images. Iin our experiments, the performance of some existing metrics did not remain consistent across different datasets, indicating that these metrics were not trained on a diverse set of images. The images included in our proposed dataset constitute a variety of scenes and were evaluated by 20 human subjects each. To our knowledge, this is the largest full-reference dataset of labeled tone mapped images, and therefore, it can be instrumental in advancing the research in tone mapping and tone mapped image quality assessment.

References

[1] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a Convolutional Neural Network", in *Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017*, 2017, https://doi.org/10.1109/ICEngTechnol.2017.8308186.

[2] T. Alotaibi, I. R. Khan, and F. Bourennani, "Quality Assessment of Tone-mapped Images Using Fundamental Color and Structural Features", *IEEE Transactions on Multimedia*, 26, 2023, 1244–54, https://doi.org/10.1109/TMM.2023.3278989.

[3] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H. P. Seidel, "Dynamic Range Independent Image Quality Assessment", *ACM Transactions on Graphics*, 27(3), 2008, https://doi.org/10.1145/1360612.1360668.

[4] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson Correlation Coefficient", *Springer Topics in Signal Processing*, 2, 2009, https://doi.org/10.1007/978-3-642-00296-0_5.

[5] M. A. V. A. Bernardo, "Quality Perception and Chromatic Changes in Digital Images", 2017, PhD Thesis, Universidade da Beira Interior (Portugal), https://search.proquest.com/openview/b485b316947e1bb9d99e63abd04f924e/1?pq-origsite=gscholar&cbl=2026366, (accessed on 03/28/2024).

[6] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment", *IEEE Transactions on image Processing*, 27(1), 2017, 206–19.

[7] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of HDR Tone Mapping Methods Using Essential Perceptual Attributes", *Computers and Graphics (Pergamon)*, 32(3), 2008, 330–49, https://doi.org/10.1016/j.cag.2008.04.003.

[8] H. Chen *et al.*, "Quality Evaluation of Arbitrary Style Transfer: Subjective Study and Objective Metric", *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, https://ieeexplore.ieee.org/abstract/document/9994780/ (accessed on 03/28/2024).

[9] P. Chen, L. Li, X. Zhang, S. Wang, and A. M. Tan, "Blind Quality Index for Tone-Mapped Images based on Luminance Partition", *Pattern Recognition*, 89, 2019, 108–18, https://doi.org/10.1016/j.patcog.2019.01.010.

[10] Y. Cui, M. Yu, G. Jiang, Z. Peng, and F. Chen, "Blind Tone-Mapped HDR Image Quality Measurement by Analysis of Low-level and High-level Perceptual Characteristics", *IEEE Transactions on Instrumentation and Measurement*, 71, 2022.

[11] S. B. Deal *et al.*, "Crowd-Sourced Assessment of Technical Skills: An Opportunity for Improvement in the Assessment of Laparoscopic Surgical Skills", *American Journal of Surgery*, 211(2), 2016, https://doi.org/10.1016/j.amjsurg.2015.09.005.

[12] P. Debevec and S. Gibson, "A Tone Mapping Algorithm for High Contrast Images", in *13th eurographics workshop on rendering: Pisa, Italy. Citeseer*, 2002, https://pages.cs.wisc.edu/~lizhang/courses/cs766-2007f/projects/hdr/Ashikhmin2002ATM.pdf (accessed on 03/28/2024).

[13] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive Logarithmic Mapping For Displaying High Contrast Scenes", *Computer Graphics Forum*, 22(3), 2003, 419–26, https://doi.org/10.1111/1467-8659.00689.

[14] Y. Fang *et al.*, "Blind Quality Assessment for Tone-Mapped Images by Analysis of Gradient and Chromatic Statistics", *IEEE Transactions on Multimedia*, 23, 2021, 955–66, https://doi.org/10.1109/TMM.2020.2991528.

[15] A. M. Fred Agarap, "Deep Learning using Rectified Linear Units (ReLU)", *arXiv:1803.08375*, 2019.

