# BLIND HIGH DYNAMIC RANGE IMAGE QUALITY ASSESSMENT USING DEEP LEARNING

*Sen Jia*

Intelligent Systems Laboratory
University of Bristol

*Yang Zhang, Dimitris Agrafiotis and David Bull*

Bristol Vision Institute
University of Bristol

## ABSTRACT

In this paper we propose a No-Reference Image Quality Assessment (NR-IQA) method on High Dynamic Range (HDR) images by combining deep Convolutional Neural Networks (CNNs) with saliency maps. The proposed method utilises the power of deep CNN architectures to extract quality features which can be applied cross HDR and Standard Dynamic Range (SDR) domains. To introduce human visual system to CNNs, a saliency map algorithm is used to select a subset of salient image patches to evaluate on. Our CNN-based method delivers a state-of-the-art performance in HDR NR-IQA experiment, competitive with full reference IQA methods.

***Index Terms*—** HDR, No-Reference Image Quality Assessment, Deep Learning, Saliency Map

## 1. INTRODUCTION

No Reference (NR) image (and video) quality metrics have until recently been associated with poor performance in estimating the perceived quality of an image or video [1]. Recently deep learning based NR metrics have been proposed to estimate the quality of images without any reference to the original [2, 3, 4]. These have produced some promising results, closing the gap to the performance delivered by full reference methods. As with Standard Dynamic Range (SDR) images, Image Quality Assessment (IQA) for HDR images is more challenging and even more so when reference image is unavailable. This paper describes a deep learning method to NR-IQA for HDR images that offers performance very close to that achieved with Full Reference (FR) HDR quality metrics. Our method combines a deep Convolutional Neural Network (CNN) architecture with the saliency map proposed in [5]. The tone-mapped images using [6] are used to compare against the result achieved on HDR images.

Many of the existing CNN-based methods for NR-IQA employ CNNs to extract image quality features. For those CNN-based methods, an input image is split into multiple patches and a quality estimation is performed based on the features presented within each patch. An early attempt at introducing CNNs for NR-IQA was proposed by Kang et al. [2]. A CNN architecture with one convolutional layer achieved competitive results with FR-IQA on the LIVE [7] dataset. Later they further extended this CNN architecture for NR-IQA and distortion type classification in [3]. It has been shown that a CNN with more layers can extract a more discrminative feature for object recognition [8, 9, 10]. Compared to [2, 3], our proposed method employs a CNN with significantly more convolutional layers (ten as opposed to one layer) for feature extraction. Recently Bosse et al. [4] proposed a CNN-based NR-IQA method also employing multiple layers on SDR images. But they tried to solve the issue of equal weight patch by learning a weight parameter from the activation of the rectified linear unit. The weight parameter is learned by CNN model itself such that the importance of a patch may not consistent with human vision systems. Our method attempts to solve the problem of equal weight patch by using the saliency map [5] in order that a CNN model is only applied on a subset of salient image patches for evaluation.

Unnoticeable qulity distortion in SDR domain may become more obvious in HDR domain [6]. Thus a tone-map is required to convert HDR image to SDR range for IQA. But cross-dataset IQA on HDR images is still challenging because each HDR dataset may have its own luminance distribution. Korshunov built a public HDR image dataset [11] on which Hanhart benchmarked most objective quality metrics [12]. The best FR-IQA on the XT dataset was HDR-VDP proposed by Narwaria [13] and the algorithm was specifically designed for HDR content. After tone-mapping HDR images to SDR using [6], the algorithm of [14] achieved the best result in the domain of Perceptually Uniform (PU) [6]. For NR-IQA on the XT dataset, the method of [15] achieved the highest result in the both domains of HDR and PU. We compare our method with state-of-the-art FR and NR IQA on the same dataset. Another HDR dataset, JPEG [16], is also used to investigate the generalisability of our method.

