Fine-Grained Image Quality Caption With Hierarchical Semantics Degradation

Wen Yang, Jinjian Wu, *Member, IEEE*, Shiwei Tian, Leida Li, *Member, IEEE*,
Weisheng Dong, *Member, IEEE*, and Guangming Shi, *Fellow, IEEE*

***Abstract*— Blind image quality assessment (BIQA), which is capable of precisely and automatically estimating human perceived image quality with no pristine image for comparison, attracts extensive attention and is of wide applications. Recently, many existing BIQA methods commonly represent image quality with a quantitative value, which is inconsistent with human cognition. Generally, human beings are good at perceiving image quality in terms of semantic description rather than quantitative value. Moreover, cognition is a needs-oriented task where humans are able to extract image contents with local to global semantics as they need. The mediocre quality value represents coarse or holistic image quality and fails to reflect degradation on hierarchical semantics. In this paper, to comply with human cognition, a novel quality caption model is inventively proposed to measure fine-grained image quality with hierarchical semantics degradation. Research on human visual system indicates there are hierarchy and reverse hierarchy correlations between hierarchical semantics. Meanwhile, empirical evidence shows that there are also bi-directional degradation dependencies between them. Thus, a novel bi-directional relationship-based network (BDRNet) is proposed for semantics degradation description, through adaptively exploring those correlations and degradation dependencies in a bi-directional manner. Extensive experiments demonstrate that our method outperforms the state-of-the-arts in terms of both evaluation performance and generalization ability.**

***Index Terms*— Image quality assessment, quality caption, hierarchical semantics degradation, deep neural network.**

I. Introduction

With the development of mobile devices and social media, we are exposed to a large amount of digital image information. However, digital images are inevitably distorted at various stages of their processes cycle, which may degrade the visual experience of human viewers. Hence, objective image quality assessment (IQA), which is capable of precisely and automatically estimating human perceived image quality, is important and has received lots of attention. According to the availability of the reference image, current objective IQA methods can be classified into full-reference (FR) IQA [1], [2], reduced-reference (RR) IQA [3], [4], and no-reference/blind (NR/B) IQA [5], [6]. Because it is generally impossible or expensive to obtain the reference image in many realistic situations, the BIQA has the broadest range of application scenarios among IQA methods and is of wide applicability. In this paper, we focus on the BIQA.

Commonly, traditional BIQA methods first extract quality-aware features based on hand-crafted descriptors, and then an additional regression model is adopted to map the features into quality score [7]–[10]. However, due to the complexities of distortions and image contents, hand-crafted features are difficult to adequately represent quality degradation. Recently, deep convolutional neural network (DCNN) has achieved great success in many computer vision tasks due to its strong representation ability. Accordingly, DCNN has also been applied to automatically extract quality-aware features for BIQA tasks [11]–[15].

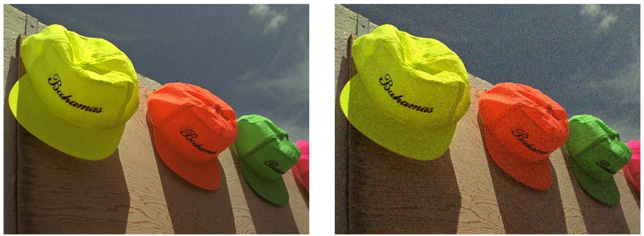
While a large variety of BIQA models have achieved promising improvements in the accuracy of the quality score prediction, there are two drawbacks to representing the image quality with a series of continuous values. On the one hand, semantic description makes up much of human cognition, allowing us better to perceive the visual world [16]. Accordingly, humans prefer to perform semantic description rather than quantitative values to perceive image quality. For example, using the description “high definition” instead of the score “6.6” makes it easier for us to cognize the quality of the image/video. Meanwhile, in [17], it points out many human learning and categorizing models involve discrete representations, leading to faster and sometimes more accurate learning. Moreover, it demonstrates that humans establish a clear distinction between discrete eigenvalues, rather than a continuous boundary. Thus, in our work, discrete description are employed for quality assessment. On the other hand, cognition is a needs-oriented task where humans extract image contents with local to global semantics as they need. For instance, local-details semantics are concerned in edge detection task, while global-concepts semantics are of interest in image classification task. Further, different cognition needs require diverse image quality, e.g., some distortion of local-details can be tolerated in image classification task, but not in edge detection. The current BIQA methods are cognition needs-agnostic and can only represent coarse or holistic image

Manuscript received July 22, 2021; revised January 11, 2022 and March 14, 2022; accepted April 12, 2022. Date of publication May 5, 2022; date of current version May 26, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62022063 and in part by the National Key Research and Development Program of China under Grant 2018AAA0101400. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sebastian Bosse. *(Corresponding authors: Jinjian Wu; Shiwei Tian.)*

Wen Yang, Jinjian Wu, Leida Li, Weisheng Dong, and Guangming Shi are with the School of Artificial Intelligence, Xidian University, Xi’an 710071, China (e-mail: wen.yang@stu.xidian.edu.cn; jinjian.wu@mail.xidian.edu.cn; ldli@xidian.edu.cn; wsdong@mail.xidian.edu.cn; gmshi@xidian.edu.cn).

