Blind Predicting Similar Quality Map for Image Quality Assessment

Da Pan, Ping Shi, Ming Hou, Zefeng Ying, Sizhe Fu, Yuan Zhang
Communication University of China
No.1 Dingfuzhuang East Street Chaoyang District, Beijing, China
{pdmeng, shiping, houming, yingzf, fusizhe, yzhang}@cuc.edu.cn

Abstract

A key problem in blind image quality assessment (BIQA) is how to effectively model the properties of human visual system in a data-driven manner. In this paper, we propose a simple and efficient BIQA model based on a novel framework which consists of a fully convolutional neural network (FCNN) and a pooling network to solve this problem. In principle, FCNN is capable of predicting a pixel-by-pixel similar quality map only from a distorted image by using the intermediate similarity maps derived from conventional full-reference image quality assessment methods. The predicted pixel-by-pixel quality maps have good consistency with the distortion correlations between the reference and distorted images. Finally, a deep pooling network regresses the quality map into a score. Experiments have demonstrated that our predictions outperform many state-of-the-art BIQA methods.

1. Introduction

Objective image quality assessment (IQA) is a fundamental problem in computer vision and plays an important role in monitoring image quality degradations, optimizing image processing systems and improving video encoding algorithms. Therefore, it is of great significance to build an accurate IQA model. In the literature, some full reference image quality assessment (FR-IQA) methods [19,20,23,27,30,32,33] which attempt to build a model simulating human visual system (HVS) can achieve good performance. For example, FSIM [33] predicts a single quality score from a generative similarity map (as shown in Fig. 1(b)). According to our analysis, two reasons bring FR-IQA methods into success. One reason is that it can access to reference image content and take the reference information by comparison. Meanwhile, this way of comparison is similar with the behavior of human vision and makes it easy to judge the image quality by FR-IQA methods [12]. The other reason is that hand-crafted features carefully designed by FR-IQA are closely related to some HVS methods properties. The difference of features on corresponding positions between reference and distorted images can well measure the distortion degree. On the other hand, some NR-IQA methods [16–18, 31] which rely on natural scene statistics do not obtain the same satisfying performance. As a result, the accuracies of most FR-IQA methods are better than those of NR-IQA when the performance is objectively evaluated.

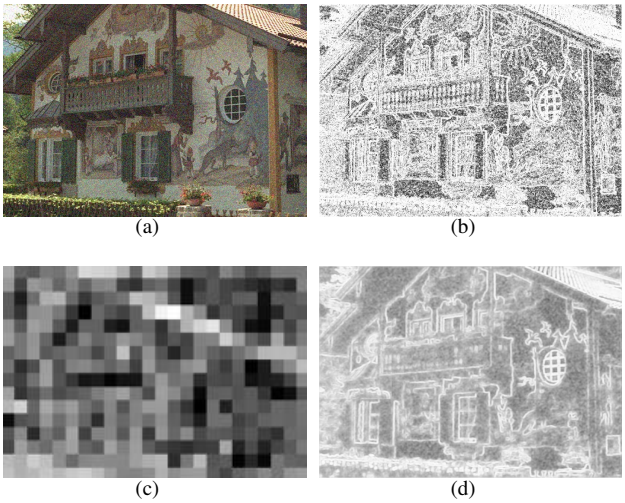


Figure 1. Examples of predicted quality maps: (a) is a distorted image; (b) is a similarity map from FSIM; (c) is a patch-based quality map from BIECON [9]; (d) is a pixel-based quality map predicted from our proposed model.

Based on these analysis, it is difficult for NR-IQA methods to build a model to imitate the behavior of HVS under the case of lacking reference information. Recently, researchers have started to harness the power of convolutional neural networks (CNNs) to learn discriminative features for various distortions types [1–3, 8]. We name these methods Deep-IQA. Most previous Deep-IQA methods consider CNN as a complicated regression function or a feature extractor, but are unaware of the importance of generating intermediate quality maps which represent the perceptual impact of image quality degradations. This training process of

deep-IQA seems not to have an explicit perceptual meaning and is always a black box for researchers. But what interests us is that BIECON [9] proposed an idea that training a CNN to replicate a conventional FR-IQA such as SSIM [27] or FSIM [33]. However, the method estimates each local patch score and patch-wise scores are pooled to an overall quality score. In essence, it visualizes a score patch map which contains spatial distribution information and the map does not reflect the distorted image in pixel level, as shown in Fig. 1(c). But we consider that the distortion value of each pixel is affected by its neighboring pixels and should not be exactly the same in the same patch. The simple patch-based scheme is not enough to correlate well with perceived quality. Therefore, how to design an effective deep learning model for blind predicting an overall pixel-by-pixel quality map related to human vision is the focus of this work.

In this paper, we propose a new deep-IQA model which consists of a fully convolutional neural network (FCNN) and a deep pooling network (DPN). We refer to this method as Blind Predicting Similar Quality Map for IQA (BPSQM). Specifically, given a similarity index map label, our proposed model can produce a HVS-related quality map to approach to the similarity index map in pixel distortion level. The predicted quality map can be a measurement map for describing the distorted image. Intuitively, the FCNN tries to simulate the process of FR-IQA methods generating similarity index maps. Then, given a subjective score label, the DPN which can be equivalent to various complicated pooling strategies predicts a global image quality score based on the predicted quality map. The primary advantage for this model is that the additional similarity map label guides FCNN to learn local pixel distortion features in the intermediate layers. Our proposed model considers assessing image quality as a problem of image-to-image. The quality maps predicted from BPSQM can reflect distorted areas in pixel level. Meanwhile, our model is simple and effective.

