Received October 31, 2020, accepted November 14, 2020, date of publication November 25, 2020,
date of current version December 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040416

End-to-End Image Patch Quality Assessment for Image/Video With Compression Artifacts

TUNG THANH PHAM[1], XIEM VAN HOANG[2], NGHIA TRUNG NGUYEN[3], DUONG TRIEU DINH[2], AND LE THANH HA[4]

[1]Faculty of Fundamental Sciences and Foreign Languages, University of Fire Prevention and Fighting, Hanoi 11415, Vietnam
[2]Faculty of Electronics and Telecommunications, University of Engineering and Technology, Vietnam National University Hanoi, Hanoi 123105, Vietnam
[3]Medical Imaging Department, Vingroup Big Data Institute, Hanoi 11622, Vietnam
[4]Faculty of Information Technology, University of Engineering and Technology, Vietnam National University Hanoi, Hanoi 123105, Vietnam

Corresponding author: Tung Thanh Pham (tungpt@daihocpccc.edu.vn)

ABSTRACT In this paper, we present an experimental image quality assessment (IQA) method for image/video patches with compression artifacts. Using the High Efficiency Video Coding (HEVC) standard, we create a new database of image patches with compression artifacts. Then, we conduct a completed subjective testing process to obtain the 'ground truth' quality scores for the mentioned database. Finally, we employ an end-to-end learning method to estimate the IQA model for the patches with HEVC compression artifacts. In such proposed method, a modified convolutional neural network (CNN) architecture is exploited for feature extraction while an adaptive moment estimation optimizer solution is used to perform the training process. Experimental results show that the proposed end-to-end IQA method significantly outperforms the relevant IQA benchmarks, especially when the compression artifacts are strongly realized in image/video patches. The proposed IQA method is expected to drive a new set of image/video compression solutions in future image/video coding and transmissions.

INDEX TERMS Image quality assessment, coding distortion, image-patch quality assessment, compression artifacts.

I. INTRODUCTION

Currently, image quality assessment (IQA) has been playing a critical part in image and video communications. IQA is a basic and important requirement in encoding images and videos [1]. Generally, either subjective or objective methods can be used to evaluate the quality of the image.

Subjective assessment methods are highly effective, but they can be infeasible in conducting the assessment in real time and on large scale. It requires the engagement of a number of human viewers who will give their views on image/video quality under a variety of test conditions. Thus, it is necessary that testing conditions be closely monitored, with careful selection of observers and processing of the findings to ensure their consistency and statistical significance. Consequently, they are costly and time consuming.

Unlike the subjective method, objective quality assessment adopts criteria that attempt to imitate the ability to perceive images via the human visual system (HVS). In some conventional methods, the absolute or squared difference between distorted images and their original ones is utilized.

The associate editor coordinating the review of this manuscript and approving it for publication was Jinjia Zhou.

The traditional methods of image compression mainly use the quality metrics based on signal-fidelity, which poorly correlate with the quality perceived by humans, such as MAE (mean absolute error), MSE (mean square error), PSNR (peak SNR) and their inheritances [2]. While these metrics have a lot of positive features, e.g. clear physical meaning and high calculation efficiency, they create adverse impact on the efficiency of compression as they fail to exclude image visual redundancies which are inconsistent with human visual perception.

A number of perceptual quality metrics have been introduced in recent years to obtain measures more consistent with human visual perception. One class of these algorithms including SSIM [3], FSIM [4], RFSIM [5] with an application of handcrafted features that supposedly capture relevant factors affecting to image quality. Although they are widely accepted and applied, the accuracy with which they reproduce human perception of image quality need to be enhanced.

Another set of methods adopt convolutional neural network (CNN) based approaches [6], [7]. In this approach, some features are extracted from the original image's pixels, which are automatically learnt and embedded within the network. Some available image quality databases have

TABLE 1. **Comparison characteristics of subjective Image Databases.**

Database	Year	Type	Data	Scores	SRC	LOD	NOD	Subjects	Ratings	Resolution	Method	PSNR
CSIQ [11]	2010	Full	DMOS+σ	866	30	5	6	25	5-7	512×512	Custom	
IVC(I) [12, 20]	2005	Full	DMOS+σ	185	10	5	3	15	15	512×512	DSIS	65%
IVC-3D [21]	2008	Full	DMOS	90	6	5	3	19	19	512×512	SAMVIQ	
IVC-Art [22, 23]	2009	Full	Raw	120	8	5	3	19	19	512×512	DSIS	
LIVE(I) [8, 17]	2006	Full	DMOS+σ	779	29	7-8	5		20-29	768×512	ACR	88%
MICT [13]	2008	Full	Raw	196	14	5	2	16	16	768×512	ACR	61%
MMSP-3D(I) [24]	2010	Full	MOS+σ	54	9	4	1	27	27	768×512	ACR	62%
TID2008 [9, 25]	2008	Full	Raw	1700	25	4	17	838	17	512×384	ACR	55%
TID2013 [10, 18]	2013	Full	MOS+σ	3000	25	4	10	971	33	512×384	ACR	
PDAP-HDDS [19]	2018	Full	MOS	12000	250	4	24	38	30	FHD	ACR	54%
HMII (Proposed)	**2018**	**Patch**	**MOS**	**40286**	**308**	**49**	**1**	**2189**	**15-20**	**HD, FHD**	**DSIS**	**65-67%**
	Scores		Number of testing images.									
	SRC		Number of source (reference) images.									
	LOD		Levels of distortions.									
	NOD		Number of distortion types.									
	PSNR		Approximate correlation between PSNR and MOS.									

been introduced in literature, including LIVE Image [8], TID2008 [9], TID2013 [10], CSIQ [11], IVC [12] and MICT [13]. Generally, these methods estimate the quality of the image patches and propose a most apparent artifact IQA metric which only well perform with desirable results on some typical image/video databases. Their architecture is more suitable for evaluating the quality of block size image in rate-distortion optimization manner which share a block-based common hybrid coding framework. However, their IQA metrics which do not synthesize the video compression features fail to compute the quality of block size video frame. Jin, in [14], has stated that there is no comprehensive method which is exactly comparable to human perception and can be equally well applied in different areas. Therefore, a subjective rating database of image/video patches with compression artifacts is necessary in constituting ground truth needed for training, testing, and benchmarking in video coding.

Based on these observations, we propose a large-scale Image-Patch Quality Assessment database with video compression distortion in this paper. State-of-the-art IQA methods on the proposed database are analysed and it can be seen that IQA's accuracy can be improved in predicting image quality. Finally, we propose a full-reference (FR) Image-Patch model to determine image patches' quality based on the CNN architecture.

In summary, this paper includes the following contributions:

1) A COMPRESSED IMAGE-PATCH DATABASE

To our best knowledge, this database is the first one to be constructed, serving as a benchmark for assessing compressed image and the quality of video's frame patch, and being beneficial for image and video compression on the basis of human perception. The existing databases with coarse-grained quality are inefficient to evaluate IQA algorithms especially patch-based methods on images with fine-grained quality differences. One of the problems in perceptual-based image and video compression is to select the optimal coding mode for each coding block according to their rate-distortion. Therefore, the proposed database can be helpful for researchers in the field of image compression to select the most appropriate IQA method to gain the perceptual based image optimization.

2) A DEEP IMAGE-PATCH NEURAL NETWORK DESIGN

We also investigate different FR methods to model the relationship between image patch and patch quality score. After multiple experiments, Deep Image-Patch Quality Assessment is proposed to address the problem of end-to-end optimization. We use the adjusted concept of Siamese networks mentioned in tasks of [15], [16] to extract features from the reference and distorted patches based on a deep convolution neural network.

This paper is organized as follows: The related studies on IQA databases and Deep Learning IQA Methods are reviewed in Section II. Then, Image/video patches with compression artifacts created and the proposed method are described in Section III. Extensive experimental results are presented in Section IV. Finally, conclusions are given in Section V.

II. RELATED WORKS

A. IQA DATABASES

A brief overview on test material and experimental details of existing databases is presented in this part. As can be seen from Table 1, experimental data are generated from original images (6-30). Various distortions are added to these images at different levels to form testing images whose quality is assessed via subjective rating by a certain number of observers (usually from 15-30). The testing methods frequently used are "Double stimulus categorical rating" (DCIS) and "Single stimulus categorical rating" (e.g., Absolute Category Rating (ACR). The assessment scores used are differential mean opinion score (DMOS) or mean opinion score (MOS) in combination with standard deviation. Among those databases, LIVE [8], CSIQ [11], TID2008 [9], TID2013 [10] are widely utilized in bench-marking, testing and training of IQA. Releasing 2 of LIVE [8], [17] consists of 779 testing images generated from 29 original images which are added with five different distortions. They are JPEG, JPEG2000, white noise, Gaussian blur, and simulated Rayleigh fading channel (JPEG2000 bitstream) of which each comprises 7-8 levels. As for Categorical Subjective Image Quality (CSIQ) Database [11], there are 30 original images, of which each is processed with the application of six types of distortions including JPEG compression, JPEG-2000 compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blurring. For each type of disortion, four to five different levels are used. The result is that 866 distorted images are produced. With The Tampere Image Database (TID 2008), 1700 distorted images are produced from 25 reference images using 17 disortion types at 4 levels of degradation for each distortion. TID 2013 [10], [18] extended from TID 2008 comprises 3000 distorted images by using 24 distortions instead of 17. Currently this image quality database has the largest number of both testing images and subjects in the public domain. Liu *et al.* [19] recently introduced the PDAP-HDDS image quality database including 2,000 high-definition resolution test images. This database consists of a total of 12000 MOS calculated from 360,000 opinions subject rating by 38 observers. Due to the fact that limited distortion levels (4-8 ones) covering the whole quality range from "Bad" to "Excellent" are available in most of the existing IQA databases, it is obviously different and easy to rank the images in adjacent distortion levels. These databases contain some distortion types that do not occur during modern image or video compression.