Shiwei Tian is with the National Innovation Institute of Defense Technology, Beijing 100010, China (e-mail: tianxwell@163.com).

Digital Object Identifier 10.1109/TIP.2022.3171445



QV:5.76; **QC**:local details are basically undistorted, regional contours are basically undistorted and **global concepts are basically undistorted**
(a)

QV:4.25; **QC**:local details are observably distorted, regional contours are slightly distorted and **global concepts are basically undistorted**
(b)

Fig. 1. The images with different quality values satisfy the quality demand for recognition. QV is the quality value of distorted image. QC is the quality caption generated from our method.

quality with mediocre quality scores, which can not reflect degradation on hierarchical semantics, failing to characterize the diverse quality demands of different cognition needs. As shown in Figure 1, though the quality values of two images are different, the global-concepts of them (i.e., understanding the general hats in two images) are basically undistorted, which satisfies the quality demand for recognition. Moreover, despite local-details and regional-contours are distorted to varying degrees in Figure 1 (b), this has little effect on understanding the global-concepts.

In this paper, to comply with human cognition, a novel quality caption model is inventively proposed for fine-grained IQA with hierarchical semantics degradation. Different from previous fine-grained IQA [18], [19], which pays attention to the holistic quality prediction for images in the same distortion level, our fine-grained IQA model is proposed to describe the degradation on hierarchical semantics of images in different distortion levels. Concretely, inspired by the neurocognition model that knowledge is distributed representations consisting of property primitives [20], image quality caption can be represented as patterns of activation distributed over the various degradation on hierarchical semantics. A novel bi-directional relationship-based network (BDRNet) is proposed to activate the distributed binary patterns for image quality description.

This paper extends our previous ACM-MM paper [21], which is named Image Quality Caption with ARSANet. The extensions are multiple folds. First, ARSANet [21] only adopts a down-up interaction mechanism to extract hierarchical semantics by exploring the correlations between them. However, there are hierarchy and reverse hierarchy correlations in visual system [22]. The proposed BDRNet designs a bi-directional hierarchical semantics interaction (BHSI) module to explore the correlations in a bi-directional manner (down-up and up-down). Second, empirical evidence shows that there are strong degradation dependencies between hierarchical semantics. For instance, when local-details are slightly degraded, the global-concepts are probably slightly degraded or not. Inversely, the degradation dependencies also hold. ARSANet [21] only explores that from local to global semantics. The proposed BDRNet designs a bi-directional recurrent degradation attractor (BRDA) module to learn the degradation dependencies with a bi-directional degradation attractor strategy (local to global, and reversed). Third, ARSANet [21] needs to crop images to a fixed size for adapting to the input of deep model. However, identifying the degradation on hierarchical semantics is sensitive to image cropping. To deal with this problem, we introduce the spatial pyramid pooling (SPP) [23] into BDRNet, that does not require cropping the input image to a fixed size, which further improves the prediction accuracy of the trained BDRNet. Finally, we conduct performance comparisons based on distortion types, considering their different effects on hierarchical semantics decay. Besides, we perform cross-database evaluations to illustrate the generalization ability of the proposed model. Quantitative analysis is further conducted to expound the needs-oriented cognition. These experiments are not provided in the ARSANet [21].

Compared to the original ARSANet [21], the proposed metric has a significant improvement in terms of both prediction accuracy and generalization ability. In summary, the main contributions of our work are as follows.

- A bi-directional relationship-based network (BDRNet) is proposed for fine-grained image quality caption with the degradation of hierarchical semantics. Extensive experiments demonstrate that our method outperforms the state-of-the-arts in terms of both evaluation performance and generalization ability.
- A novel BHSI module is proposed to explore the correlations among hierarchical semantics in a bi-directional interaction manner, which is well grounded on the hierarchy and reverse hierarchy process in visual system.
- A new BRDA module is designed to recurrently learn the degradation dependencies between different levels of semantics with a bi-directional degradation attractor strategy.

The remainder of this paper is structured as follows. In Section II, we review the related works on BIQA. Section III describes the details of the proposed model. Experimental results and analysis are presented in Section IV. Finally, we conclude this paper with a discussion in Section V.

II. Related Work

In this section, we briefly review the literature related to our approach, including some BIQA methods and IQA databases.

A. BIQA Algorithms

1) Traditional Blind Image Quality Assessment: Commonly traditional BIQA methods first extract quality-aware features based on hand-crafted descriptors, and then an additional regression model is adopted to map the features into quality score. The natural scene statistics (NSS) based approach is one of the most popular methods. Based on NSS, DIIVINE [24] proposed a 2-stage IQA framework, involving distortion identification followed by distortion-specific quality assessment. BRISQUE [25] proposed an NSS-based distortion-generic BIQA model, which operates in the spatial domain. NIQE [26] extracted quality-aware features from the NSS model and fitted them to a multivariate Gaussian (MVG) model. By integrating the features of NSS derived from multiple cues, IL-NIQE [27] learned an MVG model for quality assessment. Besides the NSS-based approaches, another common way is the human visual system (HVS)-based approaches. Based on the assumption that HVS adapts to structural information, the quality-aware features about gradient, luminance contrast, or local binary pattern are extracted for quality assessment [28]–[31]. However, due to the complexities of distortions and image contents, hand-crafted features are difficult to adequately represent quality degradation.