Our key insight is that good guided learning policies can help NR-IQA methods accurately predict global similar quality maps which agree with the distortion distribution between reference and distorted images. We use HVS-related similarity index maps derived from FR-IQA methods to navigate the learning direction of FCNN. Through guided learning, FR-IQA methods can transmit HVS-related pixel distortion feature information to NR-IQA methods. Fig. 1(d) shows a generative quality map from BPSQM. Compared to the patch-based quality map in (c), it is obvious that (d) represents pixel-wise distortions for a global distorted image. Meanwhile, the distortion distribution is generally similar with the feature map (b) from FSIM. In addition, a deep pooling network used for predicting the perceptual image quality is superior to other pooling strategies.

2. Related Work

2.1. Full-reference Image Quality Assessment

In order to effectively model the properties of HVS, many HVS-related methods have been proposed. The structure similarity index (SSIM) [27] extracted the structural, contrast and luminance information to constitute a similarity index map for assessing the perceived image quality. In [33], Zhang *et al*. proposed a feature-similarity index which calculated the phase congruency (PC) and gradient magnitude (GM) as features for the HVS perception. [30] proposed an efficient and effective standard deviation pooling strategy, which demonstrates that the image gradient magnitude alone can still achieve high consistency with the subjective evaluations. [19] used a novel deviation pooling to compute the quality score from the new gradient and chromaticity similarities, which further suggests that the gradient similarity could well measure local structural distortions. The aforementioned FR-IQA methods first compute a similarity index map to represent some properties of HVS and then design a simple pooling strategy to convert the map into a single quality score.

2.2. No-reference Image Quality Assessment

Many NR-IQA approaches model statistics of natural images and exploit parametric variation from this model to estimate perceived quality. DIIVINE [18] framework identified the distortion type firstly and applied a distortion-specific regression strategy to predict image quality degradations. BLIINDS-II [25] presented a Bayesian inference model to give image quality scores based on a statistical model of discrete cosine transform (DCT) coefficients. The CORNIA [31] learned a dictionary from a set of unlabeled raw image patches to encode features, and then adopted a max pooling scheme to predict distorted image quality. NIQE [17] used a multivariate Gaussian model to obtain features which are used to predict perceived quality in an unsupervised manner. SOM [34] focused on areas with obvious semantic information, where the patches from the object-like regions were input to CORNIA.

2.3. Deep Image Quality Assessment

With the rise of CNN for detection and segmentation tasks [4, 13, 14], more and more researchers have started to apply the deep network into IQA. Lu *et al*. [15] proposed a multi-patch aggregation network based on CNN, which integrates shared feature learning and aggregation function learning. Kang *et al*. [8] constructed a shallow CNN only consisting of one convolutional layer to predict subjective scores. [2] proposed a deeper network with 10 convolutional layers for IQA. [28] employed ResNet [5] to extract high-level features to reflect hierarchical degradation. Most deep-IQA methods only employ CNN to extract discrimi-

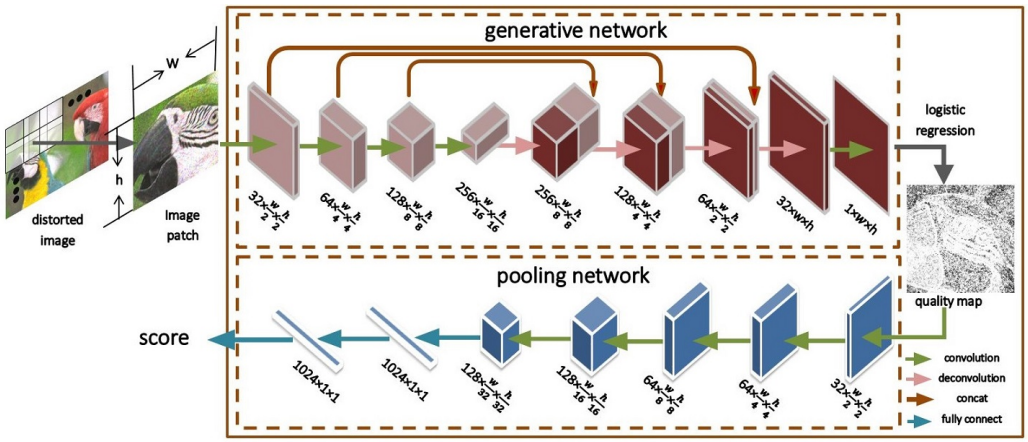


Figure 2. Architecture of the proposed BPSQM framework. The generative network takes as input a distorted image and predicts a similar quality map related to human vision. The pooling network directly regresses the generative quality map into a score.

native features and are inadequate for analyzing and visualizing the intermediate results, which makes it difficult for us to understand how to process IQA based on CNN. In [7], Kim *et al*. proposed a full reference Deep-IQA generating a perceptual error map which provides us an intuitive analysis of local artifacts for given distorted images. BIECON [9] utilized a CNN to estimate a patch score map and utilized one hidden layer to regress the extracted patch-wise features into a subjective score.