All existing databases evaluate subjective quality for full image while the quality in its regions is far different. Because of compressing or adding noise to the image, each region has its own characteristics resulting in annoying artifacts with which HVS has different sensitivity. Moreover, there is a minimum visibility threshold which no change can be perceived below [26]. Figure 1 shows that distortions between the houses and the sky regions (1) and the edge region (5) are easily observable. However, those on flat region (2) and textural regions (3, 4) are less noticeable. In addition, HVS has the ability to effortlessly identify salient objects even in a complex scene by exploiting the inherent visual attention mechanism. As stated in [27], [28], many physiological experiments have proved that only the significant portion of the scene projected onto the retina is thoroughly processed by human brain for semantic understanding. Therefore, it is inappropriate to take subjective rating of the whole image for all areas in the image.

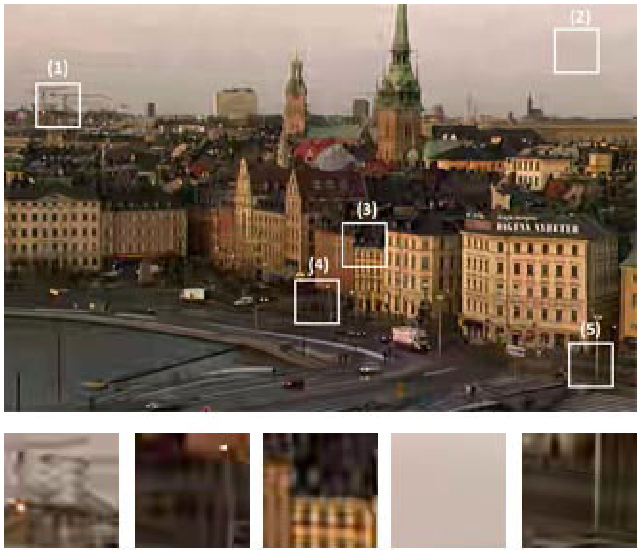


FIGURE 1. Example of distorted image.

B. DEEP LEARNING IQA METHODS

In Deep learning IQA methods, a number of patches from images are usually selected, then fed into a CNN model depending on the distortion type and distortion level information. Subsequently, the CNN model extracts features from each selected patch and evaluates its quality score. Finally, all scores of patches are weighted to obtain the quality of the image. Bosse *et al.* [6] justrandomly selected patches and acquired the image quality in two ways: simply averaging and learned weights of the patches' score. Li and Yue [29] used the idea of visual saliency to calculate the weight of each patch in an image and selected top weighting patches. Kim and Lee [30] pre-trained a CNN model using numerous patches with proxy quality scores provided by a full-reference IQA model. Recently, Zhang *et al.* [31] proposed a deep bilinear model for blind image quality assessment (BIQA) that handled both synthetic and authentic distortions. Last but not least, in [32] the authors proposed a multi-task CNN to predict the type of distortions and image quality from the last fully connected layer in the network. In summary, the afore-mentioned methods partially address the patch-based training data shortage problem, but it is difficult to extend them to ensure the subjectivity of original databases. To overcome that weakness, Wu *et al.* [33] deployed a new large-scale training dataset (including 80,000 labeled images using advanced FR-IQA metric) to develop a novel no-reference (NR) model for accessing the perceptual quality of screen content pictures.

Generally, image quality assessments (IQAs) are divided into three types, namely full-reference (FR), no-reference (NR) and reduced-reference (RR). Full-reference (FR) approaches can fully access reference images, but no-reference (NR) approaches only use distorted images. The majority of aforementioned deep learning based methods are included in NR type with the exception of the method proposed in [6] which belongs to FR IQA type. No-reference approaches yield the good performance of cross test in the training database, albeit showing poor results when being tested in others [6], [31].

The SSIM [3] can be considered as the most popular perceptual approach in FR IQA. It is determined by pooling luminance similarity, contrast similarity and structural similarity. The SSIM is not only developed into MS-SSIM

TABLE 2. Testing sequences.

No	Test sequence	Resolution	Frames	SF	SP	No	Test sequence	Resolution	Frames	SF	SP
1	Mobcal	1280 × 720	500	6	1200	21	Honey Bee	1920 × 1080	600	8	1600
2	Parkrun	1280 × 720	500	6	1200	22	Jockey	1920 × 1080	600	8	1600
3	Fhields	1280 × 720	500	6	1200	23	Old town cross	1920 × 1080	500	8	1600
4	Stockholm	1280 × 720	500	8	1600	24	Park joy	1920 × 1080	500	8	1600
5	Vidyo 1	1280 × 720	600	8	1600	25	Pedestrian area	1920 × 1080	375	6	1200
6	Vidyo 3	1280 × 720	600	8	1600	26	Ready Steady Go	1920 × 1080	600	8	1600
7	Vidyo 4	1280 × 720	600	8	1600	27	Red kayak	1920 × 1080	600	8	1600
8	Four People	1280 × 720	600	8	1600	28	River bed	1920 × 1080	250	5	1000
9	Johnny	1280 × 720	600	8	1600	29	Rush hour	1920 × 1080	500	6	1200
10	Kristen And Sara	1280 × 720	600	8	1600	30	Shake and Dry	1920 × 1080	300	5	1000
11	Aspen	1920 × 1080	570	8	1600	31	Sintel trailer	1920 × 1080	1250	10	2000
12	Basketball Drive	1920 × 1080	500	6	1200	32	Snow mountain	1920 × 1080	600	8	1600
13	Beauty	1920 × 1080	600	8	1600	33	Speed bag	1920 × 1080	600	8	1600
14	Big Buck Bunny	1920 × 1080	14315	14	2800	34	Station 2	1920 × 1080	313	5	1000
15	Blue sky	1920 × 1080	217	5	1000	35	Sun flower	1920 × 1080	500	6	1200
16	Bosphorus	1920 × 1080	600	8	1600	36	Tractor	1920 × 1080	690	9	1800
17	Controlled burn	1920 × 1080	570	8	1600	37	West wind easy	1920 × 1080	600	8	1600
18	Crowd run	1920 × 1080	500	8	1600	38	YachtRide	1920 × 1080	600	8	1600
19	Dinner	1920 × 1080	950	10	2000	39	Ducks take off	1920 × 1080	500	6	1200
20	Elephants Dream	1920 × 1080	15691	15	3000	40	In to tree	1920 × 1080	500	6	1200
		SF: Selected Frames SP: Selected Patches					**All**			**308**	**61600**

metric [34], but also be applied for FR IQA, such as the FSIM [4], the SRSIM [35] or RFSIM [5] obtaining promising results. However, testing conducted in [6] indicates that WaDIQaM-FR method shows better performance compared with the above-mentioned methods. Other FR IQAs approaches as in [36] use an adaptive representation of local patch structure yielding rewarding results, but they can only be applied to certain distortion types. Therefore, this study is only intended for comparison of performance of full-reference (FR IQA) approaches in [6].

In a recent work [37], we propose a quality assessment approach database for image patch with the desire to create a new perception-based metric to apply for each region. Then, a coding distortion modelling method for local image perception which is able to predict objective evaluation from the perceptual point of local image content is presented. Experimental results show that compressed image quality decreases depending on the visual features of image. However, the testing image database only has 600 samples, so it is not enough to cover all features of human visual perception.

III. PROPOSED IMAGE PATCH QUALITY ASSESSMENT

Given the necessity of an efficient IQA method for image/video patches with compression artifacts, we present in this section a novel end-to-end IQA method. After discussing the characteristics of image/video patches with compression artifacts, we introduce the architecture of the proposed method. Finally, we present the feature engineering and training optimization.

A. IMAGE/VIDEO PATCHES WITH COMPRESSION ARTIFACTS

All available image quality benchmark databases are only suitable for evaluating the quality of images as a whole and not able to investigate which part of the testing image contributing to the testing results or to the score of a particular patch of image. In this work, we set up an experimental database to evaluate the quality that human perceive for each image patch. The testing data are preprocessed to remove noises and outliers.

1) GROUND TRUTH SCORE ACHIEVEMENT

The goal of our study is to create a testing image database for local image perception. Due to the research orientation for video coding, testing images are extracted from the video test sequence and noise types are added to the original video by H.265/HEVC compression before extracting. There are 40 original source videos of high-definition (1280×720) and full high-definition (1920×1080) being compressed with different quantization parameters (QPs) in range from 2 to 50. For each video sequence, depending on the length of such video, a different number of original frames are selected as reference images, as summarized in Table 2. The reference frames are selected evenly throughout the video sequence to diversify the content. For each reference image, 200 pairs of 128×128 patches are randomly selected by position to crop and by quantization parameters in a frame of the compressed video which has the same index with the aforesaid reference image. We also crop the center 64×64 patches from the original pair 128×128 to evaluate in the experiments. Finally, we obtain $246,400$ images: $61,600$ pairs of 64×64 patches and $61,600$ pairs of 128×128 patches. All patches are annotated with their position.

2) TESTING METHODOLOGY

Observers may be experts or non-experts depending on specific requirements of the test. Studies have found that systematic differences may occur among different laboratories conducting similar tests [38]. One of the reasons for this is that expert observers have different view in compare with non-experts. Other factors may include gender, age, and occupation. However, the majority of consumers in reality are non-experts; hence, for the purpose of this study non-expert observers are recruited. Before the final selection round, all candidates have been checked to ensure that they possess normal visual acuity (color vision included). In [38], it is recommended that observers should not be directly involved in image quality evaluation and should not be experienced assessors. In this study, more than 2000 people (2159- undergraduates, 20-graduates, 3-researchers,7 -lecturers) are employed.

For the purpose of subject testing methodology, the International Telecommunication Union set the ITU-R BT.500-11 standard. In this standard, there are several popular subjective methodologies for testing such as ''Single stimulus categorical rating'', ''Double stimulus categorical rating'', ''Ordering by force-choice pairwise comparison'' and ''Pairwise similarity judgments''. Double stimulus categorical rating is chosen in this test. In this method, both the testing and reference images, which randomly appear, are viewed by observers for a fixed period of time. After that, the observers are asked to vote for the quality of the testing image in accordance with the scale of five categories: ''excellent'', ''good'', ''fair'', ''poor'' or ''bad''. Prior to each session, the observers have been instructed about assessment modes, assessment scale and the procedures (reference image, grey, test image, voting period).

