Blind Image Quality Assessment with Semantic Information

Weiping Ji[a], Jinjian Wu[a,*], Guangming Shi[a], Wenfei Wan[a], Xuemei Xie[a]

[a]*School of Artificial Intelligence, Xidian University, Xi'an, Shaanxi 710071, PR China*

Abstract

No-reference (NR) image quality assessment (IQA) aims to evaluate the quality of an image without reference image, which is greatly desired in the automatic visual signal processing system. Distortions degrade the visual contents and affect the semantics acquisition during the process of human perception. Although the existing methods evaluate the quality of images based on the structure, texture, or statistical characteristics, and deliver high quality prediction accuracy, they do not take the spatial semantics into account. From the perspective of human perception, distortions decrease the structural semantics that represent the structural information, and disturb the spatial semantics that describe the contents of images. Therefore, we attempt to measure the image quality by its degradation of semantics in an image. To extract the semantics of an image, a semantic network is proposed. The network contains convolutional neural networks (CNN) and Long Short-Term Memory (LSTM) that correspond to structural semantics and spatial semantics, respectively. CNN can be regarded as a coarse imitation of human visual mechanism to obtain the structural information, and LSTM can express the contents of an image. Then, by measuring the degradations of different semantics on images, a novel NR IQA is intro-

[☆]This work was partially supported by the Joint fund of the Ministry of Education (6141A020336), the NSF of China (Nos. 61772388,61632019, 61621005, 61472301), the Young Star Science and Technology Project (No. 2018KJXX-030) in Shanxi province.

[*]Corresponding author

Email addresses: weipingjileo@163.com (Weiping Ji), jinjian.wu@mail.xidian.edu.cn (Jinjian Wu), gmshi@xidian.edu.cn (Guangming Shi), wenfei.wan@stu.xidian.edu.cn (Wenfei Wan), xmxie@mail.xidian.edu.cn (Xuemei Xie)

duced. The proposed approach is evaluated on the databases of LIVE, CSIQ, TID2013, and LIVE multiply distorted database as well as LIVE in the wild image quality challenge database, and the results show superior performance to other state-of-the-art NR IQA methods. Furthermore, we explore the generalization capability of the proposed approach, and the experimental results indicate the proposed approach has a high robustness.

Keywords: No-reference image quality assessment, human perception, semantic network, structural semantics, spatial semantics.

1. Introduction

In the past two decades, images have been widely used as a mode of information description and information exchange. However, in the process of image acquisition, transmission, processing and storage, image inevitably suffers from different types and degrees of distortions. All those will cause a decline in the quality of the image which affects people's subjective feelings and information acquisition. Therefore, it is essential to assess its perceived quality in image communication and processing. The most reliable method of image quality assessment (IQA) is human subjective judgment, but it is usually time-consuming, expensive and not real-time. Hence, objective image quality evaluation is introduced at the right moment that can automatically predict image quality that is consistent with human subjective perception. Objective IQA methods are divided into three categories: full reference (FR), reduced reference (RR), and no reference (NR) [1, 2]. The FR IQA needs a full reference that is considered to be distortion-free or perfect quality in evaluating a distorted image [3, 4]. For RR IQA, certain features are extracted from the reference image instead of the full reference [5, 6]. In many practical applications, reference images are not often obtained. Therefore, it is urgent to develop a method that can evaluate image quality blindly. As no information about the primary is obtained, NR IQA [7, 8, 9, 10] is a more difficult problem, and has more practical significance.

Early NR methods mainly depend on the hand-crafted features [8, 11, 12],

which rest heavily on the designer's subjective understanding of the images. Most algorithms rely mainly on human visual system (HVS) [13] or Natural Scene Statistical (NSS) [14]. HVS-based models imitate the process of the eye gaining information based on visual attention [1], contrast sensitivity [15], and masking [1]. NSS-based models seek to capture those statistical properties that represent the distributions of certain filter responses in several different domains (i.e., spatial, wavelet, and DCT domains) [16, 11, 10]. These methods based on the texture, structure, or statistical characteristics of the whole image have made great progress, however, they are far from the human subjective perception [17].

In recent years, convolutional neural networks (CNN) are used in many vision tasks, and show a great improvement in performance as well [18, 19, 20]. A number of attempts have been made to find out whether CNN is suitable for IQA. The existing methods are mainly tried in three ways, the first one adopts the architectures and weights from the network applied to classification task followed by fine-tuning [21, 18]. CNN transforms the original image into a higher level and more abstract expression, but this is dependent on the task of classification that describes the category of images. The second kind of methods deal with image patches by assigning the subjective differential mean opinion score (DMOS) of an image to all image patches [22]. The last one uses FR-IQA models for image patches annotations [23] that is different from the second one. There are two obvious disadvantages in the latter two methods. On the one hand, they ignore global information of images; on the other hand, the quality of the image patches is not well defined. The former ignores local image quality within context that varies across spatial locations even when the distortion is homogeneous [3], and similar statistical properties may have substantially different quality. The quality of image patches directly marked by the existing FR-IQA is inaccurate in itself [24] in the latter. More importantly, those methods are easy overfitting which may be inadequate across distortion levels [23] and distortion types [25].

Most images are identified and understood by human beings, but the dis-

tortions in images may affect human subjective perception. There are two necessary processes in most definitions of human perception, i.e., recognizing and interpreting [26]. In the recognition process, human beings perceive the structural information of images to identification, and this process is hierarchical and abstract [27]. Color and luminance are the first to be perceived in the recognition process, followed by local detail information such as edges, corners, and lines. After that, more complex information that correspond to parts of familiar objects is obtained, and subsequently the concept of objects is obtained from the combinations of these parts. This is verified by the Visual Neuroscience's research on visual mechanism [28]. CNN can be seen as a simple imitation of above mechanism. It can learn the distinguishing local structural information by cascading the convolutional layers and the pooling layers, combination of these local structural information through the fully connected layers, and then the advanced attributes of an image are obtained [29]. Therefore, CNN can be used to extract structural semantics which describe the structural information of images.

In the interpreting processing, the contents of the images are supposed to understand [26]. It shows that the semantics of images not only contain the structural informantion, but also express the contents of images as well as their attributes. Human beings naturally combine visual information with language systems which is confirmed by brain inspired visual computing theory [30]. Accordingly, the contents of an image which called the spatial semantics may be expressed through sentences. The task of machine translation is to transform a source language into the target language, which provides a feasible method to generate sentences to describing an image [31]. Recent work has shown that translation can be done in a simpler way by using Long Short-Term Memory (LSTM) which reaches state-of-the-art performance [32, 33, 34]. In order to generate a sentence to describe the contents of an image, the source language is replaced by images. Over the last few years it has been shown that CNNs can produce a rich presentation of the input image by embedding it to a fixed-length vector, and such this representation can be used for a variety vision tasks.

Hence, it is natural to use a CNN to extract a representation as an input to the LSTM to generate sentences.

In this work, we assess the quality of the images based on those semantics. The structural semantics display the effects of different types and different levels of distortions on the edges, texture and geometry of the image, while the spatial semantics describe the impact on image interpretation, and combination of those semantics can accurately predict the quality of distorted images with respect to human subjective perception.