3. Quality Map Prediction

Problem formulation

Given a color or gray distorted image I_d, our goal is to estimate its quality score by modeling image distortions. Previous works for deep NR-IQA [2, 8] lets f_θ be a regression function using CNN with parameter θ. S indicates the subjective ground-truth score:

$$S = f_\theta(I_d) \tag{1}$$

In this case, the deep network simply trains on input images and directly outputs results. From the process, we can not understand how the deep network learns features related with the image distortion. In contrast, FR-IQA methods generate similarity index map firstly and then pool the map. The process can be formulated as below:

$$S = P(M(I_d, I_r)) \tag{2}$$

Where I_r represents a reference image. M indicates the way to calculate similarity index map and P denotes a pooling strategy, for example, P can be a simple average operation in SSIM or a standard deviation operation in GMSD. Given all that, we combine the advantage of FR-IQA modeling general properties of HVS with the advantage of NR deep-IQA without hand-crafted features. Our approach is first to construct a generative network G with parameters ω to predict a global quality map in pixel level. Then, a deep pooling network f_ϕ regarded as a complicated pooling strategy converts the predicted quality map into a score.

$$S = f_\phi(G_\omega(I_d)) \tag{3}$$

3.1. Architecture

The proposed overall framework is illustrated in Fig. 2. This framework consists of two main components: a generative quality map network and a quality pooling network. The requirement for the generative network is outputting a quality map of the same size with the input image. We select U-Net [24], an extension of FCNN, as a base of generative network. Because U-Net integrates the hierarchical representations in subsampling layers with the corresponding features in upsampling layers. So the degradations on both the low and high level features are considered for IQA [28].

The generative network consists of a subsampling path (SP) and an upsampling path (UP). In the SP, the distorted image goes through four convolutional layers with kernel 3×3 and padding 1×1. In the UP, there are also four corresponding deconvolution layers with kernel 2×2 and stride 2×2. The feature maps in the SP are contacted with the corresponding feature maps of the same size in UP. The last deconvolution layer outputs a pixel-wise dense prediction map with the same size as the input image. Batch normalization [6] and leaky rectified linear unit (LReLU) are used after all convolution and deconvolution operations. A 3×3 convolutional layer with padding 1×1 for keeping same size is used for reducing dimensionality into one channel feature map. U-Net can be trained with a pixel-wise logist loss against the ground-truth map. The output of logist layer is directly passed into the subsequent network. In our paper, the pooling network contains five 3×3 convolutional layers with 2×2 maximum pooling and two fully connected layers.

We perform 50% dropout after each fully connected layers so as to prevent overfitting. The pooling network ends up with a squared Euclidean loss layer. It should be noted that we crop the input image into some overlapping fixed size patches so as to adapt to the pooling network. This patch size should be large enough, which will not influence the learning of pixel distortion.

3.2. Quality Map Selection

SSIM [27], FSIM [33] and MDSI [19] are adopted to generate similarity maps as label separately. Because the luminance, contrast and structural information are treated equally in SSIM, the similarity map derived from SSIM is directly used as map label. In contrast, the FSIM method uses a pooling weight to combine the phase congruency (PC) and gradient magnitude (GM) in computing the final quality score. So we select the two features as map label separately. As for MDSI [19], the combination of gradient and chromaticity similarity maps is selected as label.

We remove pre-processing including filtering and downsampling in the process of computing the similarity index map label to guarantee the generative map same size with the input image. Specially, for SSIM, owing to the input images processed with a kernel 11×11 Gaussian filter, it leads to less 5 pixels near borders around the similarity map. To guarantee image alignment, we exclude each 5 rows and columns for each distorted image border before training SSIM labels.

3.3. Multi Types Quality Maps Fusion

For each FR-IQA method, we will train U-Net separately. Many conventional FR-IQA methods [19, 33] have demonstrated that multi complementary features are combined to increase the prediction accuracy for image quality. Thus, we also fuse the information from predicted multi types quality maps to feed into the pooling network. Different pooling strategies of quality maps are experimented:

-single pooling stream is performed by concatenating different type quality maps into a multi channels quality map followed by a single pooling network (shown in Fig. 3 (a)).

-multi pooling streams indicate that each type quality map is fed into an independent pooling network. The last convolutional layers of the pooling networks are concatenated, followed by two full connected layers (shown in Fig. 3 (b)).

3.4. Regression

The input to U-Net is an RGB patch of fixed size 144×144×3 sampled from a distortion image without any image pre-processing. We set the step of the sliding window to 120, i.e. the neighboring patches are overlapped by

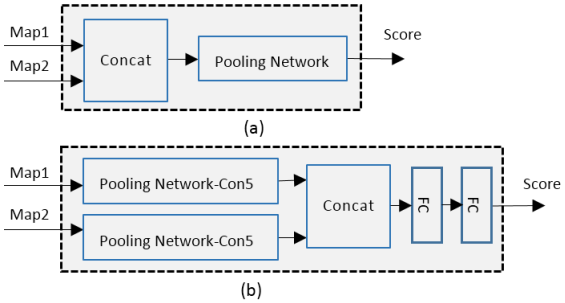


Figure 3. Different pooling strategies to combine multi types quality maps: single pooling stream (a), multi pooling streams (b).