Firstly we train a CNN model on the LIVE dataset [7] to learn SDR quality feature. We show that the trained model can extend SDR quality information on tone-mapped HDR images for NR-IQA. Secondly, when training and testing on the same HDR dataset, our method achieves competitive performance with FR-IQA on the XT dataset [11]. Thirdly, we

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

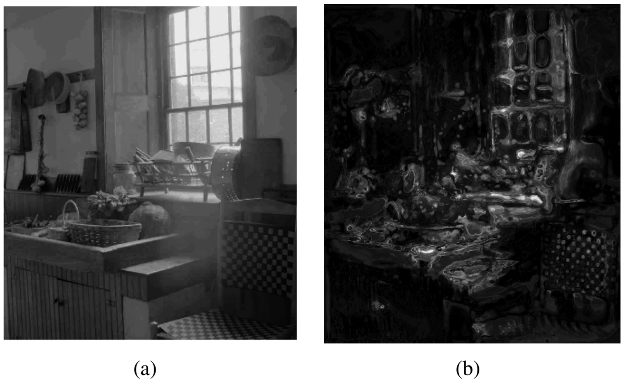


**Fig. 1**. (a) shows a tone-mapped HDR image using [6]. (b) shows the saliency map computed on the HDR image using [5].

train our method on one HDR dataset and evaluate the peroformance on the HDR dataset. The experiment shows that our method achieves good performances when directly applying on HDR images and a further improvement can be obtained by using the tone-mapping function of PU.

## 2. DATASETS

### 2.1. SDR Datasets

Two SDR datasets are used in our experiment, the LIVE [7] and the CSIQ datasets [17]. The LIVE dataset contains 799 images with five types of distortion noise. The ground truth label for the LIVE dataset is Differential Mean Opinion Scores (DMOS) in the range of [0,99]. We train an SDR CNN model on the LIVE dataset and apply it on HDR datasets for cross-domain experiment. The CSIQ dataset contains 866 images with six types of distortions and the ground truth of CSIQ is also DMOS in the range of [0,1]. Only the four shared distortion types, “JP2K”, “JPEG”, “WN”, “GBLUR”,are used to validate the proposed method between SDR datasets before applying the model on HDR images.

### 2.2. HDR Datasets

The HDR datasets used in this paper are XT [11] and JPEG [16]. The XT dataset contains 240 distorted HDR images. While the JPEG dataset contains 150 distorted HDR images. Even though there are fewer image samples comparing with SDR datasets, but the size of HDR images are normally much larger than SDR ones. That is more image patches can be obtained for CNNs to train. The average size of LIVE is (height:548, width:665) resulting in 340 ($32 \times 32$) patches. While the average size of XT is (height:1080, width:944) such that each image can deliver 957 patches for training. The ground truth of the two HDR datasets are both MOS in the range of [0,5]. We also convert the HDR datasets to SDR domain by using the tone-mapping algorithm proposed in [6]. The two tone-mapped datasets are referred as XTPU and JPEGPU respectively, see Section3.1.

## 3. METHODOLOGIES

For each dataset, we locally normalise every image using the algorithm in [2, 18]. Each image is split into a set of small patches in the size of $32 \times 32$ assigned with the same quality label. Like the work of [4], we train a CNN architecture on those image patch. But the difference in our method is that we use saliency map computed on each image to assign weight for each patch instead of learning the weight from network activation. Two measurements are applied, Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC).

### 3.1. Tone-Mapping

Comparing with SDR, HDR was designed to store a wider range of luminance values therefore an invisible distortion in SDR may become noticeable in HDR. To extend SDR quality metrics to HDR, Aydin [6] proposed a tone-mapping function that can apply SDR IQA methods on HDR, so called perceptually uniform encoding, as shown in Figure 1(a). Note that the mapping function between HDR and SDR is referred as tone-mapping in this paper to differentiate the logistic mapping applied for DMOS to MOS.

### 3.2. Local Normalization

Following the same preprocessing protocol in [18, 2, 3], a contrast normalization has been applied on each image before spliting into patches. This process might be important for cross-dataset evaluation between SDR and HDR. The pixel value range of a normalised image from either SDR or HDR is squashed into a small range centered at zero.

### 3.3. CNN Architecture

Kang [2] used only one convolutional layer followed by maxpool and minpool layers. Our proposed CNN architecture is similar to [8, 4] that ten convolutional layers with small receptive fields are stacked: conv3-32, conv3-32, maxpool2, conv3-64, conv3-64, maxpool2, conv3-128, conv3-128, maxpool2, conv3-256, conv3-256, maxpool2, conv3-512, conv3-512, fc2048, fc2048, softamx[1].