Our contributions are as follows: first, we propose a new method that assesses the quality of images from a completely new perspective. Most of the existing models are traditionally based on bottom-up or top-down approaches that ignore human subjective perception. Second, to the best of our knowledge, this is the first attempt for integrating CNN and LSTM architecture together that employed in image quality assessment. Finally, our algorithm has a good robustness, that is confirmed by cross-database evaluation in the later.

The paper is structured as follows: In Section 2 , we give the detailed description of semantic information of images. Section 3 describes the application of semantic information in NR-IQA. Experimental evaluations and comparisons to other state-of-the-art methods as well as experimental analysis are presented in Section 4 . We conclude the paper with a discussion in Section 5 .

2. Semantic network

The semantics of an image not only capture the structural information contained in an image, but also express the contens as well as their attributes. These semantics represent the discriminative information in the process of human subjective perception. The former called structural semantics can be obtained by CNN through cascading the convolutional layers and the pooling layers. The latter called spatial semantics can be expressed by a sentence that can be done in a simpler way by using LSTM. Owing to the continuity of recognition and interpretation in the process of human subjective perception, the structural se-

mantics and spatial semantics are not present independently, and the spatial semantics are influenced by the structural semantics. Thus, the semantic network should include CNN and LSTM, and all the parameters optimized by end-to-end.

2.1. Semantic network architecture

2.1.1. The sub-network of structural semantic

The architecture of CNN has great impact on the extraction of structural semantic in an image, and different architectures can bring dissimilar structural semantics. Not only motivated by its superior performance in the 2014 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) classification and localization tasks [35], but also achieve excellent performance even when used as a part of relatively simple pipelines, VGGnet [19] may be a reasonable choice for the sub-network of structural semantic. VGGnet is the first network to employ cascaded convolutional kernels small as 3×3 and 1×1, and had deeper architectures. For instance, a stack of three 3×3 convolutional layers instead of a single 7×7 layer. In other words, VGGnet incorporates three non-linear rectification layers instead of a single one, which makes the sub-network have less parameters. The 1×1 convolutional layer is a way to increase the nonlinearity without affecting the receptive fields, which results in the structural semantics are more discriminative. According to these advantages, VGGnet is the most appropriate architecture for the sub-network of structural semantics.

2.1.2. The sub-network of spatial semantic

LSTM is a memory block that contains a cell c that is always used to sequence modeling. In this work, a LSTM [36] is used to generate one word at every time which is based on the hidden state and generated words. Thus, we can obtain a sentence called spatial semantics through several LSTM series. Our implementation of LSTM is similar to [37] (see Fig. 1) that has the ability to deal with vanishing and exploding gradients, as well as can reduce overfitting. In Fig. 1, the behavior of LSTM is controlled by "gates". The forget gate f is

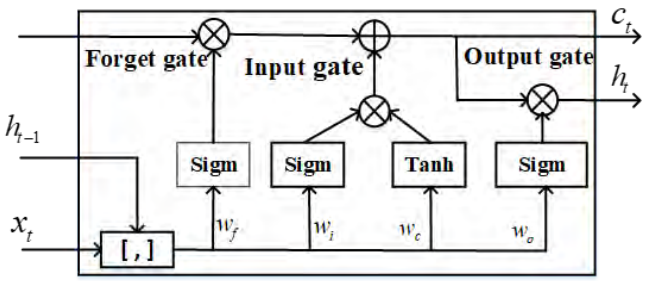

Figure 1: Internal structure of an LSTM cell used in network.

used to forget the information in the cell selectively, the input gate i controls how much new information is recorded in the cell state, and the output gate o decides the output of the new cell value. The definitions of those gates and cell update and output are as follows:

$$f_t = \sigma(W_f[x_t, h_{t-1}]) \tag{1}$$

$$i_t = \sigma(W_i[x_t, h_{t-1}]) \tag{2}$$

$$o_t = \sigma(W_o[x_t, h_{t-1}]) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c[x_t, h_{t-1}]) \tag{4}$$

$$h_t = o_t \odot c_t \tag{5}$$

For example, the output h at time $t-1$ is fed back to the memory at time t via the three gates; the cell value is fed back via the forget gate; the predicted word at time $t-1$ is fed back in addition to the memory output h at time t for word prediction. The various W matrixes are trained parameters that make the sub-network of spatial semantics robust.

2.1.3. Model

In order to maintain the continuity of semantics and obtain a sentence that describes an image, the input of LSTM at time $t = 0$ is the representation of an image, shown in the following:

$$x_0 = CNN(I) \tag{6}$$

I is an image as the input of CNN. Using an representation of images from the top layers of a convnet may have an obvious drawback, that loses the structural information of image which could be useful for obtaining richer, more descriptive spatial semantics, and low-level representations may keep more structural information. In order to obtain the abundant spatial semantics that are correspond to the position of 2-D image, we adopt the representation of an image from lower layer instead of a fully connected layer. This allows the representation focus on certain parts of an image by selecting a part of feature vectors. According to [38], the representation from layer 5 in the VGGnet not only has enough structural information of image, but also retains the location information of an image. Thus, the representation of an image from the layer 5 is more suitable as an input of LSTM. To avoid losing too much structural information of objects in images, we adopt the feature maps from the fifth convolutional layer before max pooling layer in the VGGnet as x_0. The x_0 is only used as an input of LSTM, at $t = 0$, as a priori knowledge of LSTM about the image contents.

In order to optimize all parameters by end-to-end, we directly maximize the probability of the generated sentences which describe the contents of an image by using the following formulation:

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S|I;\theta) \tag{7}$$

The θ are the parameters of the proposed semantic network, and S is its correct transcription of an image. Since the sentences are produced by words which are influenced by the former words, it is common to apply the chain rule to model the joint probability over S_1, S_2, $\cdots$, S_n, where n is the length of the sentence example as:

$$\log p(S|I) = \sum_{t=1}^{n} \log p(S_t|I, S_1, \cdots, S_{t-1}) \tag{8}$$

In more details, a true sentence $S = (S_1, \cdots, S_n)$ that describes the contents of an image, the probability is produced like that:

$$x_t = W_e S_t, t \in \{1 \cdots n\} \tag{9}$$

$$p_t = LSTM(x_{t-1}), t \in \{1 \cdots n\} \tag{10}$$

x_t $(x > 0)$ is a one-hot vector that represents the word S at time t, and the distribution of p_t represents a probability over all the words in the vocabulary. In these manners, the image by using a CNN and the sentences by using a sequence of word embedding W_e [39] are mapped to the same space.

The loss of the proposed semantic network is the sum of the negative log likelihood of the words in sentence at each step as follow:

$$L(I, S) = -\sum_{t=1}^{n} \log p_t(S_t) \tag{11}$$

We empoly end-to-end optimization to seek the optimal parameters by minimizing the loss in equation 11.

2.2. Training

In the proposed semantic network, we take a raw image I of $224 \times 224 \times 3$ and its correct transcription S as a training pair, and I is the input of CNN and S is the output of LSTM. In order to obtain a sentence that best matches the image, the proposed network is trained iteratively by backproagation over a number of epochs to optimize all parameters in the semantic network. In each epoch where samples from training set have been used once, and the training set is divided into mini-batches for batchwise optimization. In this work, we adopt the Adam optimization algorithm [40] with a mini-batch of 64 to optimization. For the training, we start with the learning rate $\alpha = 10^{-2}$ and subsequently lower it by a factor 10. Other parameters in Adam are default [40].