Since the image quality assessment methods stated in [38] are only suitable for assessing quality of image as a whole, they cannot be directly applied for our testing experiments. Therefore, we modify this image selection method into the standard so that the users can focus and only assess the local image patch instead of the whole image. Figure 3 represents the created testing software in the experiment. Each pair quality is assessed following the procedure of 2. The subjects observe the original image within the time T1 at minimum 5s, then click on the observing image patch to observe the compressed image within the time T2. After watching at least twice per image, observers are requested to score on the scale of 5.

3) DATA PREPROCESSING

In our experiment, 2189 different subjects rate 61600 image patch pairs of which each p^{th} is observed by S_p (up to 20) subjects. The differential mean opinion score (DMOS) of each patch pair is calculated by:

$$\overline{y_p} = \frac{1}{S_p} \sum_{s=p}^{S_p} y_{p,s}, \tag{1}$$

where $y_{p,s}$ is the differential opinion score of a subjective rating by subject s for patch pair p^{th}. Let Y_o denote the raw data and each image patch pair (R_p, D_p) is evaluated by at least 15 observers as folow:

$$Y_o = \left\{((R_p, D_p), \overline{y_p}) | S_p \geq 15\right\}. \tag{2}$$

FIGURE 2. Testing software.

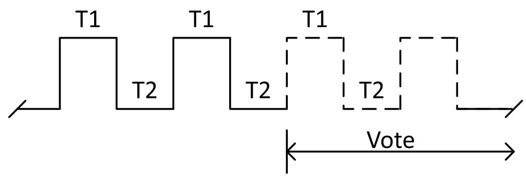


FIGURE 3. Presentation structure of test material.

The raw score database is not entirely good because some observers evaluate carelessly. To remove outlier in this data, we use z-score. The z-score of a subjective rating for patch pair p^{th} is calculated by the following formula:

$$Z_{p,s} = \frac{y_{p,s} - \mu_p}{\sigma_p}, \tag{3}$$

where μ_p denotes the mean and σ_p denotes the standard deviation of rating pairs. The figure below (Fig.4) shows that distribution of z-score is the standard normal distribution side-by-side. According to empirical rule, 95%, 98.7% and 99.7% of the values lie within 2, 2.5 and 3σ, respectively. After applying this rule we achieve the results as in table 4.

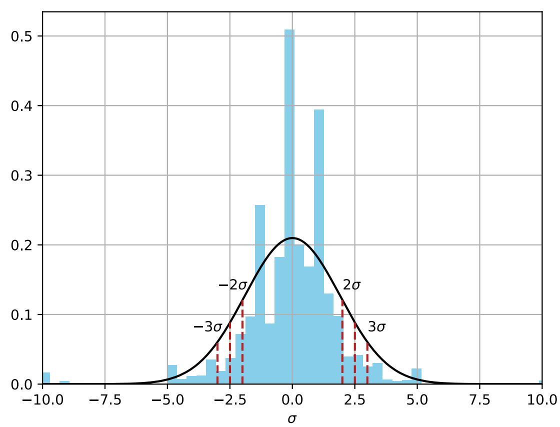


FIGURE 4. The standard normal distribution of Z-score.

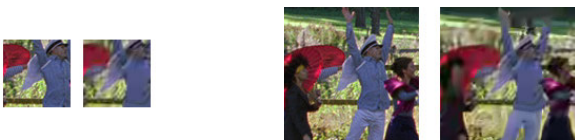
We select 2σ to minimize the error resulting in a reduction of 422 image patch pair scores in database. Fig. 6 shows the standard deviation of subjective ratings before and after outlier rejection. Most samples having deviations greater than 0.5 have been removed. The filtered data are presented as follows:

$$Y_f = \left\{((R_p, D_p), \overline{y_p}) \in Y_o | Z_{p,s} \geq 2\sigma_p; S_p \geq 15\right\}, \quad (4)$$

Finally, each pair of patches is evaluated by the average cleaned scores of at least 15 observers (with 5 levels). $N = |Y_f| = 40,286$ cleaned pairs are kept to make two final HMII (Human Machine Interaction Image) databases (Table 3). Each database comprises 40,286 quality annotated images based on 40,286 source reference image patches that are subject to different distortion levels of compression as in table. Fig. 5 is an example of image patch pair with two diference sizes. Differential mean opinion score (DMOS) for this dataset is computed for each pair, ranging from 10 to 50.

TABLE 3. HMII databases summary.

Database	Number of image patch pairs	Patch size
HMII-64	40286	64×64
HMII-128	40286	128×128



(a) 64×64 (b) 128×128

FIGURE 5. An example of image patch pairs (reference and distortion) with two diference sizes.

TABLE 4. Outlier rejection results.

Properties	2σ	2.5σ	3σ
Number of outliers	33199	8631	1991
Number of image patch pairs	422	136	37
Number subjects	21	7	3
Percentage outlier	5.0%	1.3%	0.3%

B. BENCHMARK ANALYSES

1) HMII DATABASE BENCHMARK ANALYSES

We implement of seven state-of-the-art algorithms (PSNR, UQI, VSI, SSIM, RFSIM, FSI and SRSIM) and two new methods (DIQaM-FR WaDIQaM-FR) [6] to predict objective scores for the entire HMII database. Table 5 shows that the pairwise preference consistency is evaluated using the classic correlation coefficients SRCC and PLCC. The SRCC and PLCC are the average values for the testing image patches of the same reference image patches, and the top two correlation coefficient values are in bold. It is seen that PSNR and UQI are less correlated with the quality perceived by humans, and even contrary to subjective results. This defective performance of PSNR is also mentioned in the work of Zhang *et al.* [39] about Fine-Grained Quality Assessment. Although VSI consists of HVS features and achieves more consistent results than PSNR in global image assessment, it is poorly correlated with human perceptual quality in fine-grained patch quality assessment. In general, FSIM achieves top two performances for all cases and SSIM achieves better performance with PLCC while SRSIM performs better with SRCC. For the two correlation coefficients, the above-mentioned IQA methods shows quite similar characteristics, while two new methods methods based on machine learning fails in proposed database. The reason is that each structure of a machine learning problem is only suitable for its own database.

TABLE 5. PLCC and SRCC for different IQA algorithms.

IQA ALGORITHM	HMII 64×64		HMII 128×128	
	SRCC	PLCC	SRCC	PLCC
PSNR	0.7056	0.6596	0.7233	0.6723
UQI [40]	0.0233	0.0233	0.0129	0.0124
VSI [41]	0.7659	0.7659	0.7680	0.7861
SSIM [3]	**0.7878**	0.7714	**0.7989**	0.7894
RFSIM [5]	0.7747	0.7574	0.7891	0.7596
FSIM [4]	**0.7941**	**0.7997**	**0.8241**	**0.8154**
SRSIM [35]	0.7776	**0.8030**	0.7188	**0.8035**
DIQaM-FR [6]	0.5525	0.5521	0.6075	0.6057
WaDIQaM-FR [6]	0.5648	0.6738	0.6750	0.7661

2) SIMPLE IMAGE-PATCH MODELS

In this experiment, different models are compared to find the best 'ground truth' predictor for patch quality. We use the following models with our database:

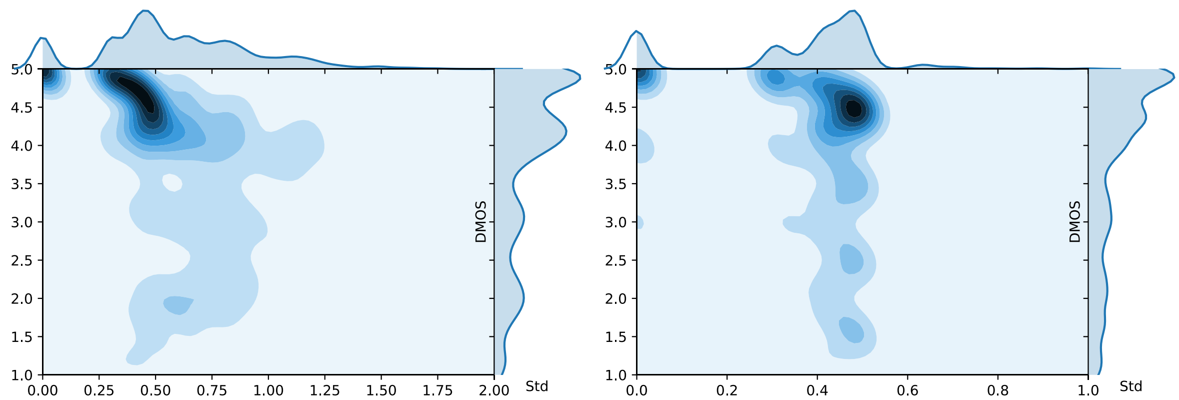
- *IPM*: Zhang in [39] assumes the curve model to predict image-patch quality is a cubic polynomial function:

$$f(\Phi(R_p); \theta) = a_1\Phi(R_p)^3 + a_2\Phi(R_p)^2 + a_3\Phi(R_p) + a_4, \quad (5)$$

  where $\theta = a_1, a_2, a_3, a_4$ is the parameter for the non-linear function of Image-Patch model and $\Phi(R_p)$ are the feature of reference patch R_p. MSE and SSIM are chosen for the design of features. In our work, we try top three FR-IQA methods: SSIM, FSIM and SRSIM.
- *DIQaM*: Bosse in [6] presents a CNN model for image quality assessment which obtains superior performance on different IQA benchmarks. We utilize the extractor architecture from this paper to train a Deep Neural Network on our database.

Firstly, we use previous works to extract SSIM, FSIM and SRSIM features for IPM. Then, the above curve model is fitted using the least mean square method to adapt the coefficients that best fit the database. DIPQA is developed based on DIQaM's extractor architecture which shares the same regression part. With the DIPQA, we use VGGNet as a feature extractor, this part is trained with the entire network.

Table 6 shows the performance of the simple models on HMII database. With any of the two correlation coefficients, DIPQA (VGG extractor) achieves superior performance to the others'. From the results of this experiment, it can be



(a) Raw database (b) After rejecting outlier

FIGURE 6. Standard deviation of subjective ratings.