The input image patch is $32 \times 32$ and the convolutional kernel is $3\times3$. A $2\times2$ maxpool layer is added and the number of kernels is doubled every two convolutional layers. Two fully connected layers are added at the end of the model, each

[1] “conv3-64” represents a convolutional layer which contains 64 kernels with a size of 3 by 3. “maxpool2” denotes a max pooling layer that in the size of 2 by 2. “fc2048” denotes a fully connected layer which contains 2048 neurons.

of which has 2048 units. Dropout is added in the two fully connected layers with ratio of 0.5. We apply exponential linear units [19] after each convolutional and fully-connected layer.

### 3.4. Saliency Map

To better mimic human vision systems, saliency map is utilised to select a subset of salient image patches to evaluabte on. We apply the algorithm of [20] to compute saliency map on SDR images and the algorithm of [5] on HDR, as shown in Figure 1(b).

Every pixel value of an saliency map is rescaled to the range of [0,1]. We define the summation of pixel value within a saliency patch represents the importance of the image patch.

$$PI_i = \sum_{m=0}^{m=M-1} \sum_{n=0}^{n=N-1} s(m,n) \tag{1}$$

where $M$, $N$ is the size of the patch, $s(m,n)$ is pixel value of the saliency patch and $PI_i$ is the importance for $ith$ image patch in the range of [0,$M \times N$] ($M = N = 32$). We set a threshold $\theta$ to select a subset of salient image patches to evaluate on. The $ith$ image patch is considered to be salient if its importance $PI_i > \theta \times M \times N$. In our experiment the threshold is chosen from $\{0,0.01,0.1,0.5\}$. Note that when $\theta = 0$, no saliency map is applied because all the image patch is considered to be salient.

## 4. EXPERIMENTS

### 4.1. SDR cross-dataset

Our first experiment is to test if the quality feature learned from SDR distribution can generalise well on HDR images. An SDR CNN NR-IQA model is trained on all images from the LIVE dataset. The total number of training epochs is 15. The start learning rate is 0.001 and the momentum is 0.9 and they both reduce every five epoch by multiplying 0.1 and subtracting 0.1 respectively. We apply the model on the CSIQ dataset [17] to evaluate its SDR cross-dataset performance. No logistic mapping is applied because they both use DMOS for annotation. Our method achieves competitive results with other state-of-the-art SDR works [3], 0.9323 LCC and 0.9331 SROCC when $\theta = 0.5$, Table 1.

We then apply the model on the two HDR datasets to investigate the generalisability of the learned feature from SDR. A logistic mapping with five parameters [2] is applied to convert the output of the model from DMOS to MOS. Each of the two HDR datasets is split into two subsets, 80% of the total for training the mapping function and 20% is for evaluation. We shuffle and repeat this process ten times to report average accuracy. As shown in Table 1, the SDR model performs poorly on HDR images. Especially on the JPEG dataset, the highest average LCC is 0.2768 and 0.3039 on SROCC. On the tone-mapped image, 0.7953 LCC and 0.7981 SROCC were obtained on the JPEGPU dataset and the performance on the XT dataset is also increased. The model performs better on the tone-mapped image because the quality information is learned from SDR dataset.

**Table 1**. Training on the LIVE dataset, testing on the CSIQ, XT, JPEG and their tone-mapped datasets.

| $\theta$ | 0 | 0.01 | 0.1 | 0.5 |
|---|---|---|---|---|
| CSIQ-LCC | .9279 | .9280 | .9280 | **.9323** |
| CSIQ-SROCC | .9321 | .9321 | **.9331** | **.9331** |
| XT-LCC | **.7466** | .7215 | .7319 | .7373 |
| XT-SROCC | **.7450** | .7342 | .7367 | .7296 |
| JPEG-LCC | .2186 | .2476 | **.2768** | .2472 |
| JPEG-SROCC | .2399 | .2587 | **.3039** | .2526 |
| XTPU-LCC | .8768 | .8633 | .8778 | **.8919** |
| XTPU-SROCC | .8642 | .8610 | .8738 | **.8849** |
| JPEGPU-LCC | .7587 | .7373 | .7424 | **.7953** |
| JPEGPU-SROCC | .7517 | .7375 | .7345 | **.7981** |

