Efficient Evaluation of Image Quality via Deep-Learning Approximation of Perceptual Metrics

Alessandro Artusi, Francesco Banterle, Fabio Carrara, and Alejandro Moreo

***Abstract*—Image metrics based on Human Visual System (HVS) play a remarkable role in the evaluation of complex image processing algorithms. However, mimicking the HVS is known to be complex and computationally expensive (both in terms of time and memory), and its usage is thus limited to a few applications and to small input data. All of this makes such metrics not fully attractive in real-world scenarios. To address these issues, we propose *Deep Image Quality Metric* (*DIQM*), a deep-learning approach to learn the global image quality feature (*mean-opinion-score*). DIQM can emulate existing visual metrics efficiently, reducing the computational costs by more than an order of magnitude with respect to existing implementations.**

***Index Terms*—Convolutional Neural Networks (CNNs), Objective Metrics, Image Evaluation, Human Visual System, JPEG-XT, and HDR Imaging.**

I. Introduction

The quality evaluation of image processing algorithms is an essential step that can be carried out either through a user study or using an objective metric. User studies are time-demanding and often impractical due to the large number of users and images required to guarantee the results to be statistically significant [1].

This issue has partially been overtaken by limiting the application of user studies to a subset of all test conditions. In this way, user studies provide a ground-truth reference for the choice of the most appropriate complex objective metric among a large set of candidates. This is typically achieved by identifying which among the objective metrics presents the highest correlation to the results of user studies, as evaluated on the same subset of test conditions [2], [3]. Although this helps to ease the tedious process of user studies, it does not cope with the fact that the findings extracted during such studies are often difficult to generalize. Notwithstanding this, the use of objective metrics is known to suffer from high computational complexity that derives from the complexity of simulating the many aspects of the *Human Visual System* (HVS) [4], [5]. *De facto*, this precludes such metrics from being applied to several quality assessment scenarios such as standardization, real-time quality assessment, etc.

This altogether motivates the need for more efficient (yet effective) computational metrics that can predict visually significant differences between any test image and its reference. A further desideratum is for this metric to be *differentiable*, so that it can be directly optimized for. In this regard, traditional error functions as, e.g., the squared L_2 norm of the pixel differences, are known to be poorly correlated with the image quality as perceived by the HVS [6].

The main focus of this work is to provide a practical solution to the aforementioned issues. In this paper, we investigate the use of deep learning to predict visual metric features of popular implementations, like the quality index Q for the *High Dynamic Range Visual Differences Predictor* (HDR-VDP), which is known to be correlated with the *mean-opinion-score* (MOS) [4] , and the *probability index* for the *Dynamic Range Independent Metric* (DRIM) metric, defined as the percentage of pixels that are above the probability detection threshold (*probability index*) [5].

We propose *Deep Image Quality metric* (*DIQM*), a model for learning visual metric features similar to other well-known existing objective metrics (e.g., HDR-VDP [4] and DRIM [5]) at a fraction of their computational costs (more than an order of magnitude faster).

We tested DIQM on a variety of scenarios designed to demonstrate its robustness and flexibility. Its real-time performance makes it suitable to be integrated as the main optimization component into different scenarios including the optimization of parameters of tone mapping, reverse tone mapping, and *High Dynamic Range* (HDR) compression (that we test in our experiments). Furthermore, the computational costs of our framework will allow standardization bodies (e.g., JPEG and MPEG) to employ substantially larger datasets than the ones being used today.

The main novelties and contributions of our work are summarized below:

- The task of evaluating image processing is formalized as an optimization problem aiming at learning visual metric features.
- DIQM generalizes to a variety of visual metrics producing results that are comparable to the ground-truth metric.
- DIQM significantly reduces the computational cost of current visual metrics, thus making visual metrics appealing for various real-world quality assessment scenarios.
- The code implementing DIQM is available online [1].

Note that, in this work, we do not attempt to predict probability maps of distortion. One reason for this decision is technical: as will be seen, the model we propose is based on a

A. Artusi is with the MRG DeepCamera Group, RISE Ltd. e-mail: (artusialessandro4@gmail.com.
F. Banterle, F. Carrara, and A. Moreo are with ISTI CNR, Italy.
Manuscript received ....; revised .....

[1] https://github.com/fabiocarrara/diqm

convolutional architecture. While convolutional processing of images is particularly robust to capturing global information in the input (thus useful in predicting a global scalar quality value), it typically presents some limitations in dealing with fine-grained aspects (that might be crucial for predicting distortion maps). Overcoming said limitations is possible, but only at the cost of sensibly augmenting the number of model parameters, and hence the number of training examples. This additional cost might not be worthwhile. The reason for this (and the main reason why we are reluctant to predict probability maps) is not merely technical: the utility of the distortion maps is recently becoming a focus of heated debate in the community. The main motivation behind this dismissal regards its high cost (i.e., from the point of view of computational cost and human effort). A typical scenario where these aspects arise concerns the comparison of several algorithms for standardization on huge image/video datasets. With the growing number of parameters and configurations to be tested, computing the distortion maps of DRIM and/or HDR-VDP might promptly become computationally intractable. Moreover, the evaluation of the distortion maps has to be carried out manually for each image or video frame. This implies the evaluation is a tedious and error-prone task. The human resources (economical and of time) this all implies rapidly becomes unaffordable, and this explains the general trend in the field (e.g., by the evaluation committees of JPEG and MPEG) of preferring scalar quality metrics to distortion maps; this is especially true in large-content datasets. This makes a natural choice for DIQM to predict the quality index Q for HDR-VDP and the probability index for DRIM, respectively.

II. Related Work

Techniques for image quality evaluation represent a cornerstone in the performance assessment of many processing algorithms spanning different areas including image encoding, acquisition, HDR imaging, or enhancement, to name a few. Image metrics can mainly be categorized into Image Quality metrics (IQMs) and visibility metrics (VMs). The former predict a single global quality score for the entire image. The latter instead predict the probability that a human observer could detect differences between a pair of images. Their output is a visibility map, in which each pixel value encodes the probability of detection.

This work focuses on learning image metrics of type IQM (i.e., a scalar value for the entire image). In this section, we will thus mainly concentrate on the discussion of state-of-the-art IQMs approaches. In general, IQMs can be subcategorized into two main classes: Fully-Reference (FR) and No-Reference (NR). FR-metrics receive as input a pair of images (the ground-truth and the distorted images), while the NR-metrics receive as input only the distorted image (i.e., without any prior information of how the free-distortions image should appear).

A. Full-Reference Metrics

Techniques within the FR class are differently characterized based on the type of approach they implement, e.g., those that directly measure differences between pixels, those that detect structural changes in the image (i.e., implement a local spatial measure of pixels value correlation), and those that model human vision aspects like contrast sensitivity, luminance adaptation, visual masking, etc.

Examples of the first category are the root mean square error (RMSE) and peak-signal-to-noise-ratio (PSNR) metrics, color based differences such as CIE-Lab color-space and its extension sCIE-Lab [7], which are typically used for comparing *Standard Dynamic Range* (SDR) content. The above objective metrics, which have been developed for SDR content, can be easily extended to work with HDR content as well by either using the Perceptually Uniform (PU) [8] or the Perceptual Quantizer (PQ) EOTF [9] to convert absolute display-referred HDR color values into perceptually uniform units.

Image structure-based quality measures are based on the observation that the Human Visual System (HVS) detects structural changes from scenes as part of the visualizing comprehension. One of the first image structure quality index are the SSIM [10] and its extension CW-SSIM [11]. While both of them focus on SDR content, the tone-mapped image quality index (TMQI) [12] can evaluate HDR content versus its tone mapped version.

HVS-based techniques can detect the visible differences between pixels so as to measure their magnitude in terms of the so-called *Just Noticeable Difference (JND). They have been developed for SDR [13] and HDR [4], [5] images.*

B. No-Reference Metrics

Several NR metrics in imaging applications have been proposed in the literature. Here, the main difficulty regards the absence of a ground-truth image as a reference. To overcome this issue, a possible approach is to model the image statistics. In this case, the assumption is that the ground-truth image occupies a subspace of the entire space of possible images, and the goal is thus to compute the distance from the distorted image [14], [15].

Other approaches have employed scene statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image due to the presence of distortions [16].

Yet another approach is to extract specific characteristics of the distortion that needs to be detected/measured. This specific knowledge can be derived from the existing know-how of the specific type of artifact and its unique characteristics; i.e., using local gradient [17], saliency map and Support Vector Regression (SVR) [18], measuring the power of the blocking signal [19], etc.

Finally, learning-based approaches can be used where extracted images' features are used to learn to distinguish from undistorted images. These techniques are further discussed in the next Section II-C.

C. CNNs-based Metrics

In this section, we discuss relevant CNN-based models for IQMs that have been recently proposed.