24 pixels, which can compensate partial distorted area continuous. Considering that the patch size is large enough to reflect the overall image quality, we set the quality score of each patch to its distorted images subjective ground-truth score. The proposed pooling network is to conduct nonlinear regression from the predicted quality map to the subjective score. To compare the performances of different network structures as regression function, we also test a simple network with only two fully connected layers and ResNet [5]. Then, the final objective function is defined as:

$$L_s(I_d;\phi,\omega) = ||(f_\phi(G_\omega(I_d)) - Eva)||^2 \qquad (4)$$

Where *Eva* denotes the human evaluation for the input distorted image. The final score of a global distorted image is averaging the cropped patches.

3.5. Training Method

The proposed network was implemented in MXNet. By outputting the intermediate results, we find that the quality maps from individually training U-Net are more close to the similarity index maps than those from the joint training of the overall framework. Thus, we first only train the generative network on similarity maps, and then fix its parameters in training process of the overall framework.

Our network was trained end-to-end by back-propagation. For optimization, the adaptive moment estimation optimizer (ADAM) [10] is employed with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$ and $\alpha = 10^{-4}$. We set an initially learning rate to 1×10^{-3} and 5×10^{-3} for the generative network and the pooling network, respectively. We set the weight decay to 1×10^{-11} for all layers to help prevent overfitting. To evaluate the performance on IQA databases, each database is randomly divided into 80% for training and 20% for testing by reference images, which ensures that the content of images in test sets never exists in train sets. We only use the horizontally flip operation to expand training data, for general data argumentation skills, such as rotation, zoom, contrast, will affect the final image quality. The models are trained for 100 epochs and we choose the model with the lowest validation error.

The partition was repeated 100 times to eliminate the bias caused by data division.

4. Experiments

4.1. Datasets

Four datasets are employed in our experiments to validate the performance of the proposed method including LIVE [26], CSIQ [11], TID2008 [22] and TID2013 [21]. LIVE consists of 982 distorted images with 5 different distortions: white Gaussian noise (WN), Gaussian blur (BLUR), JPEG, JPEG2000 (JP2K) and fast-fading distortion (FF). Each image is associated with Differential Mean Opinion Scores (DMOS) in the range [0, 100]. The CSIQ database includes 30 original images and 866 distorted images with 6 distortion types at four to five different levels of distortion. It is reported in the form of DMOS which are normalized to span the range [0, 1]. TID2008 contains 25 reference images and a total of 1700 distorted images with 17 distortion types at 4 degradation levels. TID2013 is an extension of TID2008 and includes seven new types and one more level of distortions. Mean Opinion Scores (MOS) are provided for each image in the range [0, 9]. Owing to more distortion types and images, the TID2013 is more challenge for researchers in the four databases.

To evaluate the performance of our model, two widely applied correlation criterions are applied in our experiments including the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SRCC).For both correlation metrics, a higher value indicates higher performance of a specific quality metric. The MOS values of TID2013 and the DMOS values of CSIQ have been linearly scaled to the same range [0, 100] as the DMOS values in LIVE.

4.2. Quality Map Prediction

To validate if BPSQM is consistent with human visual perception, the intermediate generative quality maps and their corresponding similarity map labels are shown in Fig. 4. The first column in Fig. 4 shows three different distortion types from TID2013, including JPEG, high frequency noise (HFN) and local block-wise distortions (LBWD). The remaining columns correspond to SSIM, the gradient magnitude of FSIM (Fg) and MDSI, respectively. A(1-3), B(1-3) and C(1-3) are the ground-truth map labels. A(4-6), B(4-6) and C(4-6) are the predicted quality maps. The dark areas indicate distorted pixels. Overall, the generative quality maps are similar with ground-truth maps on distorted degrees and areas. In case of JPEG distortion, the artifact edges caused by compression on the root are clearly shown in A(5) and A(6). But owing to SSIM similarity index maps emphasizing local structure features, this leads to some areas in the predicted quality map to be smooth without local pixels distortion, as shown in A(4). For HFN, noises spread over an overall distorted image. B(4-6) not only display the uniform distribution well but also give a clear airplane profile. LBWD is a very challenging distortion type for most BIQA methods to distinguish additional blocks and undistorted regions. Even though some wrong pixel distortion predictions appear in the undistorted regions, as shown in C(4-6), each predicted block is darker than other areas. Meanwhile, the predicted positions of local blocks can agree with those of the map labels.

In Fig. 5, the predicted quality maps of spatially correlated noise and JPEG with different distortion levels are shown. The first row denotes the spatially correlated noise, and the second row denotes the JPEG. With the noises becoming strong gradually from left to right, the predicted quality maps grow darker and darker as shown in (a)-(e). Meanwhile, when the degree of JPEG compression increases, the blocking artifact on the sculpture area was emphasized in (j). Generally, with the degree of distortion increasing, the predicted scores gradually decrease, which suggests that BPSQM predicts good pixel-by-pixel quality maps agreeing with the distortion correlations between the reference and distorted images.