Although CNN and LSTM have shown great promises in many computer vision tasks, they are data-driven approaches in essence. Purely supervised learning requires a great deal of data, and optimizing parameters in those networks require more sufficient ground truth samples. Indeed, current publicly available quality-annotated image databases not only lack enough data to train this semantic network, but also lack the annotations of the spatail semantics that describe the contents of the images. Microsoft COCO (MSCOCO) [41] is a

larger dataset in scene understanding, and each image has five annotations about the contents of the image. Compared to the publicly available quality-annotated image databases, MSCOCO has the same source image content, including face, human, animal, close-up len, wide-angle lens, natural scene and so on. Therefore, the semantic network can be trained by the database of MSCOCO, and the model can be transferred to image quality databases.

The existing databases have no ability to avoid overfiting completely so far. Thus, we adopt some tricks to deal with overfiting in the training. The most effective way is to initialize the weights of the CNN component of the semantic network by a pretrained model (VGGnet), and it helps a lot in terms of generalization. Optimizing the parameters of LSTM by fixing the parameters of CNN in the training is another valuable method. After that, all parameters in our semantic network are fine-tuned to make Equation 11 minimum. More importantly, the model level overfiting-avoiding technique is adpoted, that is dropout [42]which is proved to improve the performance effectively.

2.3. Semantics of an image

2.3.1. The structural semantic

Human subjective perception deals with the human senses that generate signals from the environment through sight, hearing, touch, smell and taste. Vision is the main source of signals, and human subjective perception based on sight mainly includes recognizing and interpreting. The process of recognization generates a shallow to deep structural information of an image. Inspired by some early discovery of the visual system and similar to artificial neural network [43], CNN can be regarded as a coarse imitation of human visual mechanism[29]. The projections of different layers in the network show the hierarchical nature of the human visual perception, such as, the structural semantics in layer 2 may accord with corners and other edges, color connection, the structural semantics in layer 3 capture similar textures, the structural semantics layer in 4 have significant variation but more class-specific, and the structural semantics in layer 5 show entire objects with significant pose variation [38]. Overall, they are represent

structural information of an image, just express the structural semantics in different levels. The structural semantics from the bottom to the top layers are a progressive relationship, from the local attributes to the whole attributes of an image.

In the visual system of human beings, a local or multiple photoreceptor signal converges to a bipolar cell; the output signal of one or more bipolar cells converges to a ganglion cell in the retina, those operations are for visual maximum sensitivty. From the perspective of CNN, the structural semantics expressed by several values in a local region replaced by a value in the pooling layers can be seen as imitation of the operations in the retina. Thus, the structural semantics extracted from the pooling layers, not only obtain the maximum invariance of an image in scalerotation and translation, but also more representative [29]. There are many kernels in a pooling layer, and every kernel is sensitive to the different structural information of an image. Therefore, the structural semantics extracted from a pooling layer are diverse, and the number of the structural semantics from a pooling layer is equal to the number of kernels. These structural semantics produced by different kernels from the same layer are a whole, and represent the structural information of this layer. Because of the bottom-to-top layers in the CNN are a progressive relationship, the structural semantics from different layers may have different image quality in assessment.

2.3.2. The spatial semantic

The contents of an image is supposed to understood in the process of interpretation. Human beings like to express what they see through a natural language. Thus, a sentence that depicts the contents of an image is needed in additional to visual understanding. Even though an image distorted with different levels or the local structures suffer different degradations, it is possible to obtain the same interpretation of images. In this situation, the biggest difference of those images is the subjective visual perception of human beings. In the sub-network of spatial semantics, at the time t, LSTM not only generates a word that represents the objects or expresses the relationship between objects,

but also obtains a loss corresponding to the word. Thus, the loss of a word can be used to indicate human subjective visual perception. In other words, the spatial semantics of an image not only contain the sentence, but also include the loss of the sentence.

3. Semantic for IQA

3.1. Structural semantics for IQA

The structural semantics from a layer in CNN are equal to $W \times F$, W is the number of the kernels in a layer, F is a feature map that represents the structural information produced from one kernel. A natural image generally has a variety of local structures in its scene. When a distortion is added to the image, the different local structural will suffer different degradations [44]. The various kernels in a layer can solve this problem because different kernels extract different structural degradations in an image, and result in the structural semantics are very rich. When those structural semantics regressed onto the human opinion score directly, it is very easy overfitting on database. Therefore, extracting a low-dimensional but distinctive representation of those structural semantics is very urgent.

Taking an average is the most common pooling strategy that is adopted in many IQA algorithms [44, 7, 3]. But this has an obvious drawback, it can not reflect how the local structural semantics degradation varies. Based on this idea that the global variation of the structural semantics can reflect its degradation of the overall structural semantics, we propose to compute the standard deviation of the structural semantics to represent the structural semantics. To represent a structural semantic comprehensively, we use the both statistical characteristics to represent the structural semantics. The two statistical characteristics are derived:

$$\mu = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} x_{i,j} \tag{12}$$

$$\sigma = \sqrt{\frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (x_{i,j} - \mu)^2} \tag{13}$$

where M, N represent the length and width of a structural semantic, respectively, and $x_{i,j}$ represents the value of a pixel in a structural semantic. So a bundle of representations about the structural semantics are obtained, the mean-pool vector $\mu = (\mu_1, \mu_2, \cdots, \mu_W)$ and standard-deviation-pooled vector $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_W)$. Then, the structural semantics from a layer in CNN can be expressed by the concatenated feature vector $V_s = (\mu, \sigma)$.

3.2. Spatial semantics for IQA

The spatial semantics are consist of the sentence and the loss correspond to the sentence. A sentence are composed of words. In order to quantization, the words should be converted into vectors. Representation of words as continuous vectors has a long history[45, 39]. In this work, a word maps to a vector through word embedding W_e, as shown in Equation 9. Generally speaking, the dimension of x_t is usually between $300 \sim 500$. As the dimension of x_t increases, the parameters of word embedding in the semantic network will increase dramatically. Indeed, current publicly available databases have no sufficient ground truth samples to optimize parameters. Thus, the accuracy of the network may be drop when have more parameters. In the experiment of training the semantic network, we try different dimension of x_t. As mentioned above, the accuracy decreases when the dimension of x_t increases. Especially, the accuracy drops five percentage points when the dimension up to 500. In order to obtain the more accurate semantic information of an image, the dimension of x_t is adopt in 300. L_t is the loss accord with the word at time t, and the dimension is 1. Obviously, there is a great difference in dimension between the word and the loss, resulting in the impact of loss in IQA is weakened. In order to reveal the importance of loss in IQA, the dimension of loss is incresed by copying. We define the dimension of the loss is m, and the vector $l_t = (L_1, L_2, \cdots, L_m)$, in this equation $L_1, L_2, \cdots, L_m$ are equal to L_t. Therefore, the represetation of the word at time t is $w_t = (x_t, l_t)$.