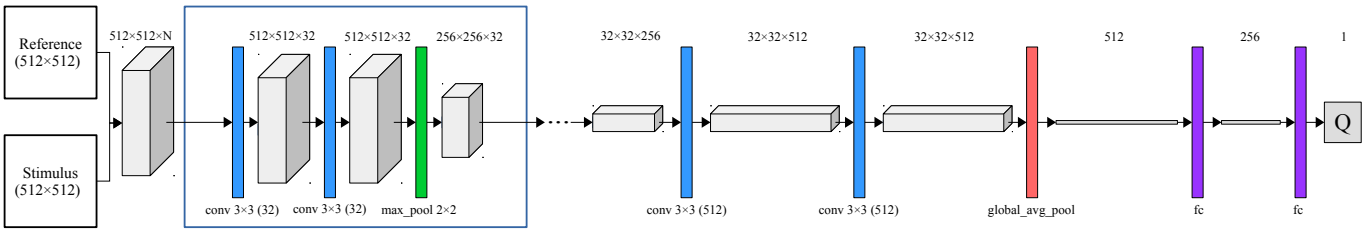


Fig. 1. The proposed DIQM architecture for computing the quality metric value Q for HDR-VDP (DIQM-Q). A sequence of convolutional layers is used to extract the image features, followed by a global average pooling and two fully connected layers that produce the final Q value for the HDR-VDP metric simulation. For the DRIM metric, the DIQM architecture (DIQM-P) changes only in the output layer, where we have 6 output nodes. Each pair of nodes output the number of pixels above the probability threshold of 75% and 95% for the three types of contrast distortions: A - contrast amplitude; L - contrast loss; R - contrast reversal.

Amirshahi et al. [20] introduced a FR image quality metric based on features extracted from CNNs. Using a pre-trained AlexNet model, they first extract feature maps of test and reference images at multiple layers and then compare their feature similarity at each layer. Finally, the similarity scores are pooled across layers to obtain an overall quality value.

More recently, Bosse et al. [21] proposed a network architecture that can be used both in NR and FR settings for IQMs. Their approach is purely data-driven and does not rely on hand-crafted features or other types of prior domain knowledge about the HVS or image statistics.

Hou et al. [22] investigated how to blindly evaluate the visual quality of an image by learning rules from linguistic descriptions. The qualitative evaluations are then converted into numerical scores to fairly benchmark objective IQMs. A discriminative deep learning model is trained to classify the features into five ordinal grades, corresponding to five explicit mental concepts: *excellent*, *good*, *fair*, *poor*, and *bad*. Finally, a quality pooling converts the qualitative labels into scores.

Two approaches to address FR [23] and NR [24] were proposed by Kim and Lee. The former uses CNNs to learn the HVS behavior from the underlying data distribution of IQMs databases. The latter approach tries to alleviate differences in quality between FR and NR approaches. The absence of ground truth in the online deployment of the NR model is solved by employing local quality maps derived by FR-IQMs as intermediate regression targets. This requires to pre-train the FR-IQM model with training datasets where the ground truth is available.

Liu et al. [25] applied a support vector regression approach to fuse scores obtained from multiple quality indices into one score. The approach is computationally expensive since it requires to compute multiple methods. These may be mitigated by reducing the number of quality indices. However, the method may fail when the input image has multiple distortions.

In a similar vein, Kang et al. [26] introduced an NR image quality metric based on CNNs. Within the network structure, feature learning and regression are integrated into one optimization process that leads to a more effective model for estimating image quality. To increase the size of the training dataset, input images are subdivided into 32×32 non-overlapping patches and labeled with a quality score. This may work well when distortions are distributed homogeneously in an image, but issues may arise when this hypothesis is not met (and this is expected to happen in many real cases).

Ye et al. [27] proposed a trained perceptually transform, similar in concept to PU or PQ EOTF, combined with PSNR for quality assessment of HDR images and video. ŋadík et al. [28] presented an analysis of feature descriptors for objective image quality assessment and proposed a data-driven FR-metric. Using this framework they optimized the parameters of popular existing metrics.

Recently, Kundu et al. [29] introduced HIGRADE, a NR metric for tone mapped images based on a large dataset of HDR images and their tone mapped versions. The method extracts different gradient-based features of the input image, which are processed by a support vector machine (SVM). This SVM is trained using a very large subjective experiment and outputs a scalar value that represents the perceptual quality of a tone mapped image.

III. Deep Image Quality Metric

A. Problem formulation and constraints

Given a pair of reference and distorted images as input, we train DIQM to produce two types of outcomes: (i) a prediction of the number of pixels above the probability threshold (probability index) of pixels changes, and (ii) to predict a single value that quantifies the quality of a processed (distorted) image with respect to the original (reference) image. Furthermore, we use this visibility information to detect structural changes as in [5].

In order to achieve this, we selected two popular existing visual objective metrics: HDR-VDP [4] and DRIM [5]. In principle, it would be possible to train a model to predict the spatially varying probability map of the per-pixel probability of distortions as generated by HDR-VDP and DRIM. However, preliminary experiments using fully convolutional networks proved that capturing fine-grained aspects, as those encountered in high frequency pixel areas of the ground-truth probability map, were difficult to reproduce; e.g., large saturated areas prevent the model to accurately infer structures and details.

It is very likely that this problem could be countered by sensibly increasing the number of training images.

Although the predicted map can still provide meaningful high-level understanding about distortions, its usage is often neglected in image quality estimation applications where a

unique parameter to define the overall quality of an image is preferred. There are two main reasons for this. First, its high cost (from the point of view of computational cost and human effort) might easily become intractable in standardization activities of large content datasets. Second, the evaluation of the distortion maps needs to be done manually for each image. This means the evaluation is tedious and error-prone since the probability of miss-interpreting the data is large.

HDR-VDP also outputs a value Q that quantifies the overall quality of the distorted image in terms of visibility [4]. Q can additionally be mapped to the *mean-opinion-score* (MOS) [4]. As described in [4], a single probability score for the entire image can be computed as the maximum value of the probability map P. HDR-VDP produces also further outputs, including the *threshold normalized contrast map* (C_{map}) and its maximum value. However, we argue these outputs are often not required when the main goal is to evaluate the overall quality of an image, in which case the Q value represents a more reliable choice.

Concerning the DRIM metric, we follow the same design principle adopted for the HDR-VDP. In this case, we train DIQM to estimate the probability index (as defined above) for the three types of contrast distortion: A – contrast amplitude, L – contrast loss, and R – contrast reversal. Note that our model can in principle be easily extended to produce additional outputs (such as TMQI[12] or HDR-VQM[30]) without any loss in generality; something we plan to investigate in future research.

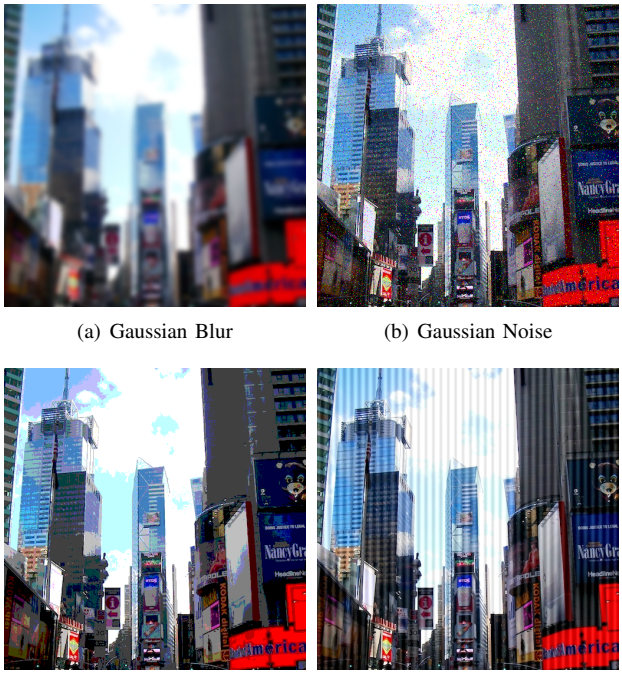
B. DIQM: Deep Image Quality Metric

We formalize the task of approximating the overall quality value Q and the probability indexes of DRIM as a regression problem.

As the model architecture, we adopted a Convolutional Neural Network (CNN) model, as this family of architectures is particularly envisioned for image related tasks [31]. Among the main advantages of CNNs in the realm of image processing, CNNs are extremely efficient because all intermediate steps are highly parallelizable. Among the broad set of variants of CNN architectures, we took the *U-Net* [32] as a starting point due to its proven success in tackling tasks similar to the one we are considering here.

U-Net is a CNN model initially introduced for image segmentation in biomedical contexts, and it is designed to make the most of the augmented training samples. We argue that solutions to the problem we are tackling here and fine-grained image segmentation are naturally interrelated as long as both are constrained to deal with similar aspects of the problem; i.e., the need to simultaneously deal with high- and low-level information of the image (commonly referred to as *context* and *location* in [32]).