2) Deep Learning-Based Blind Image Quality Assessment: In recent years, deep convolutional neural network (DCNN) has achieved great success in many computer vision tasks due to its strong representation ability. Accordingly, DCNN has also been applied to automatically extract quality-aware features for BIQA tasks. The existing deep learning-based BIQA methods are mainly constructed in two ways.

The first one only extracts the single level of feature (i.e., the last layer of the DCNN) to predict quality score. Zhang *et al.* [32] proposed a deep bilinear DCNN-based model, which can deal with both the synthetic and authentic distortions by conceptually modeling them as two-factor variations. In order to learn more effective feature representations, a two-stream DCNN that includes two subcomponents for image and gradient image was proposed for BIQA [33]. A meta-learning based BIQA method is proposed in [5], which can learn the shared prior knowledge model and then fine-tune the prior model with unknown distortions. Based on the internal generative mechanism (IGM), an active inference model based on the GAN was proposed to predict the primary content [14], and then the image quality was measured on the basis of the primary content. A Generative Adversarial Network (GAN)-based method is proposed to generate hallucinated reference, then the discrepancy map between hallucinated reference and distorted image was used for quality prediction [34]. However, different levels of distortion generate different degradation on hierarchical features. These methods mentioned above do not consider the degradation on the hierarchical features for assessing image quality.

The second type of deep learning-based BIQA methods extract multi-level features from DCNN for image quality assessment. Gao *et al.* [35] proposed to extract multi-level features from a DCNN model for learning an effective BIQA model. HFD-BIQA [36] extracted the deep semantic features from ResNet [36] to predict image quality. Inspired by the hierarchical perception in the HVS, a cascaded DCNN is proposed to extract the multi-level features for quality prediction [15]. Although hierarchical features are considered, the correlations and degradation dependencies among hierarchical features are not thoroughly explored, which weakens the precision of quality score prediction.

B. IQA Databases

In recent years, several annotated IQA databases have been constructed to provide guidance for the development of IQA algorithms. They can be divided into two main categories: 1) synthetically distorted databases and 2) authentically distorted databases.

1) Synthetically Distorted Databases: LIVE [37] is a popular IQA database containing 779 distorted images generated from 29 reference images degraded by 5 types of distortion. Image quality is labeled using a single-stimulus method with Differential Mean Opinion Score (DMOS), where the lower DMOS indicates higher quality. TID2013 [38] is a large-scale database including 3000 distorted images generated by 25 source reference images with 24 types of distortion under 5 degradation levels. Quality scores are annotated using the competition-like double stimulus procedure with Mean Opinion Score (MOS) values, where a lower MOS denotes bad visual quality. The CSIQ [39] database comprises 899 distorted images generated from 30 reference images degraded by 6 types of distortion. The quality scores are labeled by the double stimulus procedure with DMOS values.

2) Authentically Distorted Databases: LIVE-CH [40] database includes 1162 images taken under real-life conditions, and these images capture a large variety of objects and scenes that are subjected to numerous types of authentic distortions. KonIQ-10k [41] consists of 10,073 images, on which large scale crowdsourcing experiments are performed to obtain reliable quality ratings. BID [42] contains 585 images that provides a realistic scenario to evaluate algorithm effectiveness when assessing the quality of pictures in an actual application.

However, the qualities of distorted images in existing databases are labeled with continuous scores. In our work, the quality caption model is proposed to describe the degradation of hierarchical semantics for quality representation. The existing label cannot satisfy the proposed model. Thus, we will need to relabel the databases with the description of degradation on hierarchical semantics.

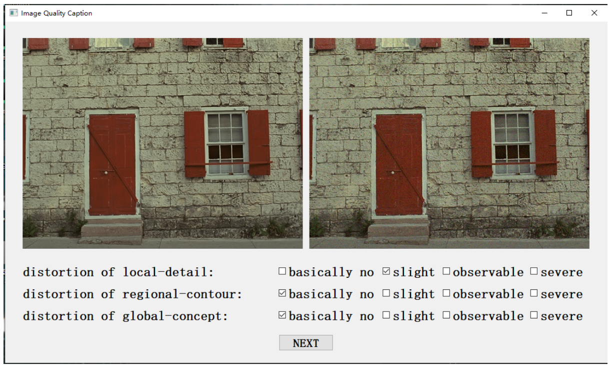
III. Image Quality Caption With BDRNet

In this section, we first introduce the preparation of quality caption databases, including initial image quality caption collection, data processing. And then we present the proposed quality caption model in detail, including the architecture and learning.

A. Database Preparation

1) Initial Image Quality Caption Collection: In this work, image quality is represented by the description of degradation on hierarchical semantics. Concretely, hierarchical semantics are characterized as local-detail semantics (edges), regional-structure semantics (contours), and global-concept semantics (categories). And the degradation levels can be described as basically no distortion, slight distortion, observable distortion, and severe distortion, similar to the ITU-R absolute category rating (ACR) scale [43]. Subjective experiments are adopted to label the image quality caption for some existing databases.