When the vectors are regressed onto the subjective score, they need the vectors have uniform dimension. Thus, a fixed length of sentences is proposed to

satisfy with the length of the spatial semantics of each image. When the length of a sentence is unable to reach the fixed length, we circle the sentence until reaching. The definition of the fixed length is N, and the spatial semantics of an image is $N \times w_t$. Either for improving the weight of human subjective vision or obtaining the same length in the spatial semantics, the spatial semantics are achieved through a large number of replicas. In this situation, the spatial semantics are very redundant, and the high dimension of the vector can easily lead to overfitting. Thus, it is inevitable to reduce the dimension of the spatial semantics, after the dimensionality reduction of $N \times w_t$, the vector V_h represents the spatial semantics of images.

3.3. Joint assessment for image quality

Distortions will affect the semantics of images. In this work, we try to assess the quality of images based on those semantics. In other words, when we imitate the process of human subjective perception to assess the quality of images, we should take the structural semantics and the spatial semantics into account. In the process of recognizing, the projection from different layers reveals the different structural information of an image, resulting in the structural semantics from different layers are different. When the structural semantics are applied to IQA, the structural semantics from different layers may have different predictions of quality about an image. So, we explore which structure semantics are the most suitable for image quality evaluation, edge or significant variation or others.

The structural semantics and the spatial semantics represent different information in the process of human subjective perception. Therefore, the semantics can be applied to image quality assessment. The structural semantics tend to texture, structural information, and the spatial semantics tend to identification, position information. In order to explore the importance of the structural semantics and the spatial semantics in the process of image quality assessment, the quality of an image is assessed by the structural semantics and the spatial semantics, respectively, then the quality from different semantics are weighted

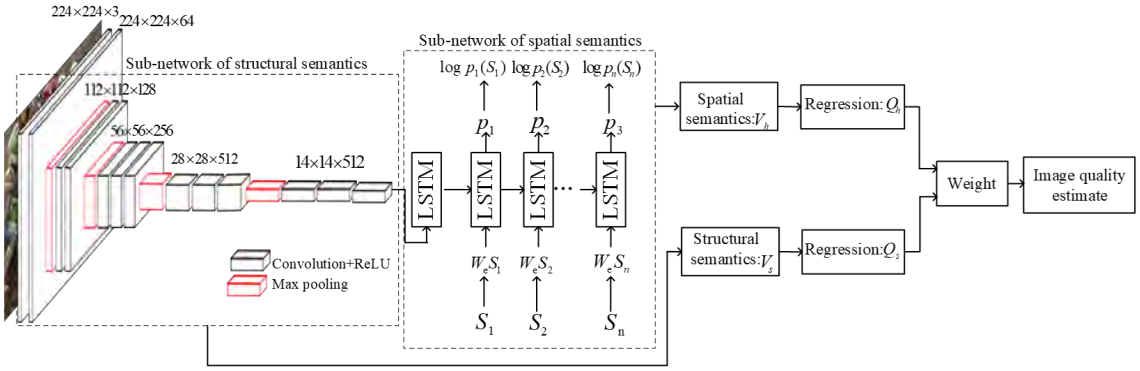

Figure 2: Illustration of the proposed method configurations for NR IQA

into a global imagewise quality estimate. A mapping is learned from the vector space to the quality scores (i.e., subjective quality value Q) by using a regression model. Q_s and Q_h are the quality of image regressed by V_s and V_h that represent the structural semantics and the spatial semantics, respectively. In this implementation, a support vector regression (SVR)is applied to image quality assessment problems frequently for regression.

$$\Re_s = SVR_{train}(V_s, Q) \tag{14}$$

$$\Re_h = SVR_{train}(V_h, Q) \tag{15}$$

When the mapping is determined, the quality of a distorted image I_d can be predicted as,

$$Q_s(I_d) = SVR_{predict}(V(I_d), \Re_s) \tag{16}$$

$$Q_h(I_d) = SVR_{predict}(V(I_d), \Re_h) \tag{17}$$

Then the global quality of an image Q is derived:

$$Q = \lambda Q_s + (1 - \lambda) Q_h \tag{18}$$

where the λ indicates the importance of the structural semantics in the image quality assessment, and $(1 - \lambda)$ measures the impact of the spatial semantics on image quality. The architecture of the proposed method is shown in Fig 2.

4. Experiments

4.1. Datasets

Experiments are evaluated on the LIVE [46], CSIQ, TID2013 [47], the LIVE multiply distorted database [48], and the LIVE In the Wild Image Quality Challenge Database [49]. Since those databases play an important role in objective image quality assessment, a brief introduction of these databases in the following.

The LIVE database is the first successful publicly available quality-annotated image database, and still used widely. The database consists of 779 quality annotated images based on 29 source reference images that subject to five different types of distortions at different distortion levels. Distortion types are JPEG compression, JPEG2000(JP2K) compression, additive white Gaussian noise, Gaussian blur and a simulated fast fading Rayleigh channel. The DMOS of each distorted image is obtained, and lie in the range of [0, 100], where a lower DMOS indicates a better visual image quality.

The CSIQ database includes 899 distorted quality-annotated image generated by 30 reference image with one of the following distortions: JPEG compression, JPEG2000(JP2K) compression, Guasian blur, Guassian white noise, Guassian pink noise or contrast change. In order to obtain a more comprehensive definition of the quality about an image, subjects were asked to assess the quality of position distorted images horizontally on a monitor. After alignment and normalization, the resulting DMOS in the range of [0, 1], similar to the above, where a lower value indicates better quality.

The TID2013 image quality database contains 3000 quality annotated images based on 25 source reference images distorted by 24 different distortion types, and each distortion have 5 distortion levels. Thus, the TID2013 is a more challenge database for evaluation of the proposed methods. Different from the one that used for the construction of LIVE, the TID2013 database employed a competition-like double stimulus procedure to obtain the annotation. The quality of this database rated by MOS, and lie in range of [0, 9], where larger

MOS indicate better quality.

The LIVE multiply distorted (MD) database is the first database that include multiply distorted images. This database contains 450 quality annotated images based on 15 source reference images distorted by two types of distortion in two combinations: simulated Guasian blur followed by JPEG compression and Guasian blur followed by additive white Gaussian noise. Each multiple distortion is used to generate 225 images for each part of the study of which 90 are singly distorted (45 of each type) and 135 are multiply distorted, and the DMOS of each distorted image is provided, lie in range of [0, 100], similar to the above, where a lower value indicates better quality.

The LIVE In the Wild Image Quality Challenge Database (CLIVE) comprises 1162 unique images that taken under real life conditions, a large variety of objects and scenes captured. The images in this database were subjected to numerous types of authentic distortions during the captured process. The distortions include, e.g., low-light blur and noise, motion blur, camera shake, overexposure, underexposure, a variety of color errors, compression errors, and many combinations of these and other impairments. This database has no reference images, because the distorted images are originals. The value of image quality obtained more rigorous. More than 8100 human subjects in a tightly monitored crowdsourced study, yielding more than 35000 human judgments. The quality of this database rated by MOS, that in range of [0, 100], where larger MOS indicate better quality. In order to compare the database, the properties of the databases are shown in Table 1.