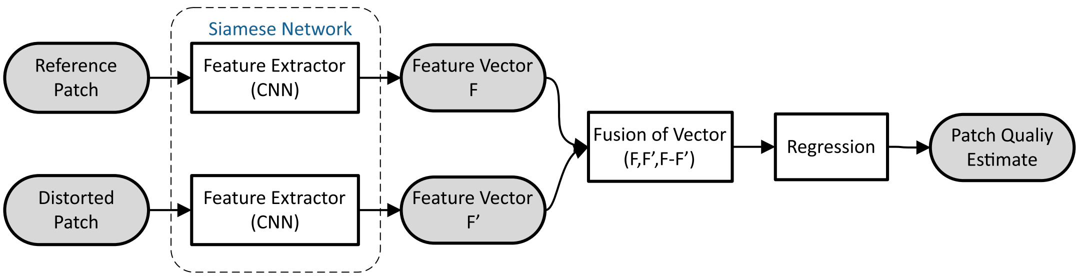


FIGURE 7. Deep Image-Patch Quality Assessment Network Architecture.

TABLE 6. Comparing different Full-Reference Image-Patch approaches.

METHOD	HMII (64x64) SRCC	HMII (64x64) PLCC	HMII (128x128) SRCC	HMII (128x128) PLCC
IPM (SSIM)	0.8362	0.7846	0.8434	0.7946
IPM (FSIM)	0.8484	0.7954	0.8715	0.8106
IPM (SRSIM)	0.8541	0.8025	0.8573	0.7980
DIPQA (VGG extractor)	0.9071	0.8382	**0.9167**	**0.8574**

seen that objective models assess image-patch quality more accurately with the larger size of the patch.

C. PROPOSED IQA METHOD

1) ARCHITECTURE OF THE PROPOSED METHOD

Being known as a designed architecture to learn the similarity relations between two given inputs, Siamese network has been applied for face verification [38] and signature [42] tasks. The main concept of this method is to process two networks that share the same architecture and weights in parallel. In this work, we employ a Siamese network for feature extraction. Before feeding the feature vector as input to the regression layers, two extracted feature vectors are combined by a fusion step. The proposed architecture of IQA method is sketched in Fig 7.

For training our IQA method, data set Y_f is randomly splitted by reference image. The training set is based on H pairs of reference and distort image, testing set on $N-H$ pairs. The training set is:

$$Y_{train} = \left\{((R_p, D_p), \overline{y_p}) \in Y_f | p = 1..L\right\}, \tag{6}$$

and testing set is:

$$Y_{test} = \left\{((R_p, D_p), \overline{y_p}) \in Y_f | p = L+1..N\right\}. \tag{7}$$

Let us describe the notations in convolution filter of CNN feature extractor which consists of a stack of convolutional layers, pooling layers, and full-connected layers. Let l denote the l^{th} layer where L is the number of layers. $H \times W$ (pixels) respectively be the width and height of input image patch x. Let $w^l_{m,n}$ denote the weight matrix between neurons of layer l and neurons of layer $l-1$. The convolved data streams at layer l plus the bias unit b^l are defined as follows:

$$x^l_{i,j} = \sum_m \sum_n w^l_{m,n} o^{l-1}_{i+m,j+n} + b^l, \tag{8}$$

where $i = 1, 2, 3 \ldots H, j = 1, 2, 3 \ldots W$ are row and column iterators of input vector; $m = 1, 2, 3 \ldots k_1, n = 1, 2, 3 \ldots k_2$ are iterators of filter w. The output vector at layer 1 $o^l_{i,j}$ are depicted as below:

$$o^l_{i,j} = f(x^l_{i,j}), \tag{9}$$

where $f(\cdot)$ denotes the activation function. Application of the activation layer to the convolved input vector at layer l is given by:

$$f(x_{i,j}^l) = \begin{cases} x_{i,j}^l, & \text{if } x_{i,j}^l > 0 \\ a_i x_{i,j}^l, & \text{if } x_{i,j}^l \leq 0 \end{cases}, \tag{10}$$

where a_i is the coefficient controlling the slope of the negative part. All layers of feature extractor are activated by a rectified linear unit (ReLu) [43] when $a_i = 0$. The resultant activation function is of the form $f(x_{i,j}^l) = max(0, x_{i,j}^l)$.

Let $\overline{y_i}$ denotes the ground-truth indicator vector, feature extractor produces the activation of the reference patches feature vector $F = [F_1, \ldots, F_P]$ and the distortion patches feature vector $F' = [F'_1, \ldots, F'_P]$. To denote the weights of last convolution filter in feature extractor by w^L, we define the feature vectors function according to the formulas:

$$\begin{aligned} F_{p,i,j} &= max(0, R_{p,i,j}^L) \\ &= max(0, \sum_m \sum_n w_{m,n}^L R_{p,i+m,j+n}^{L-1} + b^L) \end{aligned} \tag{11}$$

and

$$\begin{aligned} F'_{p,i,j} &= max(0, D_{p,i,j}^L) \\ &= max(0, \sum_m \sum_n w_{m,n}^L D_{p,i+m,j+n}^{L-1} + b^L) \end{aligned} \tag{12}$$

where R_p^l and X_p^l represent a feature map in l layer of the input patch R_p and D_p, respectively.

The feature extraction layers extract F and F' which are the feature vectors of reference and distorted patch respectively. The regression layers require the network to combine these two vectors in a feature fusion step. The simplest strategy is concatenating F and F' to an unique vector (F, F'). Besides, $F - F'$ is known as a meaningful representation for distance in feature space. Therefore, concatenating $F - F'$ is expected to contribute to learning relations between reference and distorted patch. The final output of this state is:

$$F''_p = Concat(F_p, F'_p, F_p - F'_p) \tag{13}$$

Finally, the fused features are regressed by a sequence of two fully-connected layer including: FC-512 and FC-1. The inference of fully connected layer can be represented by:

$$F'''_p = max(0, w_{L+1} * F''_p + b_{L+1}) \tag{14}$$

$$q_p = max(0, w_{L+2} * F'''_p + b_{L+2}), \tag{15}$$

where q_p represents the output IQA method, $*$ is the convolutional operation.

2) EVALUATION CRITERIA

IQA estimation algorithm's performance is measured through the deviation between the estimated and actual values. The common method to test the efficiency of IQA estimation algorithms is using Mean Absolute Error (MAE). The smaller MAE value obtained is, the smaller the error margin in prediction is made. Let $\overline{y_p}$ denotes the subjective testing IQA and q_p denotes the predicted IQA of the pair p^{th}. MAE is calculated as the average of the sum of absolute differences between two IQA variables in the following equation (15).

$$MAE = \frac{\sum_{p=1}^{N-H} |\overline{y_p} - q_p|}{N - H}, \tag{16}$$

where $N - H$ is the number of testing patch pairs.

3) TRAINING OPTIMIZATION

Our network is trained end-to-end by backpropagation, over a number of epochs. The adaptive moment estimation optimizer (ADAM) is used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data. Optimization problem is to minimize the cost function $J(.)$ as defined in equation (17).

$$min_w j(w), \tag{17}$$

where w is the weight vector. ADAM uses exponentially decaying average of past gradients, m_k (first moment) and past squared gradients, v_k (second moment) as given in equation (18) and (19) respectively. Adam weight update equation can be mathematically represented as equation (20).

$$m_l = \beta_1 m_{l-1} + (1 - \beta_1)\nabla J(w_l) \tag{18}$$

$$v_l = \beta_2 v_{l-1} + (1 - \beta_2)\nabla^2 J(w_l) \tag{19}$$

$$w_{l+1} = w_l - \alpha \cdot \frac{\sqrt{1 - \beta_1^l}}{1 - \beta_1^l} . \frac{m_l}{\sqrt{v_l} + \epsilon}, \tag{20}$$

where α is the learning rate, w_l is the weight vector, β_1 and β_2 are momentum parameters. Parameters of ADAM are chosen as recommended in [44] $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and the learning rate α is initially set to 5×10^{-4}.

4) FEATURE ENGINEERING

With the general architecture in Fig. 7, we select one of five different feature extractors as follows:

1) VGGnet: With the successful adaptation for various computer vision tasks [45], [46], especially in image quality assessment [6], VGGnet [47] is chosen as a base network for the feature extraction. The input of the VGG network is the size of 224 × 224 pixels. For the purpose of adjusting the network for 64 × 64 and 128 × 128 pixels, we have tried to change the architecture of VGG network such as: extending the network by 3 layers [6], cutting last 3 layers, last 6 layers or even replacing VGG with Resnet. Finally, we choose to cut the last 3 layers of VGGnet and achieve the best performance comparing to other approaches (Fig. 8). Our VGGnet-inspired DCNN comprises 12 weight layers as a feature extraction module and a regression module. The features are extracted in a series of conv3-64, conv3-64, maxpool, conv3-128, conv3-128, maxpool, conv3-256, conv3-256, maxpool, conv3-512, conv3-512 and maxpool layers. This results in about

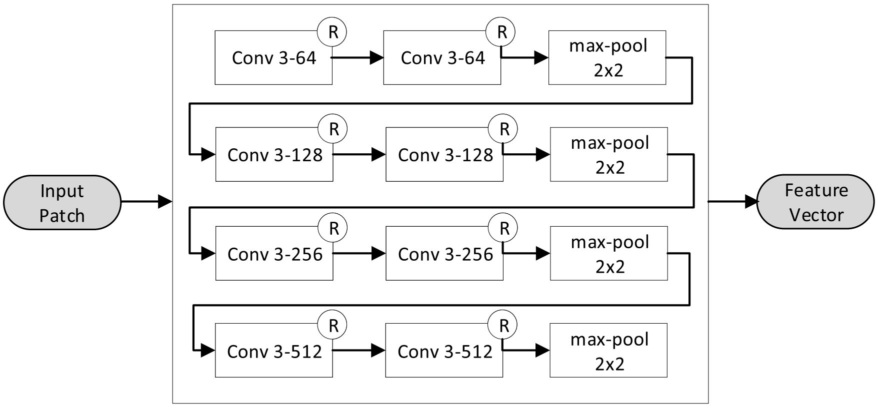


FIGURE 8. VGGnet feature extractor.