The two commonly used methods in image subjective experiment are single stimulus and double stimulus [44]. The single stimulus method directly depends on the test distorted images to obtain quality ratings, while double stimulus method collects quality ratings by comparing the perceived difference between the reference and distorted images. In our work, three synthetically distorted databases (LIVE [37], TID2013 [38], and CSIQ [39]) and one authentically distorted database (LIVE-CH [40]) are employed for subjective labeling. For the synthetically distorted database, the double stimulus method is



distortion of local-detail: □basically no ☑slight □observable □severe
distortion of regional-contour: ☑basically no □slight □observable □severe
distortion of global-concept: ☑basically no □slight □observable □severe

Fig. 2. Screenshot of interface for subjective experiment. Subjects are required to describe the degradation on hierarchical semantics of image in the right using the image on the left as a reference (if there is one).

employed because there are reference images. On the contrary, the single stimulus method is employed in the authentically distorted database due to the absence of reference images.

The subjective experiments are performed by a specifically designed interface as shown in Figure 2. 25 subjects (10 female and 15 male) are invited to describe the image quality. All participants have no experience with image quality assessment. The interface is played on a pre-programmed computer in a lab with normal illumination and the subjects' votes are recorded. Considering the constraint on survey duration to reduce the effect of the viewers' fatigue, the survey is divided into several sessions for each database (200 images in each session). Then, the subjective experiments can be performed session by session, and the subject is allowed to take as much time as needed to finish all the judgments in one session. Note that a subject can complete the evaluation of several sessions at different times. Before the start of the subjective test, a brief introduction of the objective of this survey and how to do the quality evaluation of hierarchical semantics is first presented to the viewers. Then, subjects are asked to go through some examples of semantics quality evaluation. This essential pre-session helps participants to understand hierarchical semantics and its quality assessment, as well as to stabilize their judgment.

2) Data Processing: We denote the hierarchical semantics as $\{s_1 \ldots s_K\}$ and the degradation levels as $\{w_1 \ldots w_N\}$. Inspired by the distributed model of neurocognition, which presents that knowledge is distributed representations consisting of semantic primitives [20], the diverse degradation on different levels of semantics are defined as degradation primitives $\{s_1w_1, s_1w_2, \ldots, s_Kw_N\}$. Thus, given an image, quality caption can be denoted as the binary activation patterns P distributed over those degradation primitives:

$$P\left\langle s, w\right\rangle = \left\{p_{s_1w_1}, p_{s_1w_2}, \ldots, p_{s_Kw_N}\right\}, \tag{1}$$

where $p_{s_kw_n} = 1$ $(k = 1, 2, \ldots, K; n = 1, 2, \ldots, N)$ indicates that the degradation primitive s_kw_n is activated, and 0 otherwise.

For convenience, we denote P uniformly as $\left\{P^1, P^2, \ldots, P^K\right\}$, in which $P^k = \left\{p_1^k, p_2^k, \ldots, p_N^k\right\}$ is the degradation representation of s_k.

Quality caption of semantics degradation is a relatively complex problem. Hence, there are divergences among observers on the assessment standard and degree of error tolerance. In other words, the subjective data may still be influenced by inconsistent subject behavior and it is necessary to screen the data. As mentioned above, image quality caption is represented as the discrete binary vector P. For each test image, we take the plural by place for P from all subjects to obtain the final quality caption. Under the condition where the assessment is not one-sided, we discard the relevant test images. In fact, 25 subjects reach a general agreement on assessing semantics degradation.

B. The Proposed BDRNet

In this paper, we propose a BDRNet to activate the distributed binary patterns of image quality caption. The network architecture of BDRNet is shown in Figure 3. The procedure of BDRNet mainly consists of two parts: (a) bi-directional hierarchical semantics interaction (BHSI) module, and (b) bi-directional recurrent degradation attractor (BRDA) module. The BHSI module is built to explore the hierarchy and reverse hierarchy correlations between hierarchical semantics in a bi-directional manner (down-up and up-down) for semantics extraction. The BRDA module is designed to learn the degradation dependencies with a bi-directional degradation attractor strategy (local to global, and reversed) for semantics degradation prediction.

1) Bi-Directional Hierarchical Semantics Interaction Module: The hierarchical perception of human visual system (HVS) indicates that there are hierarchy and reverse hierarchy correlations between different levels of semantics (local to global and global to local) [22]. Thus, the BHSI module is designed to explore those correlations in a bi-directional interaction manner for hierarchical semantics extraction. Figure 3 (a) shows the detailed semantics extraction process. Concretely, hierarchical semantics are extracted first from shallow to deep convolutional layers (Stage-1 to Stage-3, where each stage consists of a stack of convolutional layers and a max-pooling layer). The local-details are extracted from Stage-1, regional-contours are extracted from Stage-2 and global-concepts are extracted from Stage-3. For convenience, these hierarchical semantics can be expressed as s_k. Next, a bi-directional semantics refinement (SR) strategy is applied to regularize hierarchical semantics by capturing the hierarchy and reverse hierarchy correlations, which includes down-up semantics refinement (DUSR) and up-down semantics refinement (UDSR). For DUSR, spatial attention [45] is adopted to different levels of semantics to emphasize the region of interest in local space. For UDSR, channel attention [45] is applied to high-level semantics to weight channels during up-down semantics fusion. To avoid the image cropping, which may affect identifying the degradation on hierarchical semantics, the spatial pyramid pooling (SPP) [23] is introduced to build our model, that does not require cropping the input image to a fixed size. Finally, the refined hierarchical semantics are obtained after DUSR and UDSR respectively, denoted as d_k and u_k.