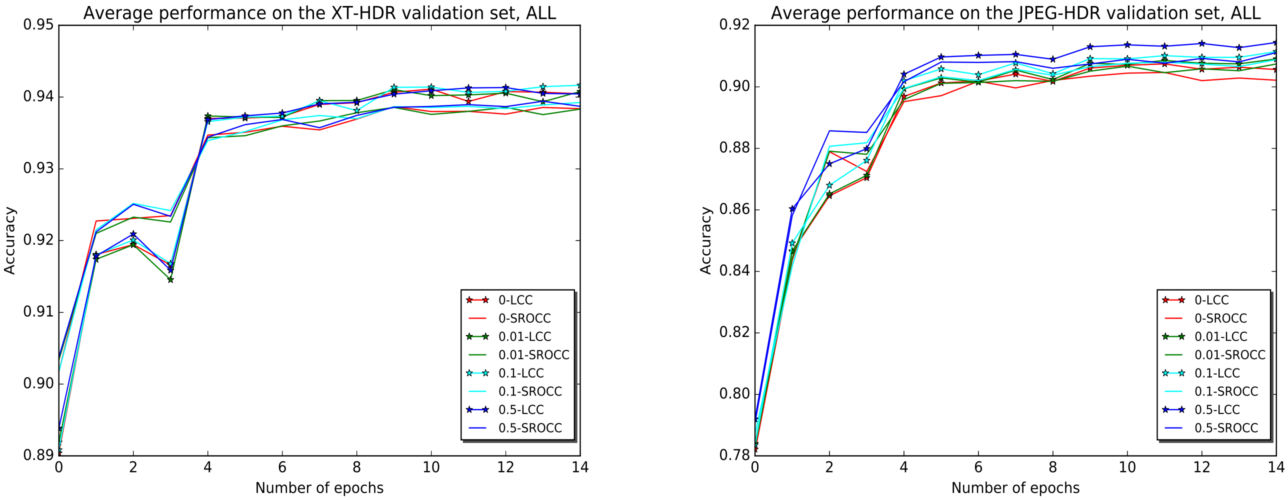
### 4.2. HDR within-dataset

The second experiment is to investigate if HDR image can offer more quality information to train a model for HDR IQA. We do not use a logistic mapping in this experiment and afterwards since the two HDR datasets share the same MOS label range. We firstly evaluate our method using within-dataset experiment setting that the training and test sets are from the same HDR dataset. For each of the two HDR datasets, we split it into 60% for training, 20% for validating and 20% for testing. The training protocol is the same as used on the LIVE dataset. In Figure 2, we show the average learning curve of ten splits on the validation set. Saliency maps delivered slightly higher LCC and SROCC performances.

the saliency map ($\theta = 0.5$) delivered a further improvement. During each split, we record the highest test accuracy based on LCC achieved on the validtion set. The average accuracy on the test set of ten random splits is reported in

**Table 2**. The average accuracy on the HDR test set.

| | |
|---|---|
| HDRVDP2-XT-LCC[12] | .9604 |
| HDRVDP2-XT-SROCC[12] | .9564 |
| MSSSIM_Y-XTPU-LCC[12] | .9447 |
| MSSSIM_Y-XTPU-SROCC[12] | .9501 |
| Marziliano_Y-XTPU-LCC[12] | .5114 |
| Marziliano_Y-XTPU-SROCC[12] | .4179 |
| Proposed-XT-LCC ($\theta = 0.5$) | .9291 |
| Proposed-XT-SROCC ($\theta = 0.5$) | .9301 |
| Proposed-JPEG-LCC ($\theta = 0.5$) | .8799 |
| Proposed-JPEG-SROCC ($\theta = 0.5$) | .8887 |



(a)

(b)

**Fig. 2**. LCC and SROCC accuracies on the validation set using different importance coefficient ($\theta = [0, 0.01, 0.1, 0.5]$).

Table 2 to compare with other methods. Note that the HDR-VDP [13] was only applied on HDR content because the algorithm was designed for absolute luminance values. The FR-IQA MSSSIM [14] and NR-IQA Marziliano [15] achieved better performance on the tone-mapped dataset. The proposed method achieves competitive result with state-of-the-art FR-IQA methods on the XT dataset, $0.9291$ LCC and $0.9301$ SROCC.

### 4.3. HDR cross-dataset

In the second experiment we show that our method can be directly applied on HDR images. But the generalisability plays a very important role for CNN-based IQA methods. Let's recall that there is no standard format of HDR luminance range. The learned quality feature by a CNN model may contain little in common with unseen HDR image. Therefore our third experiment is to train a CNN model on all images from one HDR dataset and test on the other, so called HDR cross-dataset evaluation.