4.3. Dependency on FR-IQA Similarity Map

To validate the feasibility and effectiveness of directly pooling quality maps, we compare pooling ground-truth FR-IQA maps with pooling predicted quality maps. We choose the more challenging full TID2013 database in this experiment. Here, directly pooling ground-truth map labels from SSIM and the gradient magnitude of FSIM are denoted by S_LB and Fg_LB, respectively. Directly pooling predicted maps of SSIM is denoted by S_PM. The gradient magnitude computed from FSIM is denoted by Fg_PM. We also feed the distorted image into the pooling network directly, named as D_LB. The results are shown in Table 1, we can see that the S_LB and Fg_LB both perform better than their original methods, especially for the SRCC of SSIM increasing from 0.637 to 0.904, which suggests that the deep pooling network can better fit a quality map to a subjective score than simple averaging. The D_LB performs worse than Fg_PM. We consider the primary reason is that distorted images contain too much redundant information and do not highlight distorted distribution features. Even though the deep network has strong ability of extracting discriminative features, it is still not enough to accurately present distorted patterns. For this reason, we need to firstly predict similar quality maps which correctly reveal the distorted areas and degrees.

4.4. Similarity Map Labels Comparison

To investigate the performance of different FR-IQA ground-truth maps, the similarity maps derived from SSIM,

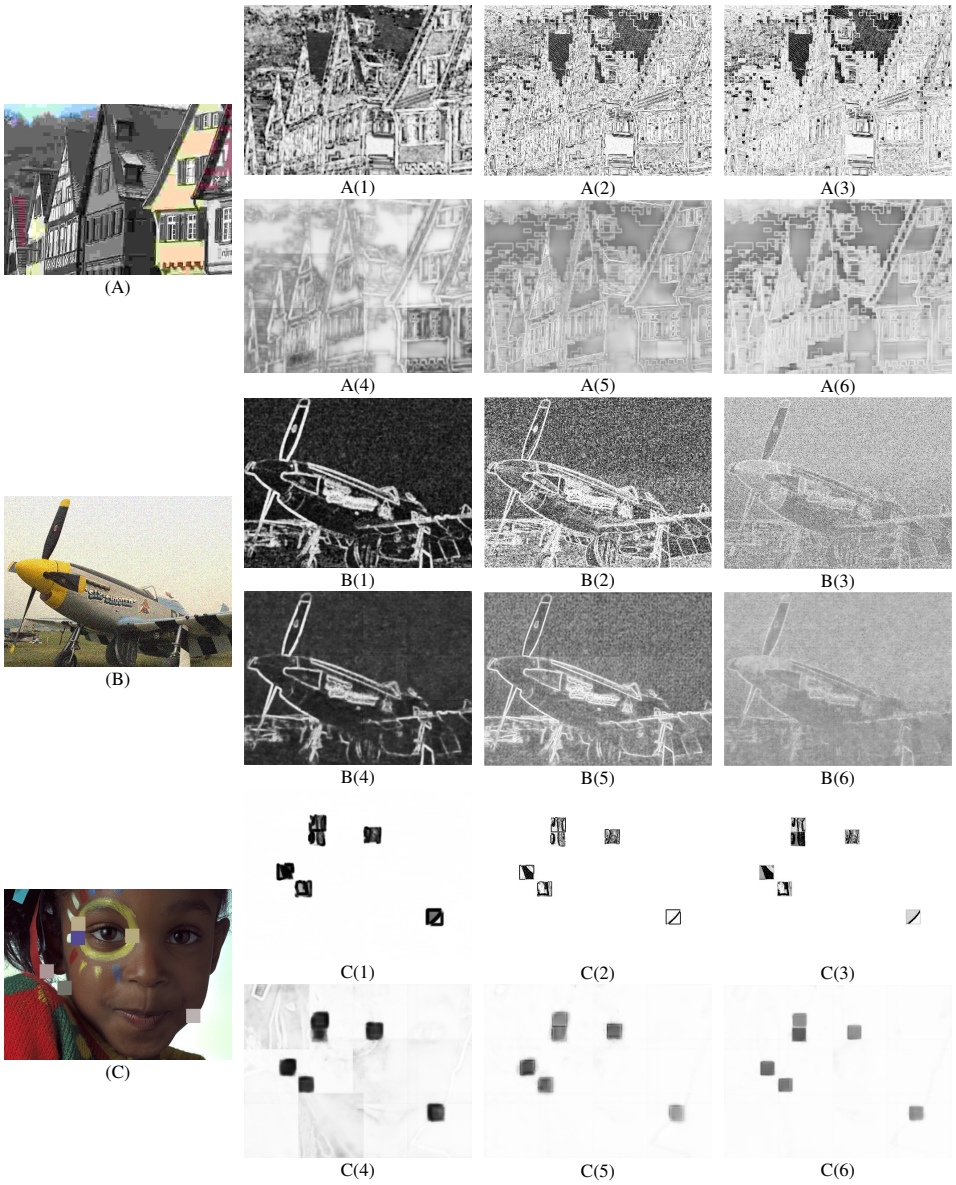


Figure 4. Predicted quality maps and ground-truth similarity maps: (A), (B) and (C) are distorted images with JPEG, HFN, LBWD, respectively; the second, the third and the forth columns indicate three FR-IQA methods which are SSIM, Fg and MDSI, respectively. A(1-3), B(1-3) and C(1-3) are ground-truth similarity maps. A(4-6), B(4-6) and C(4-6) are predicted quality maps.