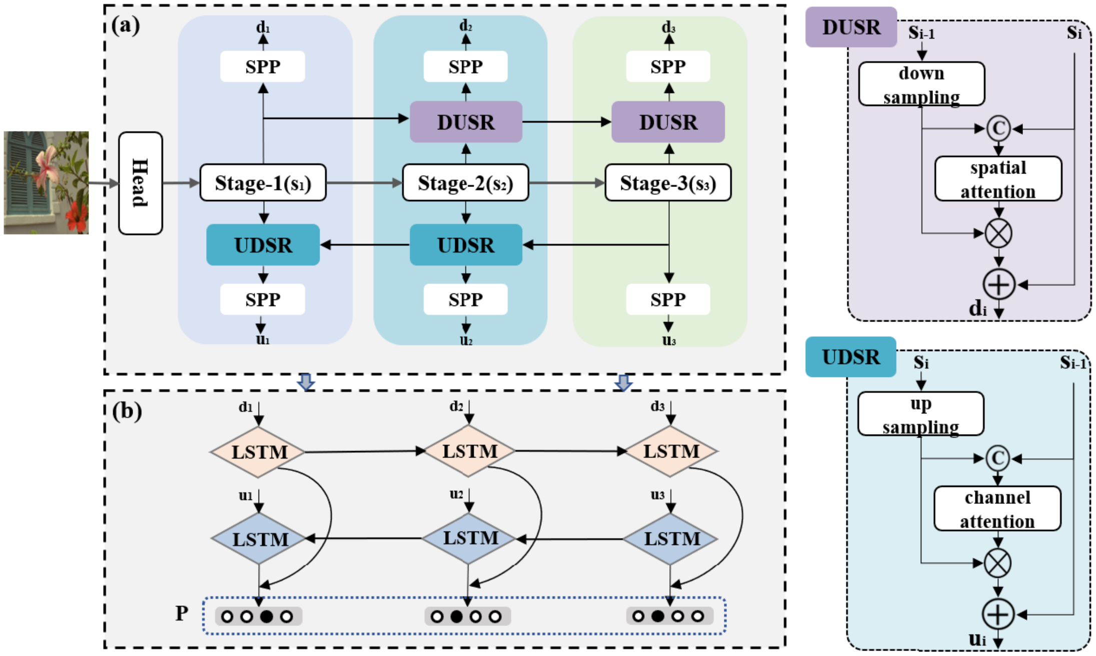


Fig. 3. The schematic of the proposed model: (a) bi-directional hierarchical semantics interaction (BHSI) module, and (b) bi-directional recurrent degradation attractor (BRDA) module.

2) Bi-Directional Recurrent Degradation Attractor Module: Empirical evidence shows that there are strong degradation dependencies between different levels of semantics. For instance, when local-details are slightly degraded, the global-concepts are probably slightly degraded or not. Inversely, this degradation dependency also holds. To mine the degradation dependencies among different levels of semantics, we use the semantics from both the local to global and from the global to local as two kinds of sequential data. Then, the LSTM-based BRDA module with a bi-directional degradation attractor strategy is proposed to exploit the preceding dependency for semantics degradation P prediction. In the following, we will introduce the BRDA module in detail.

LSTM [46] is extensively used in many tasks to capture the sequential characteristics, in which the significant information is encoded at every time step. The basic architecture of the LSTM cell is shown in Figure 4 and the classical computation of it is given as

$$\begin{aligned}
i_t &= \sigma\left(W_{ii}x_t + W_{hi}h_{t-1}\right)\\
f_t &= \sigma\left(W_{if}x_t + W_{hf}h_{t-1}\right)\\
g_t &= \tanh\left(W_{ig}x_t + W_{hg}h_{t-1}\right)\\
o_t &= \sigma\left(W_{io}x_t + W_{ho}h_{t-1}\right)\\
c_t &= \left(f_t \odot c_{t-1} + i_t \odot g_t\right)\\
h_t &= o_t \odot \tanh\left(c_t\right),
\end{aligned} \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function and $\tanh(\cdot)$ is the hyperbolic tangent function. i_t, f_t c_t, and o_t are the input, forget, memory, output state of the LSTM. g_t is the current state. h_t is the hidden state at time step t.