**Table 3**. HDR Cross-dataset LCC and SROCC accuracies with different $\theta$ values.

| $\theta$ | 0 | 0.01 | 0.1 | 0.5 |
|---|---|---|---|---|
| XT-LCC | .7322 | .7326 | **.7329** | .7312 |
| XT-SROCC | .7838 | **.7840** | .7839 | .7797 |
| JPEG-LCC | .7862 | .7894 | .7905 | **.7933** |
| JPEG-SROCC | .7765 | .7806 | .7820 | **.7844** |
| XTPU-LCC | **.8637** | .8635 | .8631 | .8634 |
| XTPU-SROCC | **.8904** | **.8904** | .8896 | .8902 |
| JPEGPU-LCC | .8551 | **.8554** | .8548 | .8509 |
| JPEGPU-SROCC | .8621 | **.8630** | .8626 | .8570 |

Two models are trained seperately on the two HDR datasets following the same protocol used in the first experiment. We show the cross-dataset result on the two HDR datasets in Table 3. Using saliency map delivers a slightly better result but the highest accuracy ($\theta = 0.5$) in Table 3 is worse than the within-dataset result in Table 2. It is interesting to see that the cross-dataset LCC and SROCC on the JPEG dataset are much higher than the SDR model obtained in Table 1. The HDR datasets share more common quality feature than it between SDR and HDR. But the CNN learned quality feature may still be dataset-specific when comparing with within-dataset experiment. To further bridge the gap of image format between the HDR datasets, we repeat the HDR cross-dataset experiment on the PU datasets. In Table 3, we can see that the performance has been increased significantly.

## 5. CONCLUSION

In this papaer we proposed a NR-IQA method on HDR images using CNNs and saliency maps. We have shown that saliency maps can further increase the performance of CNNs for HDR NR-IQA. For different HDR datasets, it is more difficult to directly apply IQA methods due to the luminance gap. A tone-mapping function is currently required to standardise the domain among different HDR datasets or between SDR and HDR. It is still an open question that how to directly learn quality features which can be applied across different luminance ranges. One of our future work could be comparing SDR saliency maps computed on tone-mapped images with HDR saliency maps for NR-IQA.

## 6. REFERENCES

[1] Sheila S. Hemami and Amy R. Reibman, “No-reference image and video quality estimation: Applications and human-motivated design,” *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 469 – 481, 2010, Special Issue on Image and Video Quality Assessment.

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

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

[4] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, “A deep neural network for image quality assessment,” in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 3773–3777.

[5] A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, “A learning-based visual saliency fusion model for high dynamic range video (LBVS-HDR),” in *2015 23rd European Signal Processing Conference (EUSIPCO)*, Aug 2015, pp. 1541–1545.

[6] T. O. Aydın, R. Mantiuk, and H. Seidel, “Extending quality metrics to full dynamic range images,” in *Human Vision and Electronic Imaging XIII*, San Jose, USA, January 2008, Proceedings of SPIE, pp. 6806–10.

[7] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A statistical evaluation of recent full reference image quality assessment algorithms,” *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov 2006.

[8] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Eprint Arxiv*, 2014.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *CoRR*, vol. abs/1409.4842, 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015.

[11] 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 *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, May 2015, pp. 1–6.

[12] P. Hanhart, M. V. Bernardo, M. Pereira, A. M. G. Pinheiro, and T. Ebrahimi, “Benchmarking of objective quality metrics for hdr image quality assessment,” *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, pp. 39, 2015.

[13] M. Narwaria, R. Mantiuk, M. Perreira Da Silva, and P. Le Callet, “Hdr-vdp-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images,” *J. Electronic Imaging*, vol. 24, pp. 010501, 2015.

[14] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, Nov 2003, vol. 2, pp. 1398–1402 Vol.2.

[15] A. V. Murthy and L. J. Karam, “A matlab-based framework for image and video quality evaluation,” in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, June 2010, pp. 242–247.

[16] M. Narwaria, M. Perreira 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. 102008–102008, 2013.

[17] E.C. Larson and D.M. Chandler, “Most apparent distortion: full-reference image quality assessment and the role of strategy,” *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006–011006, 2010.

[18] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Unsupervised feature learning framework for no-reference image quality assessment,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1098–1105.

[19] D. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (ELUs),” *CoRR*, vol. abs/1511.07289, 2015.

[20] H. J. Seo and P. Milanfar, “Static and space-time visual saliency detection by self-resemblance,” *Journal of Vision*, vol. 9, no. 12, pp. 15, 2009.