[16] K. Gu *et al.*, "Blind Quality Assessment of Tone-Mapped Images Via Analysis of Information, Naturalness, and Structure", *IEEE Transactions on Multimedia*, 18(3), 2016, 432–43, https://doi.org/10.1109/TMM.2016.2518868.

[17] H. Hadizadeh and I. V. Bajić, "Full-Reference Objective Quality Assessment of Tone-Mapped Images", *IEEE Transactions on Multimedia*, 20(2), 2018, https://doi.org/10.1109/TMM.2017.2740023.

[18] Q. He, D. Li, T. Jiang, and M. Jiang, "Quality Assessment for Tone-Mapped HDR Images using Multi-Scale and Multi-Layer Information", in *IEEE International Conference on Multimedia and Expo Workshops*, 2018, https://doi.org/10.1109/ICMEW.2018.8551502.

[19] Z. He, M. Yu, F. Chen, Z. Peng, H. Xu, and Y. Song, "Blind Tone-Mapped Image Quality Assessment Based on Regional Sparse Response and Aesthetics", *Entropy*, 22(8), 2020, 850, https://doi.org/10.3390/E22080850.

[20] S. Jia, Y. Zhang, D. Agrafiotis, and D. Bull, "Blind High Dynamic Range Image Quality Assessment using Deep Learning", in *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2017, 765–9, https://ieeexplore.ieee.org/abstract/document/8296384/ (accessed on 03/27/2024).

[21] M. Jiang, L. Shen, L. Zheng, M. Zhao, and X. Jiang, "Tone-Mapped Image Quality Assessment for Electronics Displays by Combining Luminance Partition and Colorfulness Index", *IEEE Transactions on Consumer Electronics*, 66(2), 2020, 153–62, https://doi.org/10.1109/TCE.2020.2985742.

[22] Q. Jiang *et al.*, "Single Image Super-Resolution Quality Assessment: A Real-World Dataset, Subjective Studies, and An Objective Metric", *IEEE Transactions on Image Processing*, 31, 2022, 2279–94.

[23] Q. Jiang, Y. Gu, C. Li, R. Cong, and F. Shao, "Underwater Image Enhancement Quality Evaluation: Benchmark Dataset and Objective Metric", *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9), 2022, 5959–74.

[24] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "BLIQUE-TMI: Blind Quality Evaluator for Tone-Mapped Images Based on Local and Global Feature Analyses", *IEEE Transactions on Circuits and Systems for Video Technology*, 29(2), 2019, https://doi.org/10.1109/TCSVT.2017.2783938.

[25] L. Krasula, K. Fliegel, and P. Le Callet, "FFTMI: Features Fusion for Natural Tone-Mapped Images Quality Evaluation", *IEEE Transactions on Multimedia*, 22(8), 2020, 2038–47, https://doi.org/10.1109/TMM.2019.2952256.

[26] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-scale crowdsourced study for tone-mapped HDR pictures", *IEEE Transactions*

on Image Processing, 26(10), 2017, 4725–40, https://doi.org/10.1109/TIP.2017.2713945.

[27] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-Reference Quality Assessment of Tone-Mapped HDR Pictures", *IEEE Transactions on Image Processing*, 26(6), 2017, https://doi.org/10.1109/TIP.2017.2685941.

[28] Z. Liao and G. Carneiro, "On the Importance of Normalisation Layers in Deep Learning with Piecewise Linear Activation Units", in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 2016, https://doi.org/10.1109/WACV.2016.7477624.

[29] "Logarithmic Tone Mapping Algorithm Based on Block Mapping Fusion |IEEE Conference Publication |IEEE Xplore", https://ieeexplore.ieee.org/document/8455806 (accessed on 03/28/2024).

[30] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang, "High Dynamic Range Image Compression by Optimizing Tone Mapped Image Quality Index", *IEEE Transactions on Image Processing*, 24(10), 2015, 3086–97, https://doi.org/10.1109/TIP.2015.2436340.