We propose DIQM, a variant of the U-Net architecture, which is capable of predicting visual metric features with a high level of confidence. Loosely speaking, its architecture consists of an encoding path, which subsequently extracts higher-level features from the input images so as to embed the input images into a latent representation that (differently from the original *U-Net*) is then followed by a regressor path to produce the quality values.



(a) Gaussian Blur (b) Gaussian Noise

(c) Color Quantization (d) Sinusoidal Grate

Fig. 2. An example showing the types of distortions selected for the *Scenario* 2. The distortions in this figure are enhanced for visualization purposes.

The regressor consists of 5 downsampling blocks, each of which is composed by two 3×3 convolutions with *Rectified Linear Unit* (ReLU) activation functions followed by a global average pooling and two fully connected layers with 256 and 1 neurons, respectively; see Figure 1 (DIQM-Q). In the case of DRIM, the output layer consists of 6 neurons, representing the two probability indexes P_{75} and P_{95} for each of the three types of contrast changes (A, L, and R). The suffixes 75 and 95 indicate the threshold detection probability. This variant of our DIQM will be referred throughout the whole paper as DIQM-P.

The optimization procedure of DIQM is formalized as an iterative descent of gradients in the loss function quantifying the error in predictions. As an approximation to the true gradients (whose exact computation turns out to be infeasible due to hardware limitations), we use the mini-batch stochastic gradient descent (with the Adam update rule [33]). As the loss function for a batch of n examples, we use the *Mean Square Error* (MSE):

$$\mathcal{L}(Y, \hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{Y}_i - Y_i \right)^2, \tag{1}$$

which is known to be a good default choice (and one that has become the standard in scalar regression). Note that, in the case of HDR-VDP, $\hat{Y}_i$ and Y_i are single scalars, whereas for DRIM $\hat{Y}_i$ and Y_i are vectors of length 6.

(a) HDR Image (b) Drago et al. [34] (c) Durand and Dorsey [35]

(d) Kim and Kautz [36] (e) Mertens et al. [37] (f) Reinhard et al. [38]

Fig. 3. An example showing the different TMOs selected for the *Scenario* 4.

IV. Images Dataset

In order to set a testbed to fairly train and compare our DIQM against HDR-VDP and DRIM metrics, we generated different datasets of images to cover several relevant use cases involving both HDR and SDR content. Albeit DL techniques can excel in many tasks, they typically require large quantities of training data to do so. In order to fed our model with enough data, we selected an initial number of images from available datasets including HDR and SDR images. For each type of content (HDR and SDR), we generated 6 different datasets, each of which specialized in a different use-case scenario.

In the case of HDR content, the initial dataset was composed by 387 images extracted from various available datasets [39], [40], [41] covering a variety of dynamic ranges from indoors to outdoors, from photographs to computer-generated images, and frames from 30 different videos. Since consecutive frames in the video sequence are likely very similar, we extracted a frame every 4 seconds (i.e., we skipped 96 frames, this is nine times more than in previous work [42]) as a means to prevent almost identical images from being included in the dataset. We performed data augmented to enlarge this initial HDR dataset. We applied 3 types of rotations (90°, 180°, and 270°) and horizontal flipping (also applied to each rotated image). We thus apply a total of six transformations for each image, obtaining $2,709$ HDR images (including the original ones).

In the case of SDR content, we started from an initial set of $9,227$ images extracted from [43] (i.e., the subsets *SanMarco7K* and *TimeSquare6K*). Due to a large number of images available for the SDR, data augmentation was not applied in this case.

Both HDR and SDR images were downsampled to a manageable size of 512×512 in order to speed up the network training. Note that our network is fully-convolutional and, once trained, it can process images at lower and higher resolutions than 512×512; we show this possibility in Section VI.A; see Figure 11 and Figure 12. Furthermore, we conduct experiments on images of varying resolution (up to 16 Mpixel) in order to clock execution times; see Figure 19 and Figure 20.

Note that both HDR-VDP and DRIM require physical values to process a couple of images in a meaningful way. Therefore, we converted SDR images' values into physical ones using the specification of a standard SDR sRGB LCD monitor [44]. To this aim, we linearized pixel values using the inverse sRGB curve, and we scaled these values in the range of a standard SDR LCD monitor (i.e., [1, 250] cd/m^2 [44]).

We took these two datasets as a starting point to create a series of datasets representing 6 different use-case scenarios, as described in detail in the following subsections. Tables I and II summarize the sizes of the generated datasets.

A. High Dynamic Range Visual Differences Predictor (HDR-VDP)

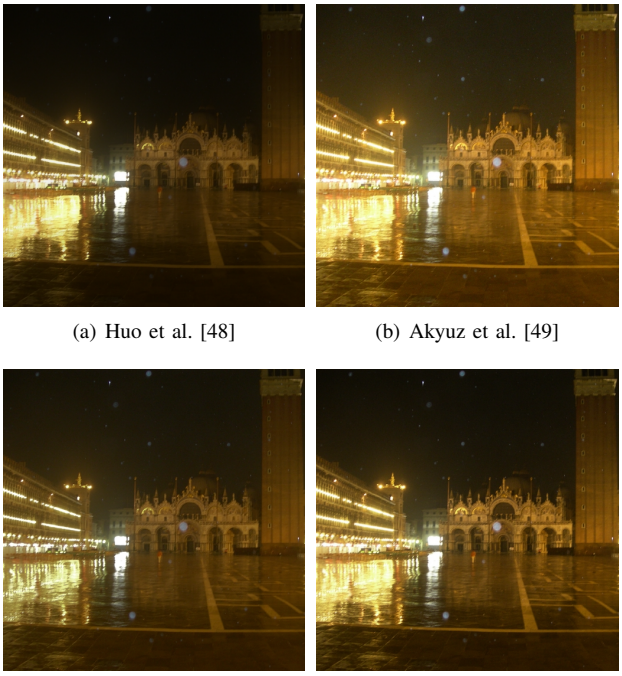
In this section, we define two different settings to test the Q score in representative scenarios, encompassing both HDR and SDR contents, for which HDR-VDP is well suited. The first scenario reproduces standardization in HDR content (Section IV-A1), while the second scenario is devoted to representing distortions in SDR content (Section IV-A2)

*1) **Scenario 1**:* In the last few years, academia and industry have been actively working on the definition of standards for the emerging HDR features added to several digital products. Valuable objective metrics have thenceforth become a fundamental tool for the standard evaluation. We used the recently proposed JPEG-XT standard coding system [2], [3] for still HDR images as a representative example.

Starting from the initial dataset of $2,709$ HDR images, we simulated the compression artifacts produced by JPEG-XT, leaving all encoding parameters set as specified in [2]. We decided to use a local TMO [38] since, as shown in [2], the compression capability of JPEG-XT is not strongly influenced by a specific tone mapping operator. The compression factor for the tone mapped image was fixed to 80, while the residual compression factor was varied from 1 to 100 at steps of 20 (we did not observe significant changes at smaller steps). We used all the three profiles available in the Part-7 of the standard, where each profile is selected randomly for all possible combinations.

*2) **Scenario 2**:* HDR-VDP can likewise be used to evaluate distortions for SDR content. Therefore, we selected four common types of distortions that are representative of various image processing tasks and randomly applied them to our set of SDR images. The distortions we considered are listed below:

- **Blur** - we selected Gaussian Blur distortion, where the sigma parameter is randomly chosen within the range $[0.5, 4]$.
- **Quantization** - we randomly applied two types of quantization distortion: JPEG compression with random quality values in the range $[10, 75]$, and bit reduction applied to each color channel of the input image by reducing encoding bits in the range $[2, 6]$.
- **Noise** - we used two types of noise: *Gaussian* noise and *salt-and-pepper* noise (or Impulse Noise). For the



(a) Huo et al. [48] (b) Akyuz et al. [49]

(c) Kovaleski and Oliveira [50] (d) Masia et al. [51]

Fig. 4. An example showing the different iTMOs selected for the *Scenario* 3. The inverse tone mapped images in this figure are divided by their maximum luminance value and gamma encoded for visualization purposes.

Gaussian noise, we considered the sigma parameter in the range $[0.0001, 0.005]$. For the salt-and-pepper noise, the intensity parameter is varied in the range $[0.001, 0.01]$. The type of noise to be applied and its parameters are decided randomly.

- **Sine gratings** - we randomly applied either a vertical or horizontal sine grates using a randomly selected intensity in the range $[0.005, 0.01]$ and a randomly selected frequency in the range $[0.008, 0.65]$.

An example of these distortions is depicted in Figure 2.