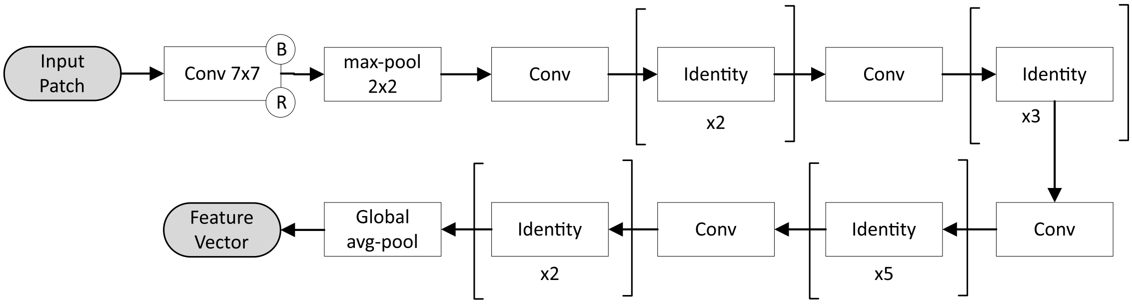


FIGURE 9. ResNeXt-50 feature extractor architecture.

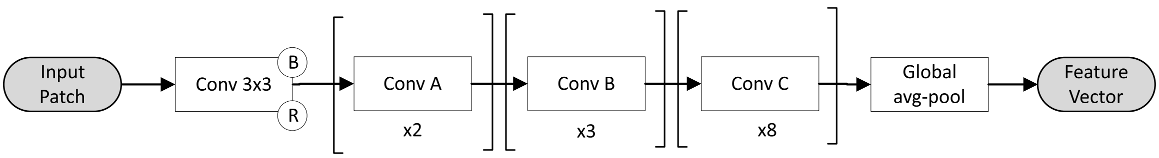


FIGURE 10. Xception feature extractor architecture.

17.3 million trainable network parameters. All convolutional layers apply 3 × 3 pixel-size convolution kernels.

2) ResNeXt-50: ResNext, also known as Aggregated Residual Transform Network, is an improvement over the Inception Network. Xie *et al.* [48] exploited the topology of the split, transformed and merged in a powerful but simple way by introducing a new term "cardinality". The model consists of 1 convolutional (7 × 7) layer, 1 maxpooling layer, 4 convolutional blocks alternated by 4 groups of identity blocks. We modify this network for extraction by removing the last Average layer and adding a Single convolutional layer instead before features fusing as shown in Fig. 9.
3) Xception: Chollet in [49] revised the idea of Inception modules and offered to use depthwise separable convolutions by maximizing the number of towers in a module. Basically we use the original, however, similar to ResNeXt-50 network, the last part is changed as shown in Fig.10.
4) Inception-v4: Inception V4 [50] is a well-known architecture developed based on the GoogLeNet platform, the input of this network is an image patch (299 × 299 pixels), the output depends on how many classes targeted to predict. In the pre-trained model used in this research, we keep Average Pooling layer and add a Single convolutional layer before features fusing in Fig. 11.
5) Inception-ResNets: Inception-ResNet-v2 is a combination of two recent networks, residual connections [51]. The Inception models are famous for their multi-branch architectures. They have a set of filters (1 × 1, 3 × 3, 5 × 5, etc.) that are merged with concatenation in each branch. The split-transform-merge architecture of the inception module is observed as a powerful

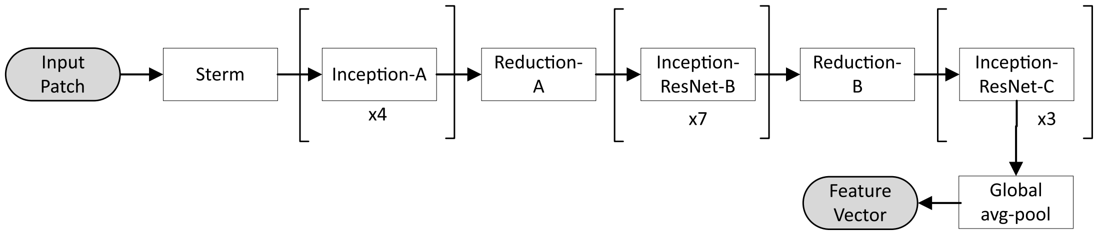


FIGURE 11. **Inception-v4 extractor architecture.**

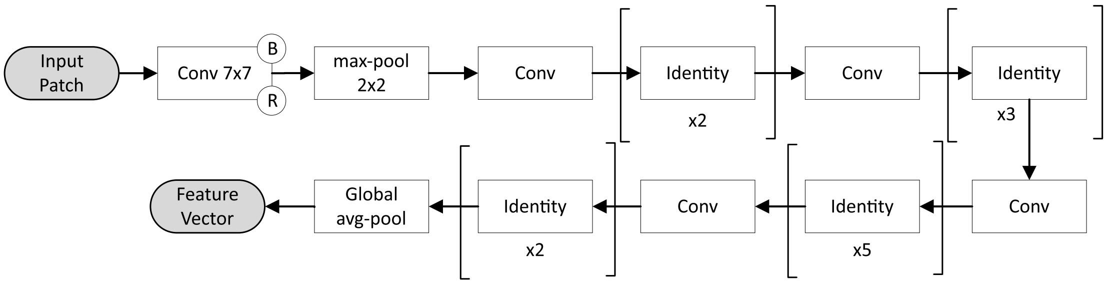


FIGURE 12. **Inception-ResNets extractor architecture.**

representational ability in its dense layers. The hybrid Inception-ResNet-v2 network shows in Fig. 12.

IV. EXPERIMENTAL RESULTS

In this section, we introduce the experimental setup including the Datasets, the Evaluation Metrics and the implementation details of IQA model. The best feature extractor applied in the proposed model (HMI-IQA) is chosen from 5 CNN architectures. In addition, we test the proposed model for 4 common datasets including [8], TID2008 [9], TID2013 [10] and CSIQ [11].

A. EXPERIMENTAL SETUP

1) DATASETS

There are 2 types of datasets used in experiments. The first one, training datasets, is used for optimizing the proposed IQA deep neuron network model and the other, cross-evaluation datasets, is for the independent evaluation of the proposed model:

- *Training Datasets*: We use datasets HMII_64 and HMII_128 as presented in Table 3 and Fig. 5 for the purpose of our IQA deep neuron nework model optimization. Reported results are based on the average performance of ten folds cross-validation. For cross-validation, a HMII database is randomly split into 8:1:1 ratio for training, validating and testing sets, respectively. Deep learning models converge after 50 epochs for each dataset.
- *Cross-evaluation Datasets*: For evaluating the generalization ability of the proposed IQA model after training step, we use the verified IQA databases in public domain. We choose four comprehensive databases mentioned above to be used as benchmarks: LIVE [8], CSIQ [11], TID2008 [9] and TID2013 [10]. All perceptual quality of Datasets are normalized in range [10;50].

2) EVALUATION METRICS

To evaluate the performances of the IQA algorithms, two measures are used including Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC). PLCC measures the linear dependence between two quantities and SRCC measures how well one quantity can be described as a monotonic function of another quantity.

B. EXPERIMENT RESULTS

1) PERFORMANCE OF PROPOSED METHOD

In our experiment, four other feature extractors in turn replace the VGG-16 extractor in the first one. Only the HMII dataset with the size of $128 \times 128 \times 3$ is used to learn models. The same number of epochs and other criteria are run with the first experiment. The default size of input patches is 224×224 for ResNext and VGGnet and 299×299 for Inception-V4, Inception-ResNet-V2 and Xception which are different from that of patches in our database. Thus, the architecture of feature extractors has been adjusted to fit our inputs while its outputs are feature vectors. Features are extracted from the distorted patch and the reference patch by a CNN and fused as difference, concatenation or concatenation supplementary with the difference vector.

Table 7 shows the results of the two experiments comparing five models with different feature extractors. The model using