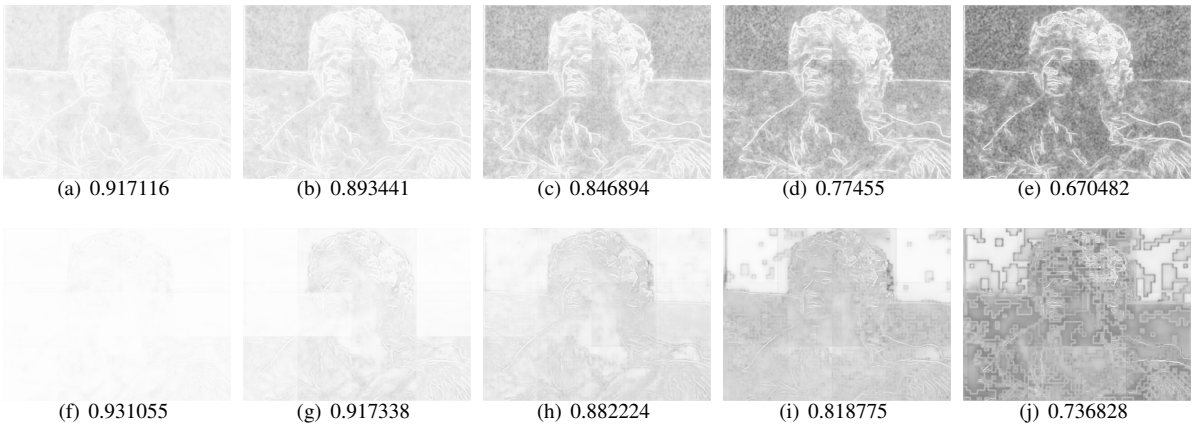
Table 1. SRCC and PLCC comparison for pooling ground-truth FR-IQA maps and pooling predicted quality maps

	NR			FR			
	D_LB	S_PM	Fg_PM	S_LB	Fg_LB	SSIM	FSIMc
SRCC	0.781	0.758	0.828	0.904	0.923	0.637	0.851
PLCC	0.837	0.803	0.856	0.913	0.930	0.691	0.877

Table 2. SRCC and PLCC comparison for each predicted quality map from different FR-IQA methods on the TID2013 database

	No_LB	S_PM	Fg_PM	Fp_PM	MD_PM
SRCC	0.736	0.758	0.828	0.723	0.863
PLCC	0.779	0.803	0.856	0.789	0.879

FSIM and MDSI were respectively chosen as labels for training the model. The TID2013 database with all distortion types was applied in this experiment. The combination of gradient and chromaticity similarity maps from MDSI are referred to MD_PM. The phase congruency (PC) from FSIM is denoted by Fp_PM. In order to analyze the effect of removing FR-IQA similarity maps, we directly employ the overall framework to perform an end-to-end training without any FR-IQA map labels, referred to No_LB.



(a) 0.917116 (b) 0.893441 (c) 0.846894 (d) 0.77455 (e) 0.670482

(f) 0.931055 (g) 0.917338 (h) 0.882224 (i) 0.818775 (j) 0.736828

Figure 5. Examples of predicted quality maps with various distortion levels of spatially correlated noise, and JPEG: (a)-(e) are distorted by spatially correlated noise; (f)-(j) are distorted by JPEG. The values indicate the predicted scores output from the pooling network. Smaller values indicate higher distortions.

The final SRCC and PLCC values are shown in Table 2, No_LB achieves worse performance among the results except for Fp_PM. Clearly, selecting FR-IQA similarity maps for training the generative network can help to learn a better model for predicting image quality, because the task improves the ability of U-Net learning discriminative features about distortion. In particular, Fp_PM achieves unfavorable performance and seems not to fit with this framework. Because when it comes to the the ability in describing local pixel distortion, the selected quality map of PC from FSIM is less than the quality map of gradient features from FSIM. In contrast, Fg_PM achieves the second rank, suggesting that gradient distortion variations learned from the U-Net is more suitable than phase congruency to apply into the framework. MD_PM performs the best result, suggesting that the chromaticity feature can be complementary to gradient features.

4.5. Effects of Pooling Network

To investigate the pooling performance of convolutional network with different depth and full connected layers. We experimented with three pooling network structures. The first network is the pooling network proposed by this paper, called DPN. The second network only contains two full connected layers with 1024 neurons each, called FC2. ResNets [5] with 18 layers and 50 layers are selected as the third and the forth pooling network. The results are shown in Table 3. The DPN performs better than FC2, which indicates that the additional convolutional layers have the strong ability of pooling quality maps. Although ResNet performs better than shallow networks on image classification and recognition, the deeper network is no necessarily to use at this experiment.

Table 3. SRCC and PLCC comparison for different pooling networks using Fg_PM on the TID2013 database

	DPN	FC2	ResNet18	ResNet50
SRCC	0.828	0.707	0.787	0.795
PLCC	0.856	0.711	0.833	0.840

Table 4. SRCC and PLCC comparison for different fusion schemes and multi predicted quality maps combinations on TID2013

		Fg_MD	S_Fg	MD_S	S_Fg_MD
Single stream	SRCC	0.862	0.825	0.853	0.855
	PLCC	0.885	0.859	0.873	0.880
Multi streams	SRCC	0.842	0.821	0.825	0.834
	PLCC	0.873	0.854	0.861	0.868

4.6. Quality Maps Fusion

We evaluate two different fusion schemes for combining multi types predicted quality maps. The first is a single pooling stream scheme (shown in Fig. 3(a)) and the second is multi pooling streams (shown in Fig. 3(b)). The results are given in Table 4. We can see that the single pooling stream is better. This further illustrates that, compared with the feature maps output from the pooling network-con5, the quality maps directly output from U-Net are more in consistent with human vision.