Typical image datasets used in the development of quality image metrics, such as TID2013[2] [45], Live IQA[3] Release 2 [46], and the ESPL Synthetic Image 2[4] could be used as training data for our model. However, as pointed out in the recent work of Kim et al. [47], they are too small for being used to train CNN-based models. We have thus decided to enlarge our original datasets by integrating the aforementioned datasets. These common datasets consist of 4,492 distorted images. In particular, the first dataset has 25 reference images at 512×384 resolution with 3,000 distorted images. The second dataset has 29 reference images at different resolutions (from 640×512 to 768×512) with 992 distorted images. Finally, the third dataset has 21 reference images at full HD resolution (1920×1080) with 500 distorted images. Note these images, which were larger than 512×512, were randomly cropped.

[2]http://www.ponomarenko.info/tid2013.htm
[3]www.live.ece.utexas.edu/research/quality/subjective.htm
[4]http://signal.ece.utexas.edu/~bevans/synthetic/

TABLE I
DATASETS GENERATED FOR FOR THE SCENARIOS WHERE THE HDR-VDP[4] METRIC IS USED.

	HDR-VDP[4]			
Scenario	Training	Evaluation	Test	Total
Scenario 1	12,768	1,596	1,638	16,002
Scenario 2	11,536	1,441	1,441	14,418

B. Dynamic Range Independent Metric (DRIM)

To evaluate the ability of our approach in approximating the contrast changes prediction of the DRIM, we defined four different scenarios that, also, in this case, include both HDR and SDR content. Given that DRIM is robust to differences in dynamic range, we distributed the scenarios of interest across two main categories: different dynamic range (Section IV-B1), and similar dynamic range (Section IV-B2).

*1) **Different Dynamic Range**:* We identified two possible scenarios where SDR content is evaluated against its corresponding HDR content:

- **Scenario 3** - This scenario takes into account the comparison of existing SDR content with respect to its expanded dynamic range version. We expanded the original SDR dataset used for the HDR-VDP simulation by randomly applying 4 inverse tone mapping operators (iTMOs) [49], [48], [50], [51]; we used the implementations of the HDR Toolbox [52] with the default parameters from the original papers. We set the maximum luminance as 3000 cd/m^2 (the typical output of an HDR display) in this case. Figure 4 shows an example of applying these operators.
- **Scenario 4** - This scenario covers the comparison of existing HDR content with respect to its tone mapped version. Taking the augmented HDR dataset as a starting point, we randomly applied 5 tone mapping operators (TMOs) [38], [35], [34], [36], [37] using the HDR Toolbox [52] implementations with the default parameters of the original papers. Figure 3 shows an example of applying these operators.

*2) **Similar Dynamic Range**:* Akin to the HDR-VDP simulation, we considered two scenarios, each of which meant to evaluate a different type of distortion in SDR and HDR contents. In this case, though, the images have a similar dynamic range. The scenarios we investigated include:

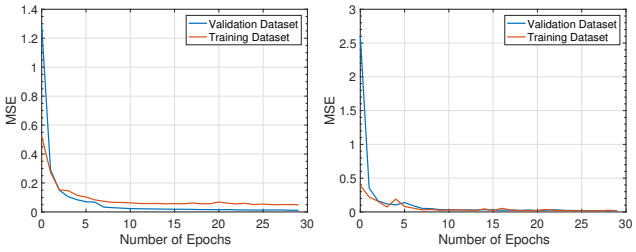
- **Scenario 5** - This scenario is analogous to *Scenario* 1 for the case of HDR-VDP, with the sole difference that the residual compression factor was varied from 1 to 100 at steps of 10 (unlike in *Scenario* 1, differences were noticeable at a smaller step).
- **Scenario 6** - In this scenario, we generated the probability indexes using DRIM in the same SDR content dataset and the same type of distortion described for *Scenario* 2.

V. TRAINING

In this section, we turn to describe the implementation details concerning the preprocessing of the network input and the training procedure. We implemented DIQM in Python using PyTorch as the deep-learning environment.

TABLE II
DATASETS GENERATED FOR THE SCENARIOS WHERE THE DRIM[5] METRIC IS USED.

	DRIM[5]			
Scenario	Training	Evaluation	Test	Total
Scenario 3	7,379	922	923	9,224
Scenario 4	2,128	266	273	2,667
Scenario 5	23,408	2,926	3,003	29,337
Scenario 6	11,536	1,441	1,441	14,418



(a) *Scenario 1* (b) *Scenario 2*

Fig. 5. An example of plots of the loss function over epochs for the training and the validation datasets for DIQM. (a) and (b) are plots for scenarios 1 and 2.

A. Input Preprocessing

We pre-process HDR and SDR contents differently. In the case of SDR images, we linearly scale the pixel values from the range $[0, 255]$ to $[0, 1]$ before feeding the net. For HDR images, we work on the logarithmic HDR pixel values x', obtained from the original values x as

$$x' = \log_{10}(x + 1). \tag{2}$$

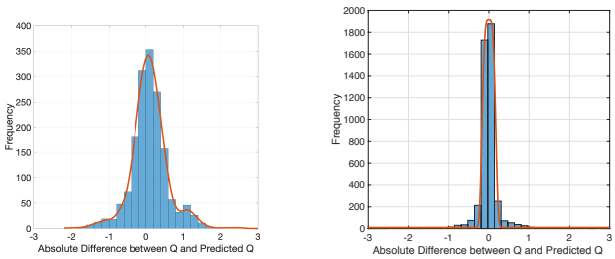
By doing so, we obtain an equilibrate scale in the positive only real values that is not biased towards large differences in high luminance values [42]. The reference and the distorted images are concatenated along the channel axis and given to the network as a unique input tensor.

B. Optimization Details

We initialized all network weights following the Xavier initialization [53]. We used stochastic optimization relying on the Adam optimizer [33] with learning rate 0.001 (leaving the rest of the parameters set to their default value; i.e., $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-8}$). The learning rate is decreased by a factor of 0.2 every time the loss function plateaus.

We set the batch size of DIQM to 32 samples, which was the largest parameters for which enough memory could be allocated in our NVIDIA GeForce GTX 1080 GPU. The training set was shuffled whenever an epoch is completed to diminish the impact of order-based biases during training. We set the maximum number of epochs to 30; the training time varies from approximately 6 hours to 3 days depending on the size of the training set.

In all cases, the final models for which results are reported to correspond to those obtaining the minimum loss as measured in a held-out validation set (described in details in Section VI).



(a) *Scenario 1* - Q (b) *Scenario 2* - *Q*

Fig. 6. Histograms for the prediction of the quality scalar value Q of HDR-VDP with the DIQM. The MSE for DIQM in *Scenario 1* is 0.144, while for *Scenario 2* is 0.275.

VI. RESULTS

In this section, we report the experimental results we obtained to validate DIQM. In particular, we trained our approach to predict the well known visual metrics HDR-VDP [4] and DRIM [5]. The aim here is to demonstrate the extent to which DIQM produces high-quality approximations of the quality factor (scalar Q) of HDR-VDP. Another goal is to show also the capability of DIQM to predict with high quality, the probability index of the DRIM contrast change maps. In both cases, we achieved significantly reduced computational cost when compared to one of the original metrics.

In Sections VI-A, we turn to describe the qualitative and quantitative evaluation of our framework in simulating HDR-VDP and DRIM metrics in 6 different scenarios that cover both HDR and SDR content. In Section VI-B, we show the evaluation process followed by tests of the computational performances of the proposed DIQM. Finally, in Section VI-C we show some possible applications where DIQM can be used to select the optimal parameters of an algorithm; e.g., TMO, iTMO, compression scheme, etc.

A. Learning Performances

1) HDR-VDP: Figure 5 shows the progress of the loss function as computed on the training and validation sets throughout the learning process. Figures 5(a) and 5(b) display examples (for Scenarios 1 and 2) of convergence trends, in which both the training and validation MSE smoothly approach zero within 30 epochs. This is also reflected on the results obtained on the test sets, depicted in Figure 6, for which DIQM-Q predictions differ only a few units from the ground-truth.

DIQM-Q model is capable to predict the scalar quality value Q of the ground truth HDR-VDP metric with high accuracy in both scenarios. Only for a few images the predicted Q value shows an absolute error greater than 1% in *Scenario* 1; see Figure 6(a). Furthermore, the absolute error is less than 1% for images in *Scenario* 2 that is far below the perceived difference by a standard observer; see Figure 6(b). This appealing feature of DIQM-Q is of the utmost importance in several applications where the interest resides in predicting the quality of an image through a unique value (e.g., image standards compression evaluation, imaging algorithms evaluation, iterative image quality improvement, image fusion,

(a) *Ground truth image* (b) *Distorted image* - $\sigma = 1.0$

(c) *Distorted image* - $\sigma = 2.5$ (d) *Distorted image* - $\sigma = 4.0$

Fig. 7. An example of different levels for the blur distortion applied to the ground truth image (a).