4.2. Experimental setup

In order to gain a semantic model, the larger database MSCOCO is used to train the semantic network with 60 epoches. Because the quality-annotated databases have the same source images when compared to MSCOCO, transferring the semantic model to the quality-annotated databases to extract the semantics is reasonable. In order to normalization, images in IQA databases are resized to the size of the input of the semantic network. Then, the trained

Table 1: THE COMPARSION OF PROPERTIES ABOUT THE DATABASES

Database	Reference image	Distorted image	Noise	Criterion	Quality
LIVE	29	779	5	DMOS	$0-100$
CSIQ	30	899	6	DMOS	$0-1$
TID2013	24	3000	24	MOS	$0-9$
LIVE MD	15	450	2	DMOS	$0-100$
CLIVE	-	1162	-	MOS	$0-100$

semantic model is used to extract semantics. The structural semantics are come from the feature maps in pool2, pool3, pool4, and conv5, and the spatial semantics which include the sentences and the loss correspond to the sentences are generated from LSTM. Through the operation of Equation 12, 13, the vectors V_{s_pool2}, V_{s_pool3}, V_{s_pool4}, V_{s_conv5} represent the structural semantics from different layers in the semantic network. In this work, the redundancy of structural information in the first pooling layer is the reason that we discard it. After the operations of weighting and dimensionality reduction are applied to the sentence and loss, the vector V_h represents the spatial semantics is obtained. Then, the vectors V_s, V_h are regressed to the quality of image Q_s and Q_h, respectively. The global quality of an image is obtained by weighting Q_s and Q_h.

In order to verify the effectiveness of our proposed method, we evaluate the proposed method on the all quality-annotated databases mentioned above. Similar to the most of the SVR based quality prediction, an 80% - 20% training-testing procedure is used, i.e., 80% distorted images in a database are chosen for training, and the rest 20% for testing. Moreover, in order to eliminate the bias caused by the data separation, the training-testing procedure is repeated for 10 times, and the median performance is used to the final results. To assess the generalization ability of the proposed method, a cross-dataset evaluation is carried out, and the model of regression is trained on LIVE, and tested on other IQA databases.

In this work, we test our algorithm on Inter i7-6700 3.4GHz CPU and Nvidia

GeForce GTX 1050Ti GPU. When including all steps (also semantic information extraction), the executing time of an image is 96.21 ms.

4.3. Quantitative analysis

In order to make a quantitative analysis of the experimental results, two common evaluation criteria are adopted.

4.3.1. Spearman's rank-order correlation coefficient (SRCC)

It is a nonparametric measure and is define as:

$$SRCC = 1 - \frac{6\sum_i d_i^2}{N(N^2-1))} \tag{19}$$

where N is the number of to be estimated images, d_i is the rank difference between the MOS and the model prediction of the $i-th$ image.

4.3.2. Pearson linear correlation coefficient (PLCC)

Another evaluation criteria is pearson linear correlation coefficient (PLCC). It is a measure of the linear correlation.

$$PLCC = \frac{\sum_i (q_i - q_m)(\hat{q}_i - \hat{q}_m)}{\sqrt{\sum_i (q_i - q_m)^2}\sqrt{\sum_i (\hat{q}_i - \hat{q}_m)^2}} \tag{20}$$

where q_i and $\hat{q}_i$ represent the MOS or DMOS of the image and the predicted quality of the $i-th$ image, respectively. For both correlation metrics a value close to 1 indicates high performance of a specific quality measure.

4.4. Performance evaluation

Performances of the proposed method are reported in this subsection. In this work, the semantics of images are used to assess the quality, and the semantics are not only contain the structural semantic but also include the spatial semantics. The Table 2 summarizes the performance of different semantics applied to assess the quality of images in TID2013 database that is the largest database with quality annotated images. From a single semantics perspective, the structural semantics perform better than the spatial semantics. In order to indicate the structural semantics is more important in IQA, we set the $\lambda = 0.6$.

Table 2: PERFORMANCE EVALUTION FOR DIFFERENT SEMANTICS ON TID2013

Semantics	PLCC	SRCC
V_{s_pool2}	0.753	0.693
V_{s_pool2}+V_h	0.825	0.815
V_{s_pool3}	0.835	0.800
V_{s_pool3}+V_h	0.874	0.859
V_{s_pool4}	0.885	0.873
V_{s_pool4}+V_h	**0.896**	**0.886**
V_{s_conv5}	0.844	0.821
V_{s_conv5}+V_h	0.871	0.862
V_h	0.717	0.740

The results in Table 2 show that the performances of the proposed method are improved when adding the spatial semantics, and the structural semantics from the fourth pooling layer combine with the spatial semantics perform best.

Because of the superior performance, the semantics that the structural semantics extracted from the fourth pooling layer combine with the spatial semantics are adopted to IQA. Evaluations are presented on all IQA databases through the proposed method. To reflect its effectiveness, the performances of the proposed method are compared to other state-of-the-art NR IQA and FR IQA methods. The results are shown in Table 3 . The PSNR and SSIM are the most common FR IQA methods, and the rest are NR IQA methods. The following conclusions are conclude in the Table 3 . The proposed approach obtains superior performance on CSIQ, LIVE MD and CLIVE in terms of PLCC and SRCC. There is a slight improve in CSIQ, LIVE MD, and the performance on CLIVE that is much more difficult than other databases has a great improvement. The proposed method performs slightly worse than BIECON and DIQAaM in LIVE, and performance on TID2013 is also slightly worse than DIIVINE.

Table 3: PERFORMANCE EVALUTION FOR DIFFERENT DATABASE

IQM	LIVE		CSIQ		TID2013		LIVE MD		CLIVE	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
PSNR [50]	0.876	0.872	0.806	0.800	0.636	0.706	0.725	0.815	N/A	N/A
SSIM [3]	0.948	0.945	0.876	0.861	0.775	0.691	0.845	0.882	N/A	N/A
DIIVINE [11]	0.928	0.926	0.876	0.896	0.908	**0.923**	0.874	0.894	0.546	0.568
NIQE [7]	0.926	0.925	0.901	0.910	0.845	0.835	0.745	0.815	0.421	0.478
BRISQUE [8]	0.939	0.942	0.756	0.797	0.572	0.651	0.897	0.921	0.607	0.585
CORNIA [51]	0.942	0.943	0.714	0.781	0.549	0.613	0.900	0.915	0.618	0.662
BIECON [52]	0.958	0.960	0.815	0.823	0.717	0.762	0.909	0.933	0.595	0.613
MEON [24]	-	-	0.932	-	**0.912**	-	-	-	-	-
DIQAaM [53]	**0.960**	**0.972**	-	-	0.835	0.855	-	-	0.606	0.601
Proposed	0.951	0.953	**0.941**	**0.954**	0.886	0.896	**0.934**	**0.933**	**0.756**	**0.798**

4.5. Cross-Database evaluation

The evaluation strategy in the above is inadequate to evaluate the generalization capability of the proposed method. In order to reflect the generalization capability, we extend cross-database experiments presented in [24] with our results. The LIVE database contains only four distortions types (JPEG, JP2K, Gaussian blur, white noise) shared between the other two databases (CSIQ, TID2013). For the three databases, one of them is chosen for training, and the rest two for testing. The performances of this strategy are shown in Table 4. The proposed method outperforms than other methods when cross-evaluated on the subset of TID2013, and the results show superior performance in LIVE when the models trained on subset of TID2013. Unfortunately, when cross-evaluated on the subset of CSIQ, the performances are great worse when compared to other methods, and the performances slightly worse than other state-of-the-art methods in LIVE when trained on CSIQ.