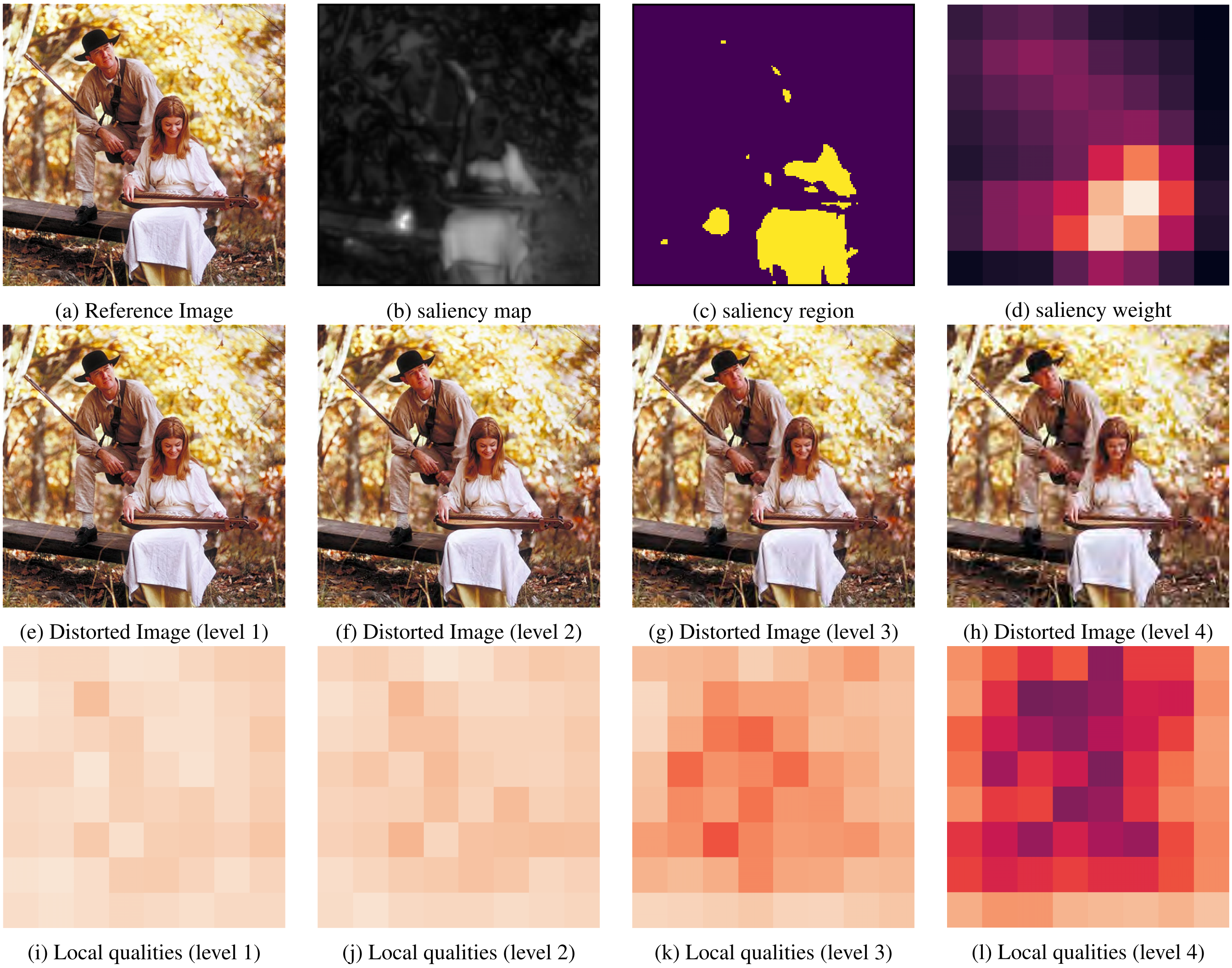


FIGURE 13. Local qualities and saliency weights for a JP2K distorted image from CSIQ. The clors black and white indicate low and high values of local qualities respectively. The DMOS values are 0.9978, 0.9274, 0.6141, 0.3693 in range [0;1], predicted qualities are 47.7512, 46.7519, 42.5574, 32.1268 in range [10;50].

TABLE 7. Comparing different feature extraction architectures.

FEATURE EXTRACTOR	SRCC	PLCC
DIPQA (VGG extractor)	0.9167	0.8574
Inception-ResNetsV2	0.8940	0.8405
Xception	0.9112	0.8582
Inception V4	0.8825	0.8146
ResNeXt-50	**0.9222**	**0.8764**

Resnext-50 named HMI-IQA has the best performance and is in bold.

2) CROSS-DATABASE EVALUATION

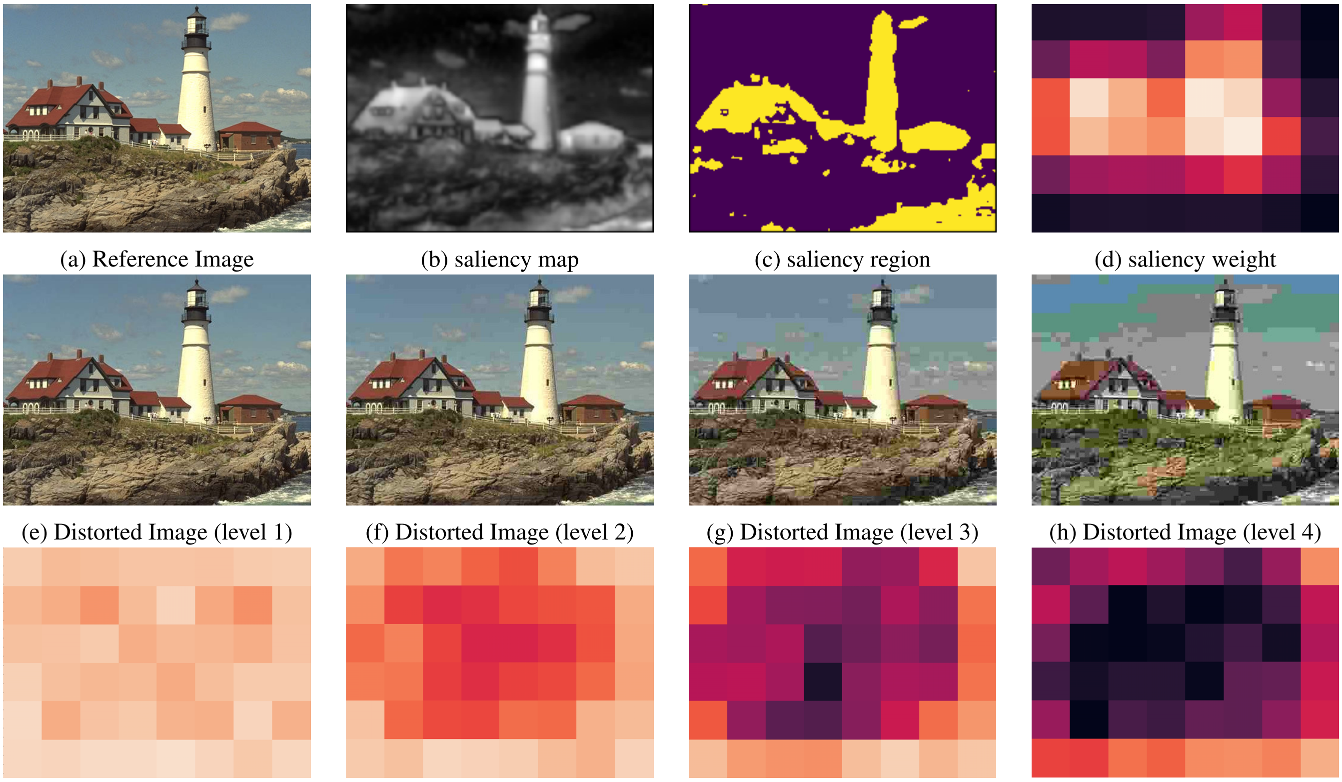
We train HMI-IQA models on HMII database and test them in the four above-mentioned Evaluation Datasets. Each referent image is divided into 64 × 64 and enlarged to 128 × 128 patches which are used to predict DMOS score of the same local distort image patch. The DMOS value of the distorted image is calculated in 2 ways: average (HMI-IQA Aver) and saliency (HMI-IQA Sal) weight of local qualities. In the first one, $\overline{q}$ is estimated by taking the average of local visual qualities q_i as following formula:

$$\overline{q} = \frac{1}{N_p} \sum_{j=1}^{N_p} q_j, \tag{21}$$

where N_p denotes the number of patches from the image.

The second way combines models of visual saliency with IQAs by weighting the local quality q_j of a region j with the corresponding local weight w_j. To determine w_j, the framework in [52] is used to detect regional saliency adjustment and measure the weight by counting pixels in this region. The overall image quality $\overline{q}$ is:

$$\overline{q} = \frac{\sum_{j=1}^{N_p} w_j q_j}{\sum_{j=1}^{N_p} w_j}, \tag{22}$$



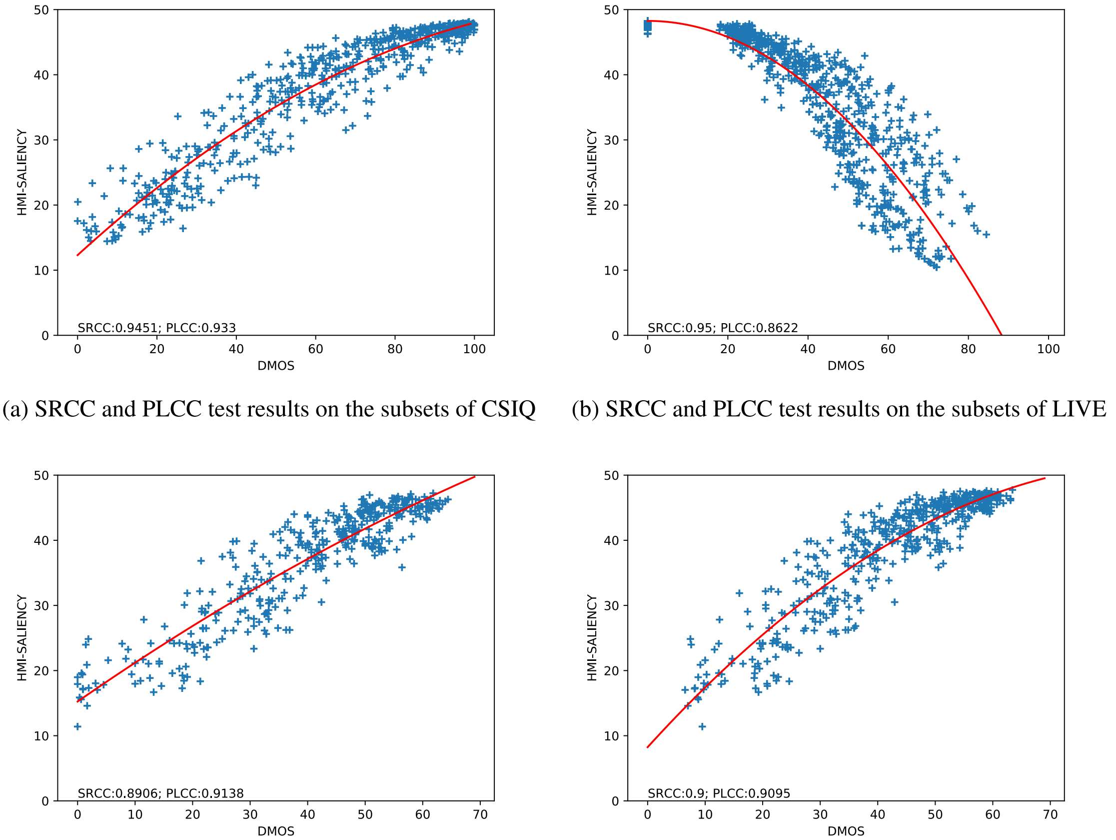
(a) Reference Image (b) saliency map (c) saliency region (d) saliency weight
(e) Distorted Image (level 1) (f) Distorted Image (level 2) (g) Distorted Image (level 3) (h) Distorted Image (level 4)
(i) Local qualities (level 1) (j) Local qualities (level 2) (k) Local qualities (level 3) (l) Local qualities (level 4)

FIGURE 14. Local qualities and saliency weights for a JPEG distorted image from TID2008. The clors black and white indicate low and high values of local qualities respectively. The MOS values are 6, 4.8065, 3.25, 1.3226 in range [0,9], predicted qualities are 45.1306, 36.4296, 26.1536, 16.6870 in range [10;50].