(a) *Ground truth image* (b) *Distorted image* - $q = 4$

(c) *Distorted image* - $q = 28$ (d) *Distorted image* - $q = 60$

Fig. 8. An example of different levels for the quantization distortion applied to the ground truth image (a). In this case, higher quantization levels mean better quality.

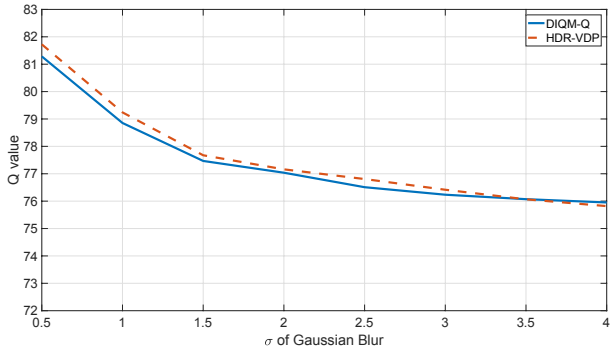


Fig. 9. Plot of the average over 5 images of the estimated Q values vs. the ground truth Q values computed using HDR-VDP varying the blur distortion using a σ value in $[0.5, 4.0]$.

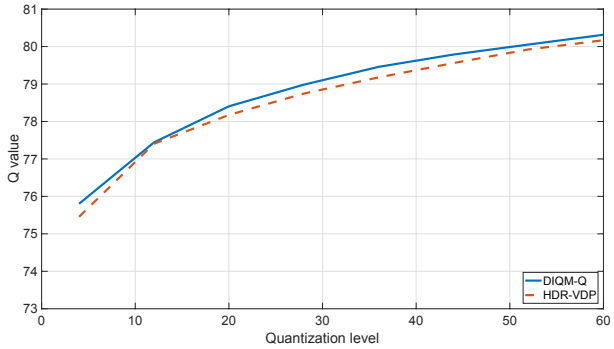


Fig. 10. The plot of the average over 5 images of the estimated Q values vs. the ground truth Q values computed using HDR-VDP varying the quantization levels with value in $[4, 60]$. Note that in this case quantization levels are directly proportional to quality.

etc.) while, at the same time, the computational time required to perform this prediction is real-time (42ms) for images at 512×512 resolution. As will be shown in Section VI-B, DIQM-Q significantly reduces the computational cost of HDR-VDP in predicting the Q value. This brings the opportunity to expand the use of HDR-VDP to more challenging scenarios for which only modest-sized datasets are currently affordable.

We also evaluated the capability of the DIQM-Q to be consistent with the level of degradation of the image. To perform this experiment, we selected a subset of images and applied different type of distortions at different level. Examples of blur and quantization distortion at different levels of degradation are shown in Figure 7 and Figure 8, respectively. Then we compared the predicted Q values of the DIQM-Q with the ground truth value of the HDR-VDP metric. The results are shown in Figure 9 and Figure 10 for the blur and quantization distortion, respectively. For all image and type of distortions, the plots show the consistency of the DIQM-Q in predicting the trend of the applied distortion with the trend of the ground truth Q value of the HDR-VDP metric.

Finally, we evaluated the capability of the DIQM-Q to be consistent with HDR-VDP when varying the image resolution. To perform this experiment, we used 30 high resolution images from a dataset [5] from DIV2K [54]. We scaled the original

[5]http://data.vision.ee.ethz.ch/cvl/DIV2K/DIV2K_valid_HR.zip visited in June 2019

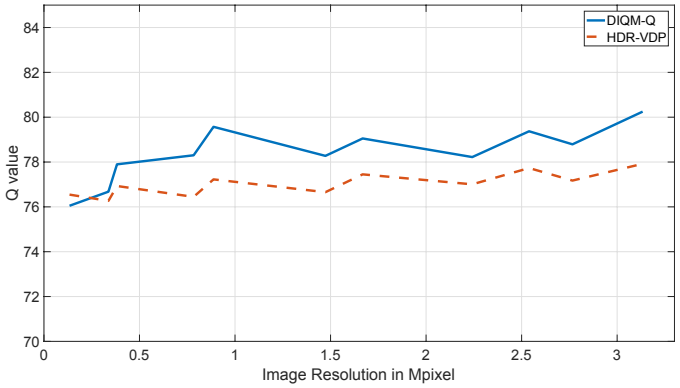


Fig. 11. Averaged Q values over 180 images as estimated by DIQM-Q vs. the ground truth Q values as computed by HDR-VDP at different image resolutions. DIQM-Q and HDR-VDP show comparable trends.

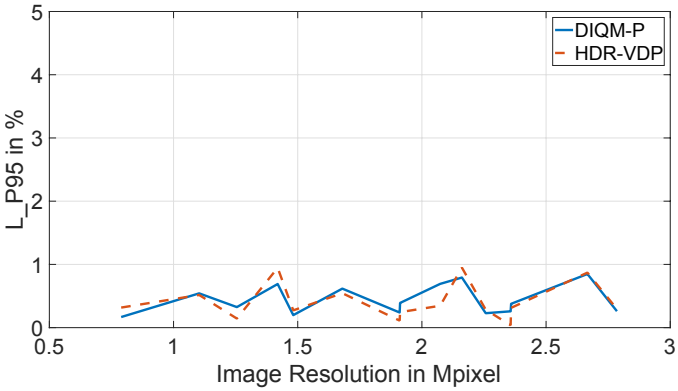


Fig. 12. Averaged L values at P_{95} over 102 images as estimated by DIQM-P vs. the ground truth computed by DRIM at different image resolutions. DIQM-P and DRIM show comparable trends.

images at different resolutions; i.e, from 3.3 Mpixels down to 0.3 Mpixels, obtaining 180 images. Then, we applied the quantization distortion with 8 levels for color channel to all the images generated. At this point, we ran both DIQM-Q and HDR-VDP on this set. This test reveals that both DIQM-Q and HDR-VDP follow similar trends when varying image resolution. DIQM-Q slightly overestimates HDR-VDP of 1-2 percentage points on average, which is a negligible error in practical applications. As an example, Figure 11 shows the average of Q values at different resolutions for both DIQM-Q and HDR-VDP; we can notice that both follow a similar trend. Similarly, we evaluated the capability of DIQM-P to be consistent with DRIM at variations of image resolution. We used 102 randomly picked images from the evaluation set of Scenario 4, and we tone mapped them using Kim and Kautz's TMO [36]. The resolution of these images ranges from 0.8 Mpixel to 2.8 Mpixel. By running both DIQM-P and DRIM on this set, we found out that even in this case DIQM-P and DRIM follow similar trends. Figure 12 shows the average of L values at P_{95} at different resolutions for both DIQM-P and DRIM.

2) DRIM: For DRIM, we opted for predicting the number of pixels of an image that are above the probability threshold of detecting a distortion (probability index) and cast aside

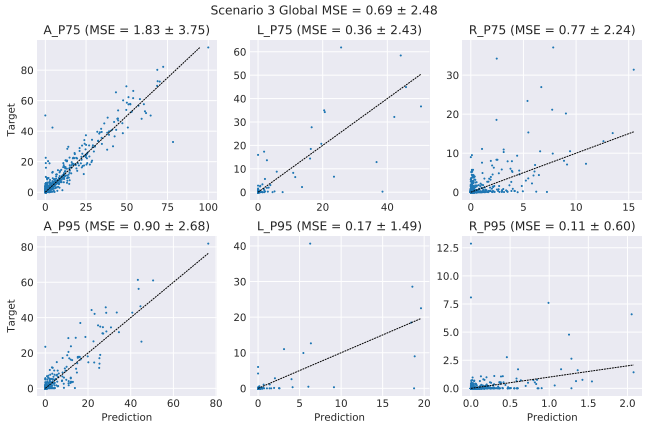


Fig. 13. Scatter plots for the DIQM predictions for scenario 3, for both probability thresholds: (top row) P_{75} - (bottom row) P_{95} and for all three types of contrast distortions detected by the DRIM (i.e., A - contrast amplitude; L - contrast loss; R - contrast reversal).