There is a clear disadvantage in the above-mentioned cross-database evaluation, that is no test on other possibly unknown distortions. Thus, the proposed

Table 4: SRCC IN CROSS-DATABASE EVALUTION. THE DATABASES ARE CONTAINING ONLY FOUR DISTORTIONS TYPES.

Trained on	LIVE		CSIQ		TID2013	
Tested on	CSIQ	TID2013	LIVE	TID2013	LIVE	CSIQ
IL-NIQE [9]	**0.880**	0.877	0.916	0.877	0.880	0.916
NIQE [7]	0.867	0.814	**0.919**	0.814	0.867	**0.919**
BRISQUE [8]	0.827	0.726	0.633	0.571	0.808	0.795
CBIQ [54]	0.842	0.817	0.811	0.804	0.794	0.618
DIIVINE [11]	0.854	0.854	0.522	0.764	0.641	0.621
Proposed	0.837	**0.887**	0.909	**0.892**	**0.912**	0.817

Table 5: SRCC COMPARTION IN CROSS-DATABASE EVALUTION. ALL MODELS ARE TRAINED ON THE FULL LIVE DADABASE AND EVALUTION ON CSIQ, TID2013, CLIVE AND LIVE MD.

IQM	CSIQ	TID2013	CLIVE	LIVE MD
BLIINDS-II [55]	0.654	0.405	0.102	0.456
DIIVINE [11]	0.553	0.487	0.342	0.662
BRISQUE [8]	0.549	0.466	0.089	0.550
NIQE [7]	0.630	0.317	0.421	0.745
HOSA [56]	0.631	0.465	0.419	0.616
FRIQUEE [57]	0.688	0.468	0.344	0.502
VIDGIQA [58]	0.641	0.415	0.315	-
DIQAam [53]	0.681	0.392	-	-
Proposed	**0.767**	**0.582**	**0.510**	**0.857**

method is evaluated by training on one database and testing on other database. For that, the proposed method trained on the full LIVE database, and evaluated on the other full database, CSIQ, TID2013, LIVE MD, CLIVE. For the full CSIQ database, two unseen distortions (i.e.,pink additive noise and contrast change) are considerably different when compared to LIVE database, and there are twenty unseen distortions in the TID2013. Moreover, CLIVE and LIVE MD that specific picture's mixture of distortions are less likely to be in the training set. The results are display in Table 5. Even though the results in Table 5 show the superior performance of the proposed method when compared to other state-of-the-art methods, the results are still very unsatisfactory. Unsurprisingly, the results suggest that learning a non-distortion-specific IQA metric using the examples in the LIVE database is hard.

4.6. Experimental analysis

In this work, we propose a method that is applied the semantics of images to IQA. When an image distorted, human beings subjective perception is change, not only the structural information, but also the interpretation of the image. Thus, the structural semantics that represent the structural information of an image and the spatial semantics that relative to the interpretation of an image are extracted to assess the quality of images. The Fig 3 displays the semantics of images with different levels distortions. In the Fig 3, (a) have litttle distortions that can not be perceived, and (b), (c) have obvious degradations in local regions that are shown with a red frame when compared to (a). By examining the pairs of (b), (e) and (c), (f), when there is a distortion in the local region of the distorted images, the corresponding regions of the structural semantic also distorted that are highlighted with red rectangular boxes. This proves that it is reasonable to use structural semantics to describe the structural information of the distorted images. Moreover, compared to (a), the distortions in (c) that painted by the red frame have lost its important structural information for identification, resulting in the word 'ground' is instead of 'motorcycle' in the spatial semantics that shown in (g) and (i). This not only reflects the consistency

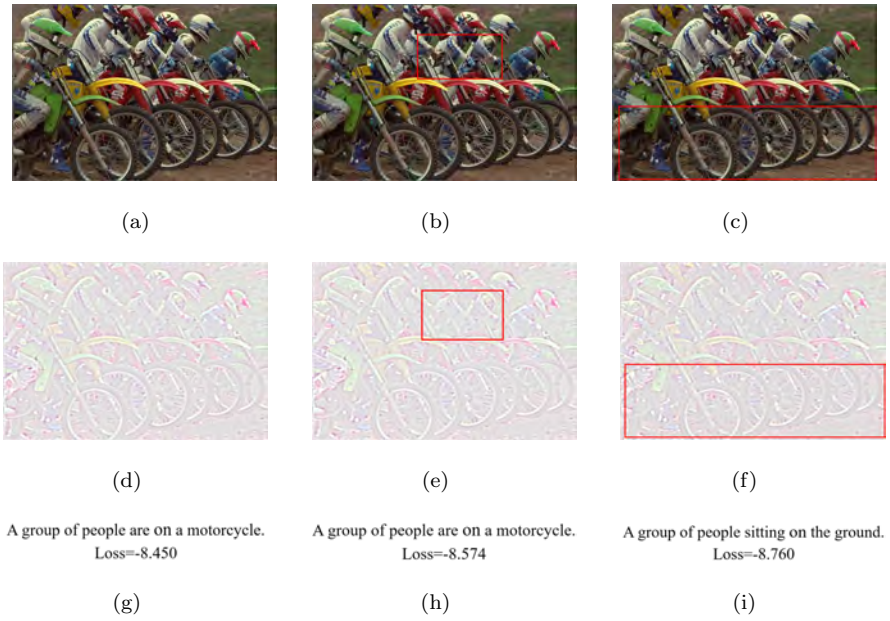

Figure 3: An example of the semantics extracted from the images with different level distortions. (a), (b), (c) are the distorted image, (d), (e), (f) are the structural semantics of the images corresponding to (a), (b), (c), and (g), (h) , (i) are the spatial semantics.

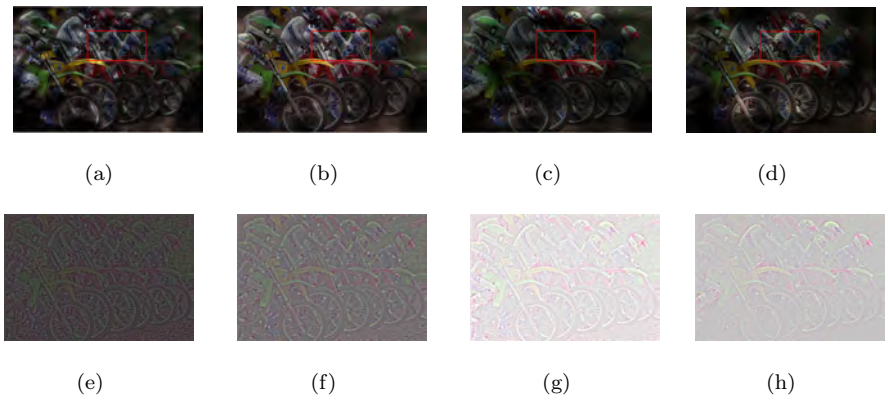

Figure 4: An example of the different structural semantics of (b) in Fig 3. (a), (b), (c), (d) show the responses of the structural semantics in (b) of Fig 3, and (a) show the response of the structural semantics in the second pooling layer, (b) show the response of the structural semantics in the third pooling layer, (c) show the response of the structural semantics in the fourth pooling layer , (d) show the response of the structural semantics in the fifth convolutional layer. (e), (f), (g), (h) are the structural semantics from different layers in semantics network, and (e) from the second pooling layer, (f) from the third pooling layer, (g) from the fourth pooling layer, (h) from the fifth convolutional layer.