The BRDA captures the degradation dependencies using a bi-directional LSTM (BiLSTM) [47], by which degradation dependencies of hierarchical semantics in both local to global and global to local directions are leveraged. The architecture of BRDA is visualized in Figure 3 (b). Take the local to global direction as an example, when the hierarchical semantics d_k are sequentially sent into the LSTM, the useful degradation information of previous $(k-1)$ levels of semantics can be encoded by the memory cell c_t. By using the memorized information of previous ones, LSTM network can easily predict degradation on the current semantics. Therefore, the degradation dependencies can be captured by the BRDA module to predict semantics degradation P. The specific process can be expressed as

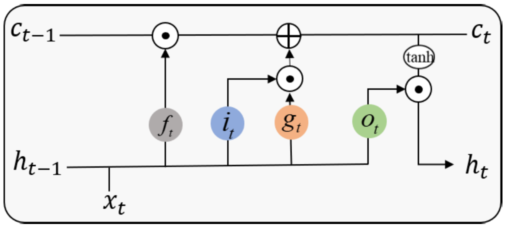


Fig. 4. A basic architecture of LSTM cell.

$$\left\{\hat{P}_d^k\right\}_{k=1}^{K} = BRDA\left(\{d_k\}_{k=1}^{K}\right), K = 3 \tag{3}$$

where $\hat{P}_d^k = \left\{\hat{p}_1^k, \hat{p}_2^k, \ldots, \hat{p}_N^k\right\} (N = 4)$ is the predicted degradation representation of hierarchical semantics s_k in local to global direction.

Similarly, for the global to local direction, we can obtain the $\left\{\hat{P}_u^k\right\}_{i=1}^{K}$. Here, unit-wise max-pooling is adopted to get the final degradation representation $\hat{P}^k$ of s_k, which is formulated as

$$\hat{P}^k = \max\left(\hat{P}_d^k, \hat{P}_u^k\right) \tag{4}$$

Finally, semantics degradation can be expressed as

$$\hat{P} = \left\{\hat{P}^k\right\}_{k=1}^{K}, K = 3 \tag{5}$$

3) Loss Function: We minimize the residuals between the predicted activation pattern $\hat{P}$ and the corresponding ground-truth activation pattern P to learn the correct binary activation pattern of degradation primitives. The cross-entropy loss is adopted to measure the residuals, which can be defined as

$$L^k(P^k, \hat{P}^k) = -\sum_{n=1}^{N} p_n^k \log\left(\hat{p}_n^k\right) + \left(1 - p_n^k\right)\log\left(1 - \hat{p}_n^k\right)$$
$$L_{all} = \sum_{k=1}^{K} L^k. \tag{6}$$

IV. Experiments and Evaluations

A. Experimental Settings

1) Implementation Details: We use PyTorch framework to implement the proposed method. For the BDRNet training, to avoid the image cropping, which may affect identifying the degradation on hierarchical semantics, the spatial pyramid pooling (SPP) [23] is introduced to build our model, which does not require cropping the input image to a fixed size. In other words, we can input the images with the original size. It should be noted that because of the inconsistent size of the images in the LIVE database, we have to crop these images to get a uniform size in order to meet the same size in a batch. Besides, to avoid overfitting when training on small-scale databases, the semantic extraction module (Head, Stage-1 to Stage-3) is first pretrained on the ImageNet [48], and then the BDRNet is fine-tuned on the IQA database to ease the overfitting problem. We use the ADAM [49] optimization algorithm with 16 mini-batches to optimize the model. The BDRNet is trained iteratively over 200 epochs with a learning rate of 10^{-5}, which is dropped by a factor of 0.1 every 100 epochs. For the BDRNet testing, we take all test images in any database at their original size as input, with a batch size of 1.

2) Evaluation Metrics: Three widely accepted metrics are used to objectively evaluate the performance of our quality caption model, which are BLEU, Meteor, and Rouge-L [50]. Based on *Precision* or *Recall*, or both, these metrics aim to measure the consistency between the generated and ground-truth descriptions. BLEU is a popular machine translation metric that analyzes the co-occurrences of n-grams between the candidate and reference sentences. It computes a corpus level clipped n-gram precision between sentences. Meteor is calculated by generating an alignment between the words in the candidate and reference sentences, with an aim of 1:1 correspondence. Rouge-L uses a measure based on the Longest Common Subsequence (LCS). An LCS is a set of words shared by two sentences that occur in the same order.

Note that, for general image caption tasks, these metrics are evaluated on the basis of 'word'. In our work, we represent the image quality caption by the binary patterns of activation distributed over the degradation primitives, i.e., a binary vector P. Thus, we use the binary value in P instead of 'word' for performance evaluation. Generally, the higher the evaluation score, the better the performance.