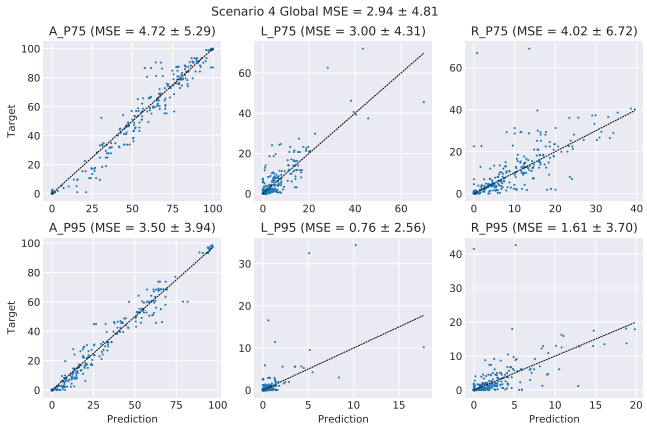


Fig. 14. Scatter plots for the DIQM predictions for scenario 4, for both probability thresholds: (top row) P_{75} - (bottom row) P_{95} and for all three types of contrast distortions detected by the DRIM (i.e., A - contrast amplitude; L - contrast loss; R - contrast reversal).

the idea of predicting the per-pixel probability maps for the reasons discussed in Section III-A. To show the capability of our approach in predicting this value, we considered two different probability thresholds 75% (P_{75}) and 95% (P_{95}), respectively.

We trained the DIQM-P independently for each of the 4 scenarios described in IV-B. Figures 13–16 show the correlations between the ground truth and the predicted values in the 4 scenarios. In all plots, the top row shows the results for the probability thresholds P_{75} and the bottom row the results for P_{95}, while each column shows the results for the three detected contrast changes A, L, and R.

In an additional experiment, we turn to test whether training one unique model on the union of the datasets concerning scenarios 3 to 5 (that share the same input range) instead of using separate datasets could have lead to some improvement. Figure 17 shows the DRIM correlations to the ground truth. Despite this leading to a unified model, which simultaneously can tackle problems coming from any of the three scenarios, the unification comes at a cost, as witnessed by the deterioration of the correlations. This comes as a surprise since a

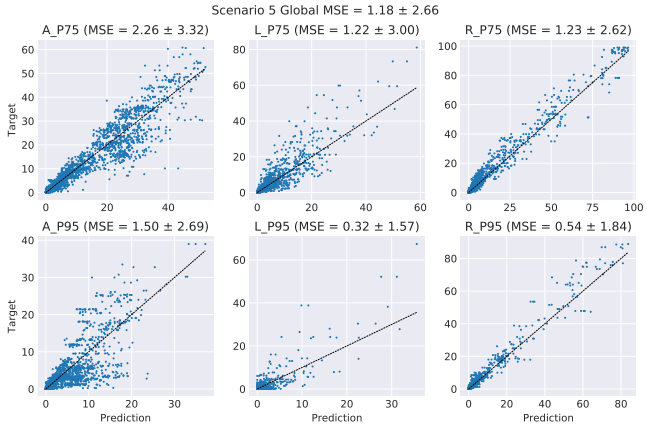


Fig. 15. Scatter plots for the DIQM-P predictions for scenario 5, for both probability thresholds: (top row) P_{75} - (bottom row) P_{95} and for all three types of contrast distortions detected by the DRIM (i.e., A - contrast amplitude; L - contrast loss; R - contrast reversal).

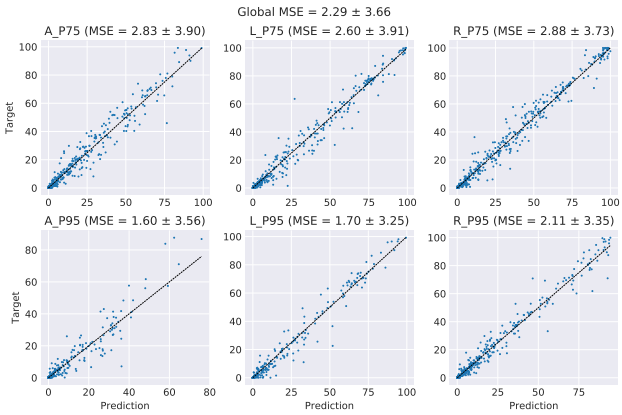


Fig. 16. Scatter plots for the DIQM-P predictions for scenario 6, for both probability thresholds: (top row) P_{75} - (bottom row) P_{95} and for all three types of contrast distortions detected by the DRIM (i.e., A - contrast amplitude; L - contrast loss; R - contrast reversal).

DL model trained on much more examples shall as expected deliver better performance. Likely, the reason for this failure can be explained by the fact that the unified model has now to divide its capacity (i.e., its parameters) in dealing with different types of distortions (which may not generalize well) while, at the same time, is constrained to understand which among the types of distortions is dealing with.

We also compared our DIQM-P (Scenario 4; i.e., tone mapped images) against HIGRADE [55] using its public available implementation[6] to study possible connections between reference and non-reference metrics. Unfortunately, we could not use the very large dataset provided by Kundu et al. [29] [7] because this dataset does not provide publicly available HDR images (note that DIQM-P does not generate probability maps from which MOS could be estimated).

We tried to find a correlation between the two metrics. As a first step, we tone mapped our dataset of 387 HDR images with

[6]http://users.ece.utexas.edu/~bevans/papers/2018/noreference/index.html visited in May 2019

[7]http://signal.ece.utexas.edu/~debarati/ESPL_LIVE_HDR_Database visited in May 2019

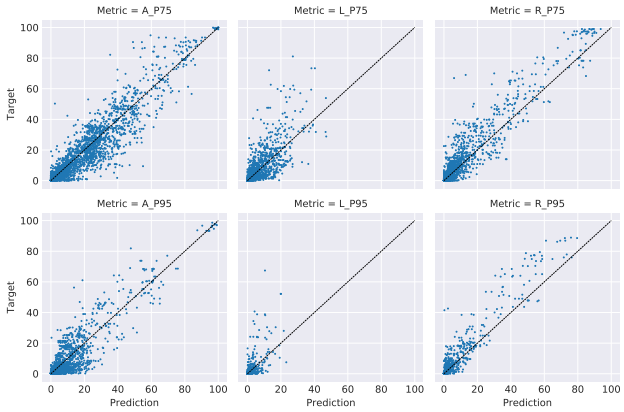


Fig. 17. Scatter plots for the DIQM-P predictions, training on all the data comprised from scenario 3 to scenario 6, for both probability thresholds: (top row) P_{75} - (bottom row) P_{95} and for all three types of contrast distortions detected by the DRIM (i.e., A - contrast amplitude; L - contrast loss; R - contrast reversal).

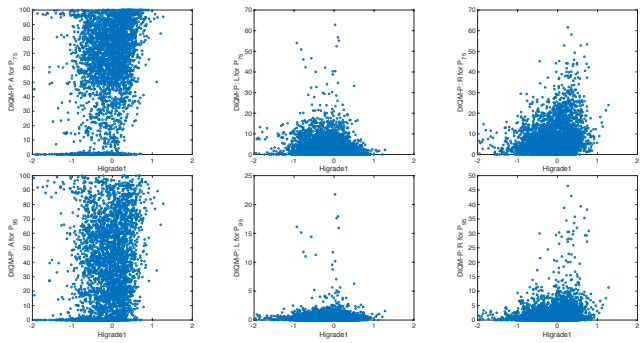


Fig. 18. Scatter plot between Higrade1 and DIQM-P for 3,378 tone-mapped images.

six tone mapping operators [35], [38], [56], [57], [58], [59] used by Kundu et al. [55], thus obtaining 2,322 tone mapped images. We then ran DIQM-P on each pair of HDR and tone mapped images. We also ran HIGRADE1 and HIGRADE2 models on the tone mapped images. Figure 18 reveals there is no evident correlation between DIQM-P and HIGRADE (we however found out that HIGRADE1 and HIGRADE2 are linearly correlated).

This may suggest that features extracted by DIQM-P and the original DRIM as *amplitude/reversal/loss of contrast* may not convey the same result of HIGRADE when evaluating the image quality of tone mapping images without a reference. This may be because the task varies from finding differences in contrast changes to no-reference image quality.

B. Timings

In order to assess the efficiency of the proposed method, we compared the computational time DIQM requires to deliver its predictions with the computational costs of the ground-truth HDR-VDP and DRIM implementations.

In these tests, we varied the size of the input images from QVGA to 16-Mpixel resolution. All experiments were run on a Linux machine equipped with an Intel CPU Core i7-7800X

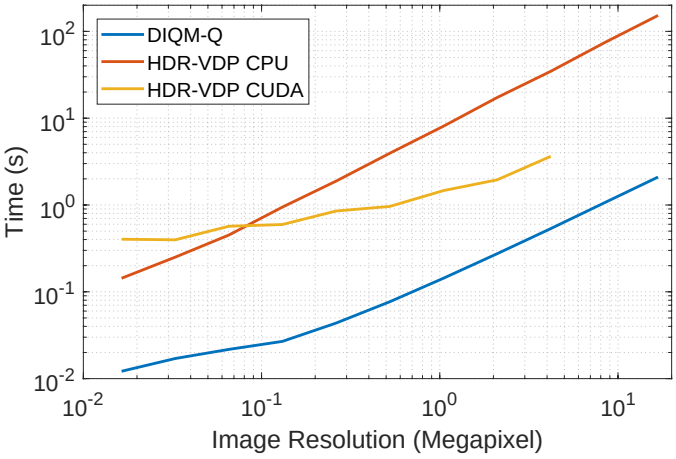


Fig. 19. Timings for the prediction of the quality value Q of DIQM with respect to the ground-truth HDR-VDP 2.2 [4] metric. The HDR-VDP CUDA version ran out of memory when processing images with a resolution equal or higher than 4-Mpixel. Note the log scale.