of the structural semantics and the spatial semantics, but also shows that the degradation of semantics can be applied to IQA. Even through (b) has more distortions than (a), they have the same sentences that express the contents of images. Fortunately, the loss of the sentences can make up for this deficiency. Therefore, the sentence and the loss corresponding to the sentence are essential in spatial semantics. Even if there are no suitable words in the thesaurus to describe the objects or relationship of the image, the sub-network of of spatial semantic has mapped the images to the same vector space in itself. In summary, the semantics contain the structural semantics and the spatial semantics are suitable for IQA.

In the Table 2 , the structural semantics from the fourth pooling layer performs better. To explain this conclusion, the different structural semantics of an image are shown in Fig 4 . The reponse of different structural semantics in images are display in (a), (b), (c) and (d), the brighter of the regions, the

Table 6: The comparion of executing time of different methods.

Method	Executing time (s)
PSNR [50]	0.021
SSIM [3]	0.032
NIQE [7]	0.249
IL-NIQE [9]	4.135
BRISQUE [8]	0.628
DIIVINE [11]	15.315
BLIINDS-II [55]	60.186
FRIQUEE [57]	23.982
HOSA [56]	0.265
Proposed	1.604

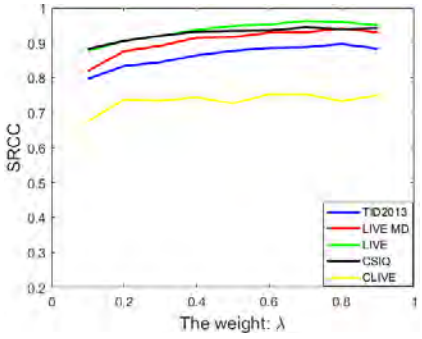

Figure 5: SRCC of the proposed method for NR IQA in dependence of the weight λ evaluated on all databases.

greater response of the regions. In the degraded regions, the structural semantics from different layers all have great response. But the response in other regions is quite different, the structural semantics scatter to the whole image and no focus in (a) and (b), resulting in the performance of IQA worse than (c), (d). Even though the latter two have obvious focus, the structural semantics in (c) are more discriminatory when compares with (d). The focus of the structural semantics in (c) are more discrete, and the degraded regions accounts for a larger proportion. It may be the reason that the structural semantics from the fourth pooling layer are more suitable for IQA. Obviously in (e), (f), (g), (h), the luminance of the structural semantics from the fourth pooling layer is brighter than the other layers, and the contents are more clearer. In other words, the greater luminance, the stronger contrast in the structural semantics, the more structural information have. The mean can reflect the overall luminance of images and standard deviation can reflects the contrast. Thus, the mean and standard deviation can be used to distinguish the different structural semantics from different layers. Finally, the conclusion that it is reasonable to use mean and variance to represent a structural semantic is obtained.

In order to compare the real-time performace of different methods when evaluate the image quality, the executing time of some methods are listed in Table 6 . As shown in Table, the time taken for the FR IQA methods is much less than the time used for the NR IQA methods. Although the executing time of NR IQA methods (i.e., NIQE[7], BRISQUE[8], HOSA[56]) are less than the proposed method when assess the image quality, the performances of those methods are perform worse than the proposed method. Moreover, the NIQE performs slight worse than the proppsed methd, and the BRISQUE, HOSA perform much worse than the proposed method in cross-database evaluation. Taking the performance and executing time into account, the proposed method is a feasible method.

In order to reveal the sensitivity of the structural semantics in IQA, a weighted approach is adopted. Fig 5 shows the influence of different weight λ on the SRCC in all databases. The overall trend is that with the weight λ

increases, the value of SRCC improves first, then decreases, and the maximum value of the SRCC is between $0.5 - 0.8$. This suggest that the structural semantic are more important than the spatial semantics in image quality assessment. As the weight *lambda* continues to increase, the phenomenon of SRCC decline may shows that the spatial semantics play an indispensable role in evaluating the quality of the image. It also proves the effectiveness of our proposed method.

5. Conclusion

In this letter, inspired by the subjective perception of human beings, a new NR IQA method based on semantics of images is proposed. The structural semantics imitated by CNN and spatial semantics produced by LSTM are corresponding to the process of recognition and interpretation of human perception, respectively. The validity of different structural semantics in IQA are studies, and the structure semantics and spatial semantics are used to evaluate image quality jointly. The experimental results show that the proposed method outperforms other state-of-the-art approached in NR IQA, and the proposed method has a good generalization capabilities on cross-database evaluation.

However, having no public training set annotated by quality and description is a principle defect of the proposed method, and larger databases hopefully to be generated in the future. In other words, the generalization performance of the proposed method can be improved for considerable room in larger database. In this work, image quality assessment is separate from the acquisition of semantics. Even through a relative generic neural network is able achieve high prediction performance, incorporating IQA with semantics to end-to-end optimization may lead to further improvements. Its relative simplicity suggests that neural networks used for IQA have lots of potential.

[1] Z. Wang, A. C. Bovik, Modern image quality assessment, Synthesis Lectures on Image, Video, and Multimedia Processing 2 (1) (2006) 1–156.

[2] S. Gabarda, G. Cristobal, N. Goel, Anisotropic blind image quality assessment: Survey and analysis with current methods, Journal of Vi-

sual Communication and Image Representation 52 (2018) 101–105. doi:10.1016/j.jvcir.2018.02.008.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612. doi:10.1109/TIP.2003.819861.

[4] Y. Wen, Y. Li, X. Zhang, W. Shi, L. Wang, J. Chen, A weighted full-reference image quality assessment based on visual saliency, Journal of Visual Communication and Image Representation 43 (2017) 119–126. doi:10.1016/j.jvcir.2016.12.005.

[5] J. Wu, G. Shi, W. Lin, X. Wang, Reduced-reference image quality assessment with orientation selectivity based visual pattern., in: ChinaSIP, Citeseer, 2015, pp. 663–666.

[6] A. Rehman, Z. Wang, Reduced-reference image quality assessment by structural similarity estimation, IEEE Transactions on Image Processing 21 (8) (2012) 3378–3389. doi:10.1109/TIP.2012.2197011.

[7] A. Mittal, R. Soundararajan, A. C. Bovik, Making a completely blind image quality analyzer, IEEE Signal Processing Letters 20 (3) (2013) 209–212. doi:10.1109/LSP.2012.2227726.

[8] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on Image Processing 21 (12) (2012) 4695–4708. doi:10.1109/TIP.2012.2214050.

[9] L. Zhang, L. Zhang, A. C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Transactions on Image Processing 24 (8) (2015) 2579–2591. doi:10.1109/TIP.2015.2426416.