In the 4.4 section, S_PM, MD_PM and Fg_PM rank the top three. In order to compare multi type quality maps combinations, we design all possible combinations which contain any two types and all types. As we can see from Table 4, the combination of more types seems not yield significant performance gains. Moreover, the predicted SSIM quality map combined with other quality maps caused a little performance degradation. Since the gradient operator is different in MDSI and FSIM, the two joint achieves the best performance, which indicates that gradient features are

Table 5. Performance comparison on the whole database (LIVE, CSIQ and TID2013). Italics indicate our proposed model.

Type		FR			NR								
Method		SSIM	FSIMc	MDSI	DIIVINE	BRISQUE	NIQE	IMNSS	HFD-BIQA	BIECON	DIQaM	*BPSQM-Fg*	*BPSQM-MD*	*BPSQM-Fg-MD*
LIVE IQA	SRCC	0.948	0.960	0.966	0.892	0.929	0.908	0.943	0.951	0.958	0.960	0.971	0.967	**0.973**
	PLCC	0.945	0.961	0.965	0.882	0.920	0.908	0.944	0.948	0.960	**0.972**	0.961	0.955	0.963
CSIQ	SRCC	0.876	0.931	0.956	0.804	0.812	0.812	0.825	0.842	0.815	-	0.862	0.860	**0.874**
	PLCC	0.861	0.919	0.953	0.776	0.748	0.629	0.789	0.890	0.823	-	0.891	0.904	**0.915**
TID2013	SRCC	0.637	0.851	0.889	0.643	0.626	0.421	0.598	0.764	0.717	0.835	0.828	**0.863**	0.862
	PLCC	0.691	0.877	0.908	0.567	0.571	0.330	0.522	0.681	0.762	0.855	0.856	0.879	**0.885**
Weighted Avg.	SRCC	0.743	0.887	0.917	0.722	0.721	0.589	0.708	0.816	0.783	-	0.863	0.884	**0.887**
	PLCC	0.773	0.902	0.928	0.668	0.673	0.500	0.655	0.772	0.813	-	0.884	0.899	**0.906**

Table 6. SRCC comparison on individual distortion types on the LIVE IQA and TID2008 databases. Italics indicate deep learning-based methods.

Method	LIVE IQA					TID2008											
	JP2K	JPEG	WN	BLUR	FF	AGN	ANMC	SCN	MN	HFN	IMN	QN	GB	DEN	JPEG	JP2K	JGTE	J2TE
SSIM	0.961	0.972	0.969	0.952	**0.956**	0.811	0.803	0.792	0.852	0.875	0.700	0.807	0.903	0.938	0.936	0.906	0.840	0.800
GMSD	0.968	0.973	**0.974**	0.957	0.942	**0.911**	**0.878**	**0.914**	0.747	0.919	0.683	0.857	0.911	**0.966**	**0.954**	**0.983**	0.852	0.873
FSIMc	**0.972**	**0.979**	0.971	**0.968**	0.950	0.910	0.864	0.890	**0.863**	**0.921**	**0.736**	**0.865**	**0.949**	0.964	0.945	0.977	**0.878**	**0.884**
BLIINDSII	0.929	0.942	0.969	0.923	0.889	0.779	0.807	0.887	0.691	0.917	0.908	0.851	0.952	0.908	0.928	0.940	0.865	**0.855**
DIIVINE	0.937	0.910	0.984	0.921	0.863	0.812	0.844	0.854	0.713	0.922	0.915	0.874	0.943	0.912	0.930	0.938	0.873	0.852
BRISQUE	0.914	0.965	0.979	0.951	0.877	0.853	0.861	0.885	0.810	0.931	0.927	0.881	0.933	**0.924**	0.934	0.944	**0.891**	0.836
NIQE	0.914	0.937	0.967	0.931	0.861	0.786	0.832	0.903	0.835	0.931	0.913	0.893	**0.953**	0.917	0.943	0.956	0.862	0.827
BIECON	0.952	**0.974**	0.980	0.956	0.923	0.913	0.835	0.903	0.835	0.931	0.913	0.893	**0.953**	0.917	0.943	0.956	0.862	0.827
BPSQM-Fg-MD	0.969	0.946	**0.993**	**0.986**	0.960	0.881	0.801	0.935	0.786	0.938	**0.933**	**0.920**	0.937	0.914	0.943	**0.967**	0.829	0.644
BPSQM-MD	**0.972**	0.929	0.985	0.977	**0.964**	**0.923**	**0.880**	**0.941**	**0.948**	**0.948**	0.892	0.909	0.908	0.878	**0.950**	**0.967**	0.836	0.756

complementary to each other.