TABLE 8. SRCC and PLCC results of individual distortion types evaluated on TID2013.

DISTORTION	DIQaM-FR [6] trained on LIVE		WaDIQaM-FR [6] trained on LIVE		HMI-IQA AVER		HMI-IQA SAL	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Additive Gaussian noise	0.8260	0.5540	**0.9192**	**0.9203**	0.8989	0.8699	0.8994	0.8739
Additive noise in color components	0.7401	0.4503	0.8083	0.8059	0.8213	0.8421	**0.8241**	**0.8510**
Spatially correlated noise	0.8000	0.7946	0.8326	0.7856	0.9187	0.9049	**0.9193**	**0.9070**
Masked noise	0.7631	0.6633	0.7901	0.7688	0.8019	0.8323	0.8084	0.8493
High frequency noise	0.8126	0.6914	**0.9125**	**0.9265**	0.8910	0.9203	0.8966	0.9258
Impulse noise	0.5706	0.4091	0.8420	0.8178	0.8630	0.8292	0.8718	0.8410
Quantization noise	0.7204	0.6046	0.8034	0.7681	0.8500	0.8406	0.8476	0.8450
Gaussian blur	0.8574	0.8442	0.9310	0.9199	0.8846	0.8247	0.8882	0.8443
Image denoising	0.7737	0.7737	0.9393	0.9616	0.9352	0.9575	0.9350	0.9548
JPEG compression	0.8258	0.8348	0.9273	0.9506	0.9381	0.9674	**0.9403**	**0.9688**
JPEG2000 compression	0.9404	0.8873	0.9484	0.9328	0.9566	0.9633	**0.9595**	**0.9712**
JPEG transmission errors	0.3123	0.2885	0.6529	0.6463	0.7564	0.7543	0.7436	0.7374
JPEG2000 transmission errors	0.5727	0.5201	0.6106	0.4860	0.7796	0.6871	0.7706	0.6785
Non eccentricity pattern noise	0.6769	0.2333	0.6433	0.6046	0.7890	0.7680	0.8014	0.7887
Local block-wise distortions of different intensity	0.5297	0.2647	0.2638	0.2471	0.5383	0.5443	0.4195	0.3870
Mean shift (intensity shift)	0.6217	0.6210	0.6494	0.5870	0.7258	0.7260	0.7299	0.7324
Contrast change	0.5493	0.6937	0.2715	0.3602	0.5150	0.6648	0.5211	0.6681
Change of color saturation	0.7100	0.6743	0.7460	0.7332	0.7853	0.7660	0.7722	0.7636
Multiplicative Gaussian noise	0.7728	0.5076	0.8932	0.8901	0.8719	0.8523	0.8784	0.8510
Comfort noise	0.4769	0.4645	0.8709	0.8403	0.9084	0.9133	0.9108	0.9161
Lossy compression of noisy images	0.8462	0.6751	0.9408	0.9228	0.9553	0.9552	**0.9548**	**0.9556**
Image color quantization with dither	0.4572	0.3920	0.8186	0.7928	0.8688	0.8886	0.8668	0.8903
Chromatic aberrations	0.8846	0.8414	0.8090	0.9092	0.8687	0.9357	0.8638	0.9388
Sparse sampling and reconstruction	0.8786	0.7794	0.9480	0.9490	0.9588	0.9651	**0.9585**	**0.9701**

Fig. 13 shows the local qualities q_j and weights w_j for an image subject to JP2K compression from CSIQ. The DMOS values of the four level distorted images are 0.9978, 0.9274, 0.6141, 0.3693 in range [0;1]; the relation between prediction accuracy of the two different ways are as expected from the previous evaluations (average weight: 45.3548, 42.6473,



(a) SRCC and PLCC test results on the subsets of CSIQ (b) SRCC and PLCC test results on the subsets of LIVE

(c) SRCC and PLCC test results on the subsets of TID2008 (d) SRCC and PLCC test results on the subsets of TI2013

FIGURE 15. SRCC and PLCC test results on the subsets of Datasets.

TABLE 9. SRCC and PLCC results of individual distortion types evaluated on TID2008.

DISTORTION	DIQaM-FR [6] trained on LIVE		WaDIQaM-FR [6] trained on LIVE		HMI-IQA AVER		HMI-IQA SAL	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Additive Gaussian noise	0.8187	0.4381	**0.9018**	**0.8906**	0.8826	0.8505	0.8821	0.8542
Additive noise in color components	0.8132	0.4562	0.8674	0.8458	**0.8929**	0.8889	0.8928	**0.8915**
Spatially correlated noise	0.7815	0.7673	0.7863	0.7508	0.9215	0.9175	**0.9228**	**0.9191**
Masked noise	0.7528	0.5824	0.7945	0.7541	0.8292	0.8315	**0.8332**	**0.8488**
High frequency noise	0.8265	0.6462	**0.9337**	**0.9329**	0.9166	0.9177	0.9199	0.9260
Impulse noise	0.4612	0.2383	0.8297	0.7808	0.8404	0.8236	**0.8408**	**0.8297**
Quantization noise	0.6993	0.5268	0.7347	0.6991	0.8487	0.8496	**0.8401**	**0.8500**
Gaussian blur	0.8316	0.7813	**0.9200**	**0.9024**	0.8183	0.7836	0.8222	0.8032
Image denoising	0.8338	0.7773	0.9515	0.9536	0.9590	0.9614	**0.9571**	**0.9573**
JPEG compression	0.8329	0.7953	0.9335	0.9392	0.9532	0.9697	**0.9553**	**0.9715**
JPEG2000 compression	0.9597	0.8672	0.9590	0.9173	0.9564	0.9565	**0.9617**	**0.9672**
JPEG transmission errors	0.2378	0.2144	0.5865	0.5861	0.7324	0.7151	**0.7151**	**0.6950**
JPEG2000 transmission errors	0.5232	0.4685	0.5144	0.3568	0.7286	0.6342	**0.7173**	**0.6264**
Non eccentricity pattern noise	0.6853	0.3973	0.6935	0.6972	0.7452	0.6968	**0.7600**	**0.7170**
Local block-wise distortions of different intensity	0.7843	0.6465	0.0869	0.1526	0.8388	0.8419	**0.7517**	**0.7425**
Mean shift (intensity shift)	0.4968	0.5005	0.5362	0.4882	0.6564	0.6313	**0.6691**	**0.6476**
Contrast change	0.7030	0.7229	0.3929	0.3951	0.6812	0.7351	**0.6801**	**0.7406**

39.6324, 33.3425; saliency weight: 47.7512, 46.7519, 42.5574, 32.1268). Similarly, Fig. 14 shows an image subject to JPEG compression from TID2008. It can be seen in Fig. 13 and Fig. 14 that the predicted quality of patches are spatially correlated as a result, the interpolation method can be used to compute the quality of the smaller patches.

The proposed model is tested on a specific distortion type and shows the results on TID2013, TID2008, LIVE and CSIQ in Table 8, 9, 10 and 11, respectively. These tables show that HMI-IQA Sal performs better than HMI-IQA Aver and HMI-IQA Sal is among the top performing models with 28 out of 52 times. Specifically, HMI-IQA

TABLE 10. SRCC and PLCC results of individual distortion types evaluated on LIVE.

DISTORTION	DIQaM-FR [6] Trained on TID2013		WaDIQaM-FR [6] Trained on TID2013		HMI-IQA AVER		HMI-IQA SAL	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Jpeg2000	0.9297	0.8888	**0.9626**	**0.8983**	0.9523	0.8777	0.9550	0.8829
Jpeg	0.7472	0.7480	**0.9665**	**0.9476**	0.9458	0.8946	0.9485	0.9016
Fastfading	0.7846	0.7698	**0.9343**	**0.8935**	0.9065	0.7998	0.9074	0.8170
Blur	0.7249	0.7444	**0.9277**	**0.8854**	0.9048	0.8603	0.9154	0.8714
Wn	0.8505	0.8015	**0.9616**	0.9367	0.9850	0.9134	**0.9875**	0.9155

TABLE 11. SRCC and PLCC results of individual distortion types evaluated on CSIQ.

DISTORTION	DIQaM-FR [6] Trained on TID 2013		WaDIQaM-FR [6] Trained on TID 2013		DIQaM-FR [6] Trained on LIVE		WaDIQaM-FR [6] Trained on LIVE		HMI-IQA AVER		HMI-IQA SAL	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
AWGN	0.9170	0.9315	0.9067	0.9172	0.8838	0.8740	0.9370	0.9448	0.9349	0.9019	0.9391	0.9101
Jpeg	0.8548	0.9429	0.9479	**0.9750**	0.8872	0.9153	0.9376	0.9551	0.9418	0.9652	0.9435	0.9679
Jpeg2000	0.8463	0.9012	**0.9696**	0.9653	0.8954	0.8990	0.9581	0.9671	0.9673	0.9714	**0.9691**	**0.9747**
Fnoise	0.8967	0.8982	0.9277	0.9340	0.8973	0.9013	0.8630	0.8469	0.9379	0.9439	**0.9314**	0.9378
Blur	0.9004	0.9163	0.9469	0.9399	0.9322	0.9293	**0.9580**	**0.9600**	0.9414	0.9267	0.9496	0.9380
Contrast	0.9150	0.9239	0.9180	0.9128	**0.9287**	0.9166	0.9185	**0.9221**	0.9163	0.9118	0.9169	0.9193

TABLE 12. SRCC and PLCC test results on the subsets of Datasets contain only the 4 distortions.