[10] L. Tang, L. Li, K. Sun, Z. Xia, K. Gu, J. Qian, An efficient and effective blind camera image quality metric via modeling quaternion wavelet coefficients, Journal of Visual Communication and Image Representation 49 (2017) 204–212. doi:10.1016/j.jvcir.2017.09.010.

[11] A. K. Moorthy, A. C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Transactions on Image Processing 20 (12) (2011) 3350–3364. doi:10.1109/TIP.2011.2147325.

[12] Q. Wu, H. Li, F. Meng, K. N. Ngan, S. Zhu, No reference image quality assessment metric via multi-domain structural information and piecewise regression, Journal of Visual Communication and Image Representation 32 (2015) 205–216. doi:10.1016/j.jvcir.2015.08.009.

[13] S. J. Thomas, Foundations of vision., Psyccritiques 42 (7).

[14] E. P. Simoncelli, B. A. Olshausen, Natural image statistics and neural representation, Annual Review of Neuroscience 24 (1) (2001) 1193–1216. doi:10.1146/annurev.neuro.24.1.1193.

[15] G. M. Johnson, On contrast sensitivity in an image difference model, Pics April (2002) 18–23.

[16] C. Wang, M. Shen, C. Yao, No-reference quality assessment for dct-based compressed image, Journal of Visual Communication and Image Representation 28 (2015) 53–59.

[17] V. Tyagi, Texture feature, in: Content-Based Image Retrieval, Springer, 2017, pp. 161–182.

[18] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, COMMUNICATIONS OF THE ACM 60 (6) (2017) 84–90. doi:10.1145/3065386.

[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, international conference on learning representations.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, computer vision and pattern recognition (2016) 770–778doi:10.1109/CVPR.2016.90.

[21] S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, Signal, Image and Video Processing 12 (2) (2018) 355–362. doi:10.1007/s11760-017-1166-8.

[22] J. Kim, S. Lee, Deep learning of human visual sensitivity in image quality assessment framework, 30TH IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2017) (2017) 1969–1977doi:10.1109/CVPR.2017.213.

[23] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, L. Zhang, Group mad competition? a new methodology to compare objective image quality models, in: Computer Vision and Pattern Recognition, 2016, pp. 1664–1673.

[24] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-end blind image quality assessment using deep neural networks, IEEE Transactions on Image Processing 27 (3) (2018) 1202–1213. doi:10.1109/TIP.2017.2774045.

[25] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, Image database tid2013: Peculiarities, results and perspectives, Signal Processing Image Communication 30 (2015) 57–77. doi:10.1016/j.image.2014.10.009.

[26] P. H. Lindsay, D. A. Norman, Human information processing; an introduction to psychology, ACADEMIC, 1972.

[27] J. Sweller, J. J. G. V. Merrienboer, F. G. W. C. Paas, Cognitive architecture and instructional design, in: Educational Psychology Review, 1998, pp. 251–296.

[28] D. H. Hubel, T. N. Wiesel, Receptive fields of single neurones in the cat's striate cortex, Journal of Physiology 148 (3) (1959) 574. doi:10.1113/jphysiol.1959.sp006308.

[29] L. C. Yann, B. Yoshua, H. Geoffrey, Deep learning, Nature 521 (7553) (2015) 436–44. doi:10.1038/nature14539.

[30] M. K. Tanenhaus, M. J. Spiveyknowlton, K. M. Eberhard, J. C. Sedivy, Integration of visual and linguistic information in spoken language comprehension, Science 268 (5217) (1995) 1632–1634. doi:10.1126/science.7777863.

[31] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, international conference on learning representations.

[32] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, neural information processing systems (2014) 3104–3112.

[33] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, computer vision and pattern recognition (2015) 3156–3164.

[34] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, international conference on machine learning (2015) 2048–2057.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252. doi:10.1007/s11263-015-0816-y.

[36] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[37] W. Zaremba, I. Sutskever, O. Vinyals, Recurrent neural network regularization, Eprint Arxiv.

[38] M. D. Zeiler, R. Fergus, Visualizing and Understanding Convolutional Networks, Springer International Publishing, 2014.

[39] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv: Computation and Language.

[40] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, Computer Science.

[41] X. Chen, H. Fang, T. Y. Lin, R. Vedantam, S. Gupta, P. Dollar, C. L. Zitnick, Microsoft coco captions: Data collection and evaluation server, Computer Science.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (1) (2014) 1929–1958.

[43] N. Kriegeskorte, Deep neural networks: A new framework for modeling biological vision and brain information processing., Annual Review of Vision Science 1 (1) (2015) 417. doi:10.1146/annurev-vision-082114-035447.

[44] W. Xue, L. Zhang, X. Mou, A. C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE Transactions on Image Processing 23 (2) (2014) 684–695. doi:10.1109/TIP.2013.2293423.

[45] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors., MIT Press, 1988.

[46] H. R. Sheikh, M. F. Sabir, A. C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE TRANSACTIONS ON IMAGE PROCESSING 15 (11) (2006) 3440–51. doi:10.1109/TIP.2006.881959.

[47] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, Color image database tid2013:

Peculiarities and preliminary results, in: European Workshop on Visual Information Processing, 2013, pp. 106–111.

[48] D. Jayaraman, A. Mittal, A. K. Moorthy, A. C. Bovik, Objective quality assessment of multiply distorted images, in: Signals, Systems and Computers, 2013, pp. 1693–1697.

[49] D. Ghadiyaram, A. Bovik, Massive online crowdsourced study of subjective and objective picture quality., IEEE Transactions on Image Processing 25 (1) (2015) 372–387. doi:10.1109/TIP.2015.2500021.

[50] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, A. C. Bovik, Image quality assessment based on a degradation model, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 9 (4) (2000) 636–650. doi:10.1109/83.841940.

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

[52] J. Kim, S. Lee, Fully deep blind image quality predictor, IEEE Journal of Selected Topics in Signal Processing 11 (1) (2017) 206–220. doi:10.1109/JSTSP.2016.2639328.

[53] S. Bosse, D. Maniry, K. Muller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Transactions on Image Processing 27 (1) (2018) 206–219. doi:10.1109/TIP.2017.2760518.

[54] P. Ye, D. S. Doermann, No-reference image quality assessment based on visual codebook., in: ICIP, Citeseer, 2011, pp. 3089–3092.

[55] M. A. Saad, A. C. Bovik, C. Charrier, Dct statistics model-based blind image quality assessment, in: Image Processing (ICIP), 2011 18th IEEE International Conference on, IEEE, 2011, pp. 3093–3096.

[56] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, D. S. Doermann, Blind image quality assessment based on high order statistics aggregation, IEEE Transactions on Image Processing 25 (9) (2016) 4444–4457. doi:10.1109/TIP.2016.2585880.

[57] D. Ghadiyaram, A. C. Bovik, Perceptual quality prediction on authentically distorted images using a bag of features approach., Journal of Vision 17 (1). doi:10.1167/17.1.32.

[58] J. Guan, S. Yi, X. Zeng, W. Cham, X. Wang, Visual importance and distortion guided deep image quality assessment framework, IEEE Transactions on Multimedia 19 (11) (2017) 2505–2520. doi:10.1109/TMM.2017.2703148.