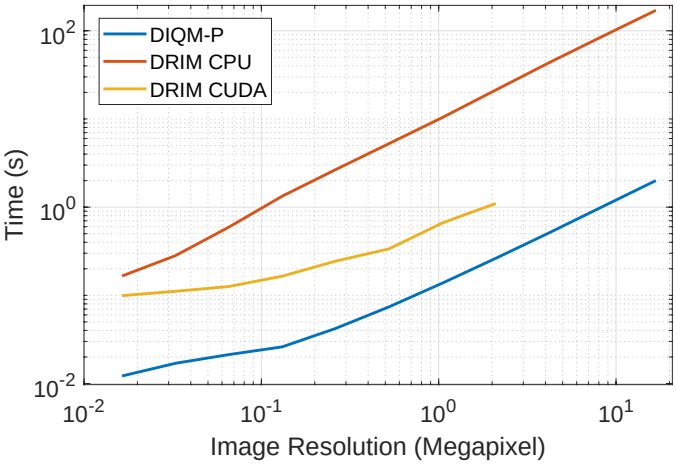


Fig. 20. Timings for the prediction of the probability values of DIQM with respect to the ground-truth DRIM [5] metric. The DRIM CUDA version ran out of memory when processing images with a resolution equal or higher than 2-Mpixel. Note the log scale.

(3.50 GHz) with 64 Gb of memory and an NVIDIA GeForce GTX 1080 GPU with 8 Gb of memory.

For HDR-VDP and DRIM metrics, we used the available MATLAB implementations provided by their authors. For the sake of fairness, we modified their code to exploit GPU acceleration using the highly optimized NVIDIA CUDA libraries [8] for convolutions and FFT (which account for most of the computational time).

As shown in Figure 19 and Figure 20, DIQM drastically reduces the computational costs of HDR-VDP and DRIM. For example, while the (original) CPU version of HDR-VDP takes 8.08s to process an image at 1024×1024 resolution, and its CUDA version takes 1.46s, DIQM requires only 0.14s to undertake the same task. This represents a $57\times$ speed-up compared to the original version and a $10\times$ speed-up compared to the CUDA version. Similar gains are obtained also for DRIM. Furthermore, DIQM can process images at much higher resolutions than the CUDA versions of HDR-VDP and DRIM, which ran out of memory.

[8]https://www.mathworks.com/solutions/gpu-computing.html visited in May 2019.

(a) (b)

Fig. 21. An example of a developed application, TMO Opt., for optimizing the parameters of a TMO (i.e., Reinhard et al. [38]): (a) An HDR image tone mapped using the default parameters of Reinhard et al.'s TMO [38]. (b) An HDR image tone mapped using optimized parameters using DIQM. Note that the look of this rendering looks more natural than (a).

C. Applications

In order to show the usefulness of DIQM, we designed and implemented three applications that exploits it. The first one (HDR Comp.) is a simple compression scheme, which reduces the file size of an HDR image based on a companding scheme (tone-mapping for compressing the signal and inverse tone mapping for expanding it) followed by a choice compression, e.g. JPEG. The application uses a perceptual metric (both HDR-VDP or DRIM) to determine the best parameters during the companding steps to achieve high quality. The second application (TMO Opt.) is a TMO based on Reinhard et al.'s TMO [38] in which parameters of the global tone curve (i.e., a and L_{white}) are optimized using DRIM for reducing all possible distortions that the metric can detect. The third application (iTMO Opt.) is an iTMO based on Masia et al.'s TMO [51] in which the gamma parameter is optimized using DRIM as for the second application.

Figure 21 shows a tone mapping result of TMO Opt; note that optimizing parameters using DIQM lead to more natural images than those obtained using the standard parameters of TMO. The parameters of Figure 21 were optimized in only 2.5 seconds using DIQM, while DRIM takes 150 seconds with the original MATLAB implementation and 29 seconds with the CUDA version.

We gained similar results for the other two applications. For example, HDR Comp. can compress high-quality 512×512 images with JPEG-XT in only 5.38 seconds instead of 151.83 seconds of the original implementation. Finally, iTMO Opt. can expand the dynamic range of SDR 512×512 images in only 5.23 seconds instead of 187.82 seconds of the original implementation.

VII. Conclusions and Future Works

Object visual metrics based on the HVS mechanisms provide a useful tool in the quality assessment of image/video processing techniques. However, their usage is precluded or

limited in certain relevant applicative scenarios due to their high computational costs. To overcome this problem, we have presented DIQM, a deep-learning-based objective metric which can predict scalar quality values comparable to those obtained by the traditional HDR-VDP and DRIM algorithms.

DIQM has been tested on a large dataset covering 6 different representative scenarios, including standardization and distortions, both for HDR and SDR contents. We have empirically demonstrated DIQM can predict the quality value Q and the probability index with high accuracy. DIQM is also significantly computationally cheaper, thus making it feasible to apply such visual metrics to scenarios that remained out of reach up to date.

Examples of applications that can benefit from the use of DIQM include optimization processes for selecting optimal parameters for TMOs, iTMOs, and JPEG-XT.

As future work, we plan to apply this approach in the domain of video content, where providing fast predictions is of the utmost importance. Typically, video metrics can be extremely cumbersome in terms of computational resources; e.g., 5 minutes for a video of 2 seconds [60]. Finally, we plan to investigate novel edge-aware network architectures such as the bilateral neural networks [61]. This may improve the quality of the predictions and might enable a reliable computation of per-pixel probability maps.

ACKNOWLEDGMENT

This work has been supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 739578 for Dr. Artusi and No 820434 (ENCORE) for Dr. Banterle respectively. The funding for the work of Dr. Artusi are also complemented by the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development. The work for Dr. Carrara was partially supported by "Automatic Data and documents Analysis to enhance human-based processes" (ADA), CUP CIPE D55F17000290009. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Tesla K40 GPU used for the early stage of this research.

REFERENCES

[1] A. Artusi, F. Banterle, T. O. Aydı n, D. Panozzo, and O. Sorkine-Hournung, *Image Content Retargeting: Maintaining Color, Tone, and Spatial Consistency*. CRC Press, september 2016.

[2] A. Artusi, R. Mantiuk, T. Richter, P. Hanhart, P. Korshunov, M. Agostinelli, A. Ten, and T. Ebrahimi, "Overview and evaluation of the JPEG XT HDR image compression standard," *Real Time Image Processing Journal*, 2015.

[3] A. Artusi, R. Mantiuk, T. Richter, P. Korshunov, P. Hanhart, T. Ebrahimi, and M. Agostinelli, "JPEG XT: A Compression Standard for HDR and WCG Images [Standards in a Nutshell]," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 118–124, 2016.

[4] 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," *Journal of Electronic Imaging*, vol. 24, no. 1, 2015.

[5] T. O. Aydın, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 69:1–69:10, Aug. 2008. [Online]. Available: http://doi.acm.org/10.1145/1360612.1360668

[6] H. Zhao, O. Gallo, I. Frosio, and J. kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.

[7] X. Zhang, D. A. Silverstein, J. E. Farrell, and B. A. Wandell, "Color image quality metric s-cielab and its application on halftone texture visibility," in *Proceedings of the 42Nd IEEE International Computer Conference*, ser. COMPCON '97. Washington, DC, USA: IEEE Computer Society, 1997, pp. 44–. [Online]. Available: http://dl.acm.org/citation.cfm?id=792770.793734

[8] H.-P. S. Tunç Ozan Aydı n, Rafal Mantiuk, "Extending quality metrics to full luminance range images," in *Proceedings of Human Vision and Electronic Imaging XIII*, vol. 6806, 2008. [Online]. Available: https://doi.org/10.1117/12.765095

[9] SMPTE, ""ST 2084:2014" - Society of Motion Picture and Television Engineers," 2014.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Trans. Img. Proc.*, vol. 13, no. 4, pp. 600–612, Apr. 2004. [Online]. Available: http://dx.doi.org/10.1109/TIP.2003.819861

[11] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 2, March 2005, pp. 573–576.

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

[13] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA, USA: MIT Press, 1993, pp. 179–206. [Online]. Available: http://dl.acm.org/citation.cfm?id=197765.197783

[14] T. Brandão and M. P. Queluz, "No-reference image quality assessment based on dct domain statistics," *Signal Process.*, vol. 88, no. 4, pp. 822–833, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2007.09.017

[15] M. A. Saad, A. C. Bovik, and C. Charrier, "A dct statistics-based blind image quality index," *IEEE Signal Processing Letters*, vol. 17, pp. 583–586, 2010.