DISTORTION	AVERAGE		SALIENCY	
	SRCC	PLCC	SRCC	PLCC
TID2013	0.8965	0.9016	0.9000	0.9095
TID2008	0.8844	0.9069	0.8906	0.9138
LIVE	0.9449	0.8493	0.9500	0.8622
CSIQ	0.9404	0.9330	0.9451	0.9417

model outperforms in compressed distortions by JPEG and JPEG2000 because of the effectiveness of pre-training in HMII database. We observed that HMI-IQA method performs well on unseen distortion types, including lossy compression of noisy images, sparse sampling and reconstruction, spatially correlated noise and comfort noise. In addition, proposed model fails in three distortion types on TID2013, i.e., compression transmission errors, local block-wise distortions, and contrast change, whose characteristics are difficult to model. However, the performance of HMI-IQA model is only equivalent to WaDIQaM but insignificantly better than other methods. The reason is that this study only focuses on patch quality without considering its contribution to the whole image quality. HMI-IQA Sal method uses simple saliency model only to test the effectiveness of the patch-based method. To tackle this issue, we will use whole image quality databases to develop a weighted estimate method instead of a replacement of existing visual saliency models.

The proposed HMI-IQA model is evaluated on subsets of CSIQ, LIVE, TID2008 and TID2013, containing only the four distortions types shared among the four databases (JPEG, JP2K, Gaussian blur and white noise). Table 12 shows that performances of the proposed model are relatively stable for all subsets. Fig.15 shows the scattering distributions of subjective DMOS versus the predicted scores obtained by the HMI-IQA Sal on four subset databases. The proposed model still anticipates some other common types of distortion quite well as various types of distortion are generated in training database during the process of video compression. However, the prediction is inadequately good in case of high noise level. As can be seen in Fig.15 when the image quality is reduced, the prediction is less accurate. In further study, we will collect more experimental data of other distortion types to enhance the accuracy.

V. CONCLUSION

In this paper, we present an experimental image quality assessment solution for image/video with compression artifacts. First, the subject quality rating database considering image patch quality assessment method for image/video with compression artifacts are introduced. Due to the lack of 'ground truth' quality of patches, we expect the proposed image patches database to be useful for further investigation. Second, we introduce an efficient deep neural network based image patch quality assessment solution with several feature engineering options. Experimental results conducted for a rich set of image/video database shows that the proposed IQA method is particularly suitable for image/video with compression artifacts, not only under the video HEVC compression but also with image JPEG or JPEG-2000 compression standards. For future works, we will explore the proposed IQA model to improve the image/video compression efficiency by directly integrating it into those standards.

REFERENCES

[1] Y. Shi and H. Sun, *Image and Video Compression for Multimedia Engineering: Fundamentals, Algorithms, and Standards*, 2nd ed. Boca Raton, FL, USA: CRC Press, Jan. 2017.

[2] Y. Wang, Y.-Q. Zhang, and J. Ostermann, *Video Processing and Communications*, 1st ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[3] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 516–529, Apr. 2012.

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

[5] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using riesz transforms," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 321–324.
[6] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.
[7] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2791–2795.
[8] L. C. H. R. Sheikh, Z. Wang, and A. C. Bovik. (2006). *Live Image Quality Assessment Database Release 2*. [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm
[9] N. N. Ponomarenko, V. V. Lukin, A. A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. (2008). *TID2008—A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics*. [Online]. Available: http://www.ponomarenko.info/tid2008.htm
[10] N. N. Ponomarenko, V. V. Lukin, A. A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. (2013). *Tampere Image Database 2013*.[Online]. Available: http://www.ponomarenko.info/tid2008.htm
[11] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.
[12] P. L. Callet and F. Autrusseau. (2005). *Subjective Quality Assessment Irccyn/IVC Database*. [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/
[13] S. Tourancheau, F. Autrusseau, Z. M. P. Sazzad, and Y. Horita, "Impact of subjective dataset on the performance of image quality metrics," in *Proc. 15th IEEE Int. Conf. Image Process.*, San Diego, CA, USA, Oct. 2008, pp. 365–368. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00321663
[14] L. Jin, K. Egiazarian, and C.-C.-J. Kuo, "Perceptual image quality assessment using block-based multi-metric fusion (BMMF)," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1145–1148.
[15] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, pp. 669–688, Aug. 1993.
[16] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 539–546.
[17] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[18] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, p. 57–77, Jan. 2015.
[19] T. Liu, H. Liu, S. Pei, and K. Liu, "A high-definition diversity-scene database for image quality assessment," *IEEE Access*, vol. 6, pp. 45427–45438, 2018.
[20] A. Ninassi, F. Autrusseau, and P. Le Callet, "Pseudo no reference image quality metric using perceptual data hiding," *Human Vis. Electron. Imag.*, vol. 6057, Feb. 2006, Art. no. 60570G.
[21] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP J. Image Video Process.*, vol. 2008, Dec. 2008, Art. no. 659024. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00362891
[22] F. Autrusseau and M. Babel. (2009). *Subjective Quality Assessment of LAR Coded Art Images*. [Online]. Available: http://www.irccyn.ec–nantes.fr/~autrusse/Databases/LAR/
[23] C. Strauss, F. Pasteau, F. Autrusseau, M. Babel, L. Bedat, and O. Deforges, "Subjective and objective quality evaluation of lar coded art images," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2009, pp. 674–677.
[24] L. Goldmann, F. De Simone, and T. Ebrahimi, "Impact of acquisition distortion on the quality of stereoscopic images," in *Proc. Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, Sep. 2010.
[25] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Mod. Radioelectronics*, vol. 10, pp. 30–45, Jan. 2009.
[26] Z. Luo, L. Song, S. Zheng, and N. Ling, "H.264/advanced video control perceptual optimization coding based on JND-directed coefficient suppression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 935–948, Jun. 2013.
[27] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 193–222, Mar. 1995.
[28] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vis. Res.*, vol. 42, pp. 107–123, Feb. 2002.
[29] J. Li and Z. Yue, "Visual saliency based blind image quality assessment via convolutional neural network," in *Proc. Int. Conf. Neural Inf. Process.*, Oct. 2017, pp. 550–557.
[30] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.
[31] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.
[32] S. Golestaneh and K. Kitani, "No-reference image quality assessment via feature fusion and multi-task learning," 2020, *arXiv:2006.03783*. [Online]. Available: https://arxiv.org/abs/2006.03783
[33] J. Wu, Z. Xia, H. Zhang, and H. Li, "Blind quality assessment for screen content images by combining local and global features," *Digit. Signal Process.*, vol. 91, pp. 31–40, Aug. 2019.
[34] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, 2003, pp. 1398–1402.
[35] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1473–1476.
[36] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2387–2390, Dec. 2015.
[37] T. T. Pham, T. D. Dinh, V. X. Hoang, T. V. Huu, and T. H. Le, "Distortion model based on perceptual of local image content," in *Proc. 4th Int. Conf. Consum. Electron. Asia*, Jun. 2019, pp. 1–4.
[38] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Int. Telecommun. Union, Geneva, Switzerland, 2003.
[39] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, and W. Gao, "Fine-grained quality assessment for compressed images," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1163–1175, Mar. 2019.
[40] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[41] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.
[42] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 539–546.
[43] V. Nair and G. Hinton, "Rectified linear units improve restricted Boltzmann machines Vinod Nair," in *Proc. ICML*, vol. 27, Jun. 2010, pp. 807–814.
[44] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," in *Proc. ICLR, Int. Conf. Learn. Represent.*, 2015.
[45] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[46] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[48] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," 2016, *arXiv:1611.05431*. [Online]. Available: http://arxiv.org/abs/1611.05431
[49] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[50] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *AAAI Conference on Artificial Intelligence*, 02 2016.
[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[52] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

TUNG THANH PHAM was born in Hoabinh, Vietnam, in 1982. He received the B.E. degree in fire safety from the University of Fire Fighting and Prevention, Vietnam, in 2005, the B.E. degree in computer science from the Hanoi University of Science and Technology, Vietnam, in 2008 and the M.Sc. degree in software engineering from the University of Engineering and Technology, VNU, Vietnam, in 2014, where he is currently pursuing the Ph.D. degree in computer science, under the supervision of Prof. Dr. L. T. Ha. He is also with the University of Fire Fighting and Prevention.

His research interests include video coding, image processing, fire alarm systems, and simulation.

XIEM VAN HOANG received the B.S. degree in electrical engineering from the Hanoi University of Science and Technology, Vietnam, in 2009, and the M.S. degree from Sungkyunkwan University, Suwon, South Korea, in 2011, and the Ph.D. degree in electrical engineering from Lisbon University, Portugal, in 2015.

He is currently a Faculty Member of the Vietnam National University–University of Engineering and Technology. His research interests include image, video processing, and embedded systems development. He has received several awards for his contributions on video coding, including the Best Paper Award at the Picture Coding Symposium 2015 and the Ph.D. Award from the Fraunhofer Portugal Institute.

NGHIA TRUNG NGUYEN received the B.E. degree in computer science from the University of Engineering and Technology, VNU, Vietnam, in 2019.

He is currently working with the Medical Imaging Department, Vingroup Big Data Institute (VinBDI), Hanoi, Vietnam, aiming to conduct research on important areas of big data. His research interests include computer vision, computer graphics, artificial intelligence, and human–computer interaction.

DUONG TRIEU DINH received the B.S. and M.S. degrees in electrical engineering from the College of Technology, Vietnam National University, Hanoi, and the Ph.D. degree from Korea University, South Korea, in 2010.

He is currently a Lecture with the Faculty of Electronics and Telecommunications, Vietnam National University–University of Engineering and Technology, Hanoi, Vietnam. His research interests includes telecommunication, video coding, and communication.

LE THANH HA received the B.S. and M.S. degrees in information technology from the College of Technology, Vietnam National University, Hanoi, and the Ph.D. degree, in 2010.

In 2005, he received a Korean Government Scholarship for Ph.D. program with the Department of Electronics Engineering, Korea University. He is currently an Assistant Professor with the Faculty of Information Technology, University of Engineering and Technology, Vietnam National University, Hanoi. His research interests include computer vision, computer graphics, artificial intelligence, and human–computer interaction. He has also been principle and a main Investigator of many fundamental research and technology development projects funded by both domestic and international organizations. He also makes contributions in serving many domestic and international ICT academic conferences, including KSE, NICS, ATC, SoICT, and ICEIC. In addition, he is a member of the Institute of Electrical and Electronics Engineers (IEEE), The Institute of Electronics, Information, and Communication Engineers (IEICE), and The Vietnamese Association for Pattern Recognition (VAPR).