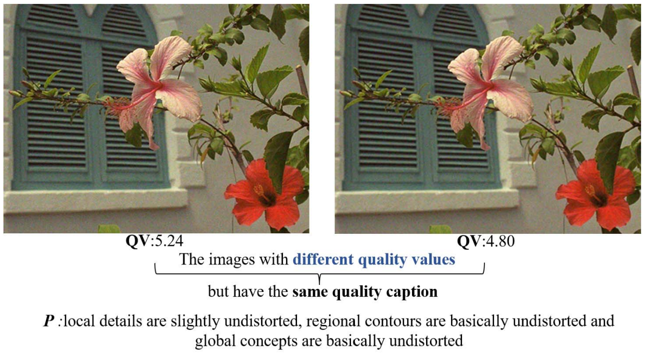

QV:5.24
QV:4.80
The images with different quality values
but have the same quality caption
P :local details are slightly undistorted, regional contours are basically undistorted and global concepts are basically undistorted

Fig. 5. The images with different quality values have the same quality caption label P. QV is the quality value of distorted image. P is quality caption generated from subjective experiment.

B. How to Compare the Continuous Metrics and Discrete Quality Caption?

The existing BIQA metrics typically produce objective scores to represent image quality. To compare with existing BIQA methods, the correlation between the objective continuous scores Q and the discrete quality captions P should be evaluated. As introduced in [52], the rate of classification errors is one parameter that can evaluate the effectiveness of the metrics. The classification errors occur when the objective quality score Q is converted to quality caption P' which is inconsistent with the subjective quality caption P. A BIQA metric which provides a higher Correct Classification Rate (CCR) is a better match with discrete quality caption.

In general, even if the quality scores of the two distorted images are not exactly same, they have the same subjective quality caption P (shown in Figure 5). This means that the quality scores might have uncertain accuracy when converted to P. In other words, a higher quality score does not necessarily lead to a better quality description.

In order to minimize that uncertainty in the quality scores, the continuous scores are quantified into multiple interval ranges $\{\Delta_i\}$, where i is the number of types of P. Then, based on the multiple interval ranges, objective quality scores Q can be classified into different quality caption P. It is clear that as the interval ranges $\{\Delta_i\}$ change, the CCR also changes. For instance, with very large Δ_i, most of the test images are classified as the same P. To select the best $\{\Delta_i\}$, Hanhart *et al.* [53] suggested setting $\{\Delta_i\}$ to achieve the Maximum Correct Classification Rate (MCCR). We first give an initial $\{\Delta_{i0}\}$, and then adjust the $\{\Delta_i\}$ by traversing the exhaustive objective scores of each test image to acquire the MCCR.

We obtain the MCCR in two approaches.

1) Approach 1: A common $\{\Delta_i\}$ is applied for all types of distortion in a database.

2) Approach 2: The different $\{\Delta_i\}$ is applied for different types of distortion in a database. In this approach, associated with each distortion type in a database, we have one $\{\Delta_i\}$.

TABLE I
MCCR BASED COMPARISON ON TID2013 DATABASE (APPROACH 1). THE FORMAT OF THE RESULT IS MEAN (STD)

	BLEU-1	BLEU-2	BLEU-3	Rouge-L	Meteor
ILNIQE [27]	0.650 (0.0093)	0.556 (0.0105)	0.325 (0.0126)	0.673 (0.0091)	0.599 (0.0096)
NIQE [26]	0.618 (**0.0088**)	0.461 (**0.0097**)	0.217 (0.0101)	0.641 (**0.0084**)	0.536 (**0.0090**)
HOSA [51]	0.751 (0.0524)	0.652 (0.0608)	0.422 (0.0707)	0.774 (0.0444)	0.690 (0.0554)
DIIVINE [24]	0.720 (0.0429)	0.562 (0.0615)	0.317 (0.0626)	0.750 (0.0348)	0.657 (0.0511)
BRISQUE [25]	0.700 (0.0240)	0.583 (0.0332)	0.319 (0.0456)	0.743 (0.0204)	0.650 (0.0278)
CaHDC [15]	0.777 (0.0511)	0.654 (0.0876)	0.470 (0.1503)	0.810 (0.0637)	0.713 (0.0885)
DBCNN [32]	0.719 (0.0409)	0.659 (0.0549)	0.477 (0.0467)	0.753 (0.0463)	0.685 (0.0714)
Two-stream [33]	0.822 (0.0251)	0.743 (0.0269)	0.583 (0.0523)	0.843 (0.0263)	0.779 (0.0302)
WaDIQaM [11]	0.817 (0.0327)	0.733 (0.0552)	0.565 (0.1288)	0.825 (0.0054)	0.738 (0.0340)
ARSANet [21]	0.894 (0.0142)	0.821 (0.0235)	0.765 (0.0377)	0.892 (0.0130)	0.856 (0.0554)
BDRNet	**0.928** (0.0110)	**0.890** (0.0102)	**0.835** (**0.0091**)	**0.935** (0.0094)	**0.899** (0.0176)

TABLE II
MCCR BASED COMPARISON ON CSIQ DATABASE (APPROACH 1). THE FORMAT OF THE RESULT IS MEAN (STD)