[16] 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, Dec 2012.

[17] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *2009 International Workshop on Quality of Multimedia Experience*, July 2009, pp. 64–69.

[18] Z. Gu, L. Zhang, and H. Li, "Learning a blind image quality index based on visual saliency guided sampling and gabor filtering," in *ICIP*. IEEE, 2013, pp. 186–190.

[19] Zhou Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101)*, vol. 3, Sep. 2000, pp. 981–984 vol.3.

[20] A. Seyed Ali, P. Marius, and X. Y. Stella, "Image quality assessment by comparing cnn features between images," *Journal of Imaging Science and Technology*, vol. 60, no. 6, pp. 060 410:1–060 410:10, 2016.

[21] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Processing*, vol. 27, no. 1, pp. 206–219, 2018. [Online]. Available: https://doi.org/10.1109/TIP.2017.2760518

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

[23] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[24] ——, "Fully deep blind image quality predictor," *J. Sel. Topics Signal Processing*, vol. 11, no. 1, pp. 206–220, 2017.

[25] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Ieee transactions on image processing," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1793 – 1807, 12 2012.

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

[27] N. Ye, M. Pérez-Ortiz, and R. K. Mantiuk, "Trained perceptual transform for quality assessment of high dynamic range images and video," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 1718–1722.

[28] M. Čadík, R. Herzog, R. Mantiuk, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Learning to predict localized distortions in rendered images," *Computer Graphics Forum*, 2013.

[29] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "Large-scale crowdsourced study for tone-mapped HDR pictures," *IEEE Trans. Image Processing*, vol. 26, no. 10, pp. 4725–4740, 2017.
[30] M. Narwaria, M. P. 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.
[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.
[34] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," *Computer Graphics Forum*, vol. 22, no. 3, pp. 419–426, September 2003.
[35] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, Jul. 2002. [Online]. Available: http://doi.acm.org/10.1145/566654.566574
[36] M. H. Kim and J. Kautz, "Consistent tone reproduction," in *Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging*, ser. CGIM '08. Anaheim, CA, USA: ACTA Press, 2008, pp. 152–159. [Online]. Available: http://dl.acm.org/citation.cfm?id=1722302.1722332
[37] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion," in *Proceedings of the 15th Pacific Conference on Computer Graphics and Applications*, ser. PG '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 382–390. [Online]. Available: http://dx.doi.org/10.1109/PG.2007.23
[38] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, p. 267, Jul. 2002.
[39] M. Fairchild, "Fairchild's hdr photographic survey," 2008. [Online]. Available: http://rit-mcsl.org/fairchild/HDR.html
[40] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays," in *Proc. SPIE 9023, Digital Photography X, 90230X*, vol. 9023, 2014. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9023/90230X/Creating-cinematic-wide-gamut-HDR-video-for-the-evaluation-of/10.1117/12.2040003.short
[41] H. Nemoto, P. Korshunov, and T. Hanhart, Philippe ad Ebrahimi, "Visual attention in ldr and hdr images," in *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Chandler, Arizona, USA, February 5-6*, 2015. [Online]. Available: http://mmspg.epfl.ch/hdr-eye
[42] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "Hdr image reconstruction from a single exposure using deep cnns," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 178:1–178:15, Nov. 2017. [Online]. Available: http://doi.acm.org/10.1145/3130800.3130816
[43] S. Cao and N. Snavely, "Learning to match images in large-scale collections," in *Computer Vision - ECCV 2012. Workshops and Demonstrations - Florence, Italy, October 7-13, 2012, Proceedings*, 2012, pp. 259–270.
[44] R. Mantiuk, S. Daly, and L. Kerofsky, "Display adaptive tone mapping," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 68:1–68:10, Aug. 2008. [Online]. Available: http://doi.acm.org/10.1145/1360612.1360667
[45] N. Ponomarenko, O. Ieremeiev, V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. C. Kuo, "A new color image database tid2013: Innovations and results," in *15th International Conference on Advanced Concepts for Intelligent Vision Systems - Volume 8192*, ser. ACIVS 2013. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 402–413. [Online]. Available: https://doi.org/10.1007/978-3-319-02895-8_36
[46] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.
[47] J. Kim, H. Zeng, D. Ghadiyaram, . S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture quality prediction," *IEEE Signal Processing Magazine, Special Issue on Deep Learning for Visual Understanding*, vol. 34, pp. 131–141, 11 2017.
[48] Y. Huo, F. Yang, L. Dong, and V. Brost, "Physiological inverse tone mapping based on retina response," *The Visual Computer*, vol. 30, pp. 507–517, May 2014.
[49] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, "Do hdr displays support ldr content?: a psychophysical evaluation," *ACM Trans. Graph.*, vol. 26, no. 3, p. 38, 2007.
[50] R. P. Kovaleski and M. M. Oliveira, "High-quality reverse tone mapping for a wide range of exposures," in *Graphics, Patterns and Images (SIBGRAPI), 2014 27th SIBGRAPI Conference on*. IEEE, 2014, pp. 49–56.
[51] B. Masia, A. Serrano, and D. Gutierrez, "Dynamic range expansion based on image statistics," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 631–648, Jan 2017. [Online]. Available: https://doi.org/10.1007/s11042-015-3036-0
[52] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice (2nd Edition)*. Natick, MA, USA: AK Peters (CRC Press), July 2017.
[53] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.
[54] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
[55] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Processing*, vol. 26, no. 6, pp. 2957–2971, 2017.
[56] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 249–256, 2002.
[57] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 3, no. 4, pp. 291–306, 1997.
[58] S. Raman and S. Chaudhuri, "Bilateral Filter Based Compositing for Variable Exposure Photography," in *Eurographics 2009 - Short Papers*, P. Alliez and M. Magnor, Eds. The Eurographics Association, 2009.
[59] F. Pece and J. Kautz, "Bitmap movement detection: HDR for dynamic scenes," in *Visual Media Production (CVMP), 2010 Conference on*. IEEE, 2010, pp. 1–8.
[60] T. O. Aydı n, M. Čadík, K. Myszkowski, and H.-P. Seidel, "Video quality assessment for computer graphics applications," *ACM Trans. Graph.*, vol. 29, no. 6, pp. 161:1–161:12, Dec. 2010. [Online]. Available: http://doi.acm.org/10.1145/1882261.1866187
[61] V. Jampani, M. Kiefel, and P. V. Gehler, "Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Alessandro Artusi received a PhD in Computer Science from the Vienna University of Technology in 2004. He is the Team Leader of the DeepCamera MRG group at RISE, Research Center on Interactive Media, Smart Systems and Interactive Technologies, in Cyprus. His research interests include image/video processing, computer graphics, computer vision and color science, with particular focus to deploy the next generation of imaging/video pipeline. Contact him at artusialessandro4@gmail.com

Francesco Banterle is a full-time researcher at the Visual Computing Laboratory at ISTI-CNR, Pisa, Italy. He received a Ph.D. in Engineering from Warwick University in 2009 under prof. Alan Chalmers supervision. During his Ph.D. he developed Inverse Tone Mapping which bridges the gap between Low/Standard Dynamic Range Imaging and High Dynamic Range (HDR) Imaging. He also holds a B.Sc. and an M.Sc. in Computer Science from Verona University. He is the first author of the book "Advanced High Dynamic Range Imaging", a reference book for HDR imaging research, and co-author of the book "Image Content Retargeting". His main research interests are in the field of HDR imaging, Computer Graphics (rendering and image-based lighting), Computer Vision, and recently the application of Imaging to Deep Learning.

Fabio Carrara graduated in Computer Engineering in 2015 and received the Ph.D. in Computer Engineering in 2019 at the University of Pisa (Italy). Since 2015, he is a graduate research fellow at the Networked Multimedia Information System Laboratory (NeMIS) of the Information Science and Technologies Institute (ISTI) of the National Research Council (CNR) in Pisa, Italy, and a member of the Artificial Intelligence and Multimedia Information Retrieval (AIMIR) group. His works cover deep learning for multimedia data with a focus on visual perception, image classification, content-based and cross-media image retrieval, analysis and detection of visual adversarial examples.

Alejandro Moreo Fernández received a PhD in Computer Sciences and Information Technologies from the University of Granada in 2013. He is a researcher at the Networked Multimedia Information System Laboratory (NeMIS) of the Information Science and Technologies Institute (ISTI) of the National Research Council (CNR) in Pisa, Italy, and a member of the Text Learning group. His research interests include deep learning and representation learning, with particular focus on word embeddings and transfer learning for text classification. Contact him at alejandro.moreo@isti.cnr.it