4.7. Performance Comparison

In Table 5, the proposed BPSQM is compared with 7 state-of-the-art NR-IQA methods (DIIVINE [18], BRISQUE [16], NIQE [17], IMNSS [29], HFD-BIQA [28], BIECON [9] and DIQaM [2]) and 3 FR-IQA methods (MDSI [19], SSIM [27], FSIM [33]). All distortion types are considered over the three databases. The best PLCC and SRCC for the NR IQA methods are highlighted. The weighted average in the last column is proportional to the number of distorted images of each database. We can see that the BPSQM obtains superior performance to state-of-the-art BIQA methods, except for DIQaM evaluated by PLCC on LIVE. Especially for the challenging TID2013, BPSQM achieves a remarkable improvement against BEICON. It is obvious that predicted global quality maps in pixel level helps the model extract more useful features to achieve a good accuracy. Meanwhile, the BPSQM-Fg-MD achieves competitive performance to some FR-IQA methods.

Table 6 shows the SRCC performance of the competing BIQA methods for individual distortion type on LIVE and TID2008 database. The best results are in bold. In general, BPSQM-MD and BPSQM-Fg-MD achieve the competitive performances among most distortion types on the two databases. Compared with BIECON, BPSQM is more capable in dealing with the distortion of SCN, HFN, QN and JP2K. By contrast, for the distortions of GB, JGTE and J2TE, BEICON performs better than BPSQM-MD and BPSQM-Fg-MD. Moreover, the achieved scores on some distortion types are close to state-of-the-art FR-IQA methods.

Table 7. SRCC comparison of the models trained using LIVE IQA database and tested on the TID2008 database

	Metrices	JP2K	JPEG	WN	BLUR	ALL
FR	SSIM	0.963	0.935	0.817	0.960	0.902
NR	BRISQUE	0.832	0.924	0.829	0.881	0.896
	BIECON	0.878	**0.941**	0.842	**0.913**	**0.923**
	BPSQM-Fg-MD	**0.947**	0.909	**0.886**	0.874	0.910

4.8. Cross Database Test

To evaluate the generalization ability of BPSQM, we trained it on the LIVE IQA database and tested it using the TID2008 database. Since the TID2008 includes more distortion types, we only chose the common types between the two databases, including JP2K, JPEG, WN and BLUR. Table 7 shows that BPSQM performs well on the four distortion types. These results suggest that our proposed method does not depend on the database and shows good generalization capabilities.

5. Conclusion

In this paper, we developed a simple yet effective blind predicting quality map for IQA that generates the map in pixel distortion level under the guidance of similarity maps derived by FR-IQA methods. Meanwhile, we also compare how to fuse the multi features information for predicting image quality. We believe that this proposed model could achieve better performance if where there is a better similarity index map navigating the generative network training. Optimizing our generative network to predict more accurate pixel distortion is a potential direction for future work.

Acknowledgement This work was supported by the National Natural Science Foundation of China under Contract 61472389 and by "Double Tops" construction project.

References

[1] S. Bianco, L. Celona, P. Napoletano, and R. Schettini. On the use of deep learning for blind image quality assessment. *arXiv preprint arXiv:1602.05531*, 2016. 1

[2] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 2017. 1, 2, 3, 8

[3] S. Bosse, D. Maniry, T. Wiegand, and W. Samek. A deep neural network for image quality assessment. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3773–3777. IEEE, 2016. 1

[4] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. *IEEE International Conference on Computer Vision*, 2017. 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *computer vision and pattern recognition*, pages 770–778, 2016. 2, 4, 7

[6] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *international conference on machine learning*, pages 448–456, 2015. 3

[7] S. L. Jongyoo Kim. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[8] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 1, 2, 3

[9] J. Kim and S. Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):206–220, 2017. 1, 2, 3, 8

[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[11] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010. 5

[12] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng. Image quality assessment using similar scene as reference. In *European Conference on Computer Vision*, pages 3–18. Springer, 2016. 1

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *arXiv preprint arXiv:1612.03144*, 2016. 2

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 2

[15] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *IEEE International Conference on Computer Vision*, pages 990–998, 2015. 2

[16] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. 1, 8

[17] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 1, 2, 8

[18] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 1, 2, 8

[19] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M. Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016. 1, 2, 4, 8

[20] S.-C. Pei and L.-H. Chen. Image quality assessment using human visual dog model fused with random forest. *IEEE Transactions on Image Processing*, 24(11):3282–3292, 2015. 1

[21] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. 5

[22] 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. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009. 5

[23] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *arXiv preprint arXiv:1607.06140*, 2016. 1

[24] 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*, pages 234–241. Springer, 2015. 3

[25] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 2

[26] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 5

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1, 2, 4, 8

[28] J. Wu, J. Zeng, Y. Liu, G. Shi, and W. Lin. Hierarchical feature degradation based blind image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517, 2017. 2, 3, 8

[29] X. Xie, Y. Zhang, J. Wu, G. Shi, and W. Dong. Bag-of-words feature representation for blind image quality assessment with local quantized pattern. *Neurocomputing*, 2017. 8

[30] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014. 1, 2

[31] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1098–1105. IEEE, 2012. 1, 2

[32] L. Zhang, Y. Shen, and H. Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014. 1

[33] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 1, 2, 4, 8

[34] P. Zhang, W. Zhou, L. Wu, and H. Li. Som: Semantic obviousness metric for image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2394–2402, 2015. 2