[31] A. I. McLeod, "Kendall Rank Correlation and Mann-Kendall Trend Test", *R Package "Kendall."*, 602, 2011.

[32] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind Image Quality Estimation via Distortion Aggravation", *IEEE Transactions on Broadcasting*, 64(2), 2018, 508–17, https://doi.org/10.1109/TBC.2018.2816783.

[33] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain", *IEEE Transactions on Image Processing*, 21(12), 2012, 4695–708, https://doi.org/10.1109/TIP.2012.2214050.

[34] L. Myers and M. J. Sirois, "Spearman Correlation Coefficients, Differences between", *Wiley StatsRef: Statistics Reference Online*, 2014, https://doi.org/10.1002/9781118445112.stat02802.

[35] Z. H. Nafchi, A. Shahkolaei, R. Farrahi Moghaddam, and M. Cheriet, "FSITM: A Feature Similarity Index For Tone-Mapped Images", *IEEE Signal Processing Letters*, 22(8), 2015, 1026–9, https://doi.org/10.1109/LSP.2014.2381458.

[36] J.-S. Pang, "Serial and Parallel Computation of Karush–Kuhn–Tucker Points via Nonsmooth Equations", *SIAM J. Optim*, 4(4), 1994, 872–93, https://doi.org/10.1137/0804050.

[37] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic Tone Reproduction for Digital Images", in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, ed. M. C. Whitton, New York, NY, USA: ACM, 2023, 661–70, https://doi.org/10.1145/3596711.3596781.

[38] A. Rosenfeld, "Image Analysis and Computer Vision: 1993", *CVGIP: Image Understanding*, 59(3), 1994, 367–404.

[39] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain", *IEEE Transactions on Image Processing*, 21(8), 2012, 3339–52, https://doi.org/10.1109/TIP.2012.2191563..

[40] Y. Salih, A. S. Malik, and N. Saad, "Tone Mapping of HDR Images: A Review", in *2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012)*, IEEE, 2012, 368–73, https://ieeexplore.ieee.org/abstract/document/6306220/ (accessed on 03/28/2024).

[41] Y. Song, G. Jiang, M. Yu, Z. Peng, and F. Chen, "Quality Assessment Method Based on Exposure Condition Analysis for Tone-Mapped High-Dynamic-Range Images", *Signal Processing*, 146, 2018, 33–40, https://doi.org/10.1016/j.sigpro.2017.12.020.

[42] "The study of logarithmic image processing model and its application to image enhancement - PubMed", https://pubmed.ncbi.nlm.nih.gov/18290000/ (accessed on 03/28/2024).

[43] D. Varga, "No-Reference Image Quality Assessment with Global Statistical Features", *Journal of Imaging*, 7(2), 2021, https://doi.org/10.3390/jimaging7020029.

[44] X. Wang, Q. Jiang, F. Shao, K. Gu, G. Zhai, and X. Yang, "Exploiting Local Degradation Characteristics and Global Statistical Properties for Blind Quality Assessment of Tone-Mapped HDR Images", *IEEE Transactions on Multimedia*, 23, 2021, https://doi.org/10.1109/TMM.2020.2986583.

[45] 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*, 13(4), 2004, 600–12, https://doi.org/10.1109/TIP.2003.819861.

[46] H. Yeganeh and Z. Wang, "Objective Quality Assessment of Tone-Mapped Images", *IEEE Transactions on Image Processing*, 22(2), 2013, 657–67, https://doi.org/10.1109/TIP.2012.2221725.

[47] G. Yue, W. Yan, and T. Zhou, "Referenceless Quality Evaluation of Tone-Mapped HDR and Multiexposure Fused Images", *IEEE Transactions on Industrial Informatics*, 16(3), 2020, 1764–75, https://doi.org/10.1109/TII.2019.2927527.

[48] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment", *IEEE Transactions on Image Processing*, 20(8), 2011, 2378–86, https://doi.org/10.1109/TIP.2011.2109730.

[49] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network", *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1), 2020, 36–47.