	BLEU-1	BLEU-2	BLEU-3	Rouge-L	Meteor
ILNIQE [27]	0.642 (0.0204)	0.600 (0.0211)	0.450 (0.0203)	0.645 (0.0204)	0.617 (0.0203)
NIQE [26]	0.458 (0.0388)	0.365 (0.0348)	0.235 (0.0408)	0.498 (0.0360)	0.410 (0.0370)
HOSA [51]	0.702 (0.0301)	0.607 (0.0405)	0.405 (0.0470)	0.736 (0.0268)	0.655 (0.0347)
DIIVINE [24]	0.755 (0.0486)	0.681 (0.0527)	0.541 (0.0477)	0.786 (0.0455)	0.717 (0.0501)
BRISQUE [25]	0.748 (0.0252)	0.670 (0.0356)	0.498 (0.0457)	0.777 (0.0233)	0.709 (0.0292)
CaHDC [15]	0.760 (0.0504)	0.694 (0.0533)	0.573 (0.0717)	0.783 (0.0495)	0.718 (0.0521)
DBCNN [32]	0.695 (0.0561)	0.625 (0.0535)	0.428 (0.0333)	0.722 (0.0555)	0.652 (0.0384)
Two-stream [33]	0.812 (0.0222)	0.719 (0.0664)	0.590 (0.1393)	0.842 (0.0240)	0.777 (0.0914)
WaDIQaM [11]	0.822 (0.0327)	0.746 (0.0552)	0.610 (0.1288)	0.849 (**0.0054**)	0.789 (0.0340)
ARSANet [21]	0.910 (**0.0102**)	0.865 (0.0193)	0.800 (0.0181)	0.918 (0.0105)	0.881 (**0.0124**)
BDRNet	**0.937** (0.0123)	**0.899** (**0.0178**)	**0.835** (**0.0178**)	**0.941** (0.0104)	**0.908** (0.0136)

TABLE III
MCCR BASED COMPARISON ON LIVE DATABASE (APPROACH 1). THE FORMAT OF THE RESULT IS MEAN (STD)

	BLEU-1	BLEU-2	BLEU-3	Rouge-L	Meteor
ILNIQE [27]	0.581 (0.0192)	0.434 (0.0249)	0.289 (0.0236)	0.550 (0.0177)	0.529 (0.0195)
NIQE [26]	0.458 (0.0196)	0.371 (0.0266)	0.282 (0.0172)	0.494 (0.0180)	0.417 (0.0226)
HOSA [51]	0.711 (0.0426)	0.593 (0.0532)	0.394 (0.0620)	0.746 (0.0303)	0.635 (0.0430)
DIIVINE [24]	0.688 (0.0450)	0.450 (0.0438)	0.337 (0.0479)	0.729 (0.0360)	0.572 (0.0424)
BRISQUE [25]	0.724 (0.0306)	0.500 (0.0447)	0.403 (0.0513)	0.764 (0.0244)	0.636 (0.0314)
CaHDC [15]	0.752 (0.0519)	0.555 (0.0850)	0.417 (0.0619)	0.780 (0.0519)	0.662 (0.0142)
DBCNN [32]	0.714 (0.0817)	0.517 (0.0799)	0.410 (0.0709)	0.763 (0.0720)	0.641 (0.0865)
Two-stream [33]	0.799 (0.0294)	0.585 (0.0464)	0.474 (0.0674)	0.837 (0.0326)	0.714 (0.0343)
WaDIQaM [11]	0.803 (0.0416)	0.612 (0.0448)	0.519 (0.0093)	0.824 (0.0166)	0.719 (0.0468)
ARSANet [21]	0.902 (**0.0135**)	0.853 (0.0213)	0.771 (0.0299)	0.919 (0.0109)	0.871 (0.0157)
BDRNet	**0.914** (0.0156)	**0.880** (**0.0206**)	**0.812** (**0.0164**)	**0.923** (**0.0108**)	**0.887** (**0.0153**)

C. Comparison of Discrete BDRNet and Continuous BIQA Methods Based on MCCR

Some state-of-the-art BIQA methods are chosen for performance comparison, including traditional BIQA methods (NIQE [26], ILNIQE [27], HOSA [51], DIIVINE [24] and BRISQUE [25]), and DCNN-based BIQA methods (CaHDC [15], DBCNN [32], Two-stream [33], WaDIQaM [11] and our previous ARSANet [21]). The objective scores of traditional BIQA methods are calculated by the source codes downloaded from the authors' homepage and those of deep learning-based methods are generated from publicly available trained models.

For each database, we randomly select 80% distorted images for training and the remaining 20% for testing by reference images for no overlapping in context. For the BIQA methods being compared, having computed the objective scores on the testing set, the MCCR based performances are obtained. Our previous ARSANet [21] and proposed BDRNet is trained on the training set and tested on the testing set. To avoid the bias due to the data separating way, the random split of the dataset is repeated 100 sessions, and both the average value and the standard deviation (STD) value are reported.

1) Comparison Based on Approach 1: We first compare performance based on the above stated Approach 1. The performance comparisons among the four databases are shown in Tables I-IV. And the best results are in bold and the second best results are underlined. It can be easily observed from Tables I-IV that the proposed BDRNet is significantly

TABLE IV
MCCR BASED COMPARISON ON LIVE-CH DATABASE (APPROACH 1). THE FORMAT OF THE RESULT IS MEAN (STD)

	BLEU-1	BLEU-2	BLEU-3	Rouge-L	Meteor
ILNIQE [27]	0.512 (0.0139)	0.234 (0.0201)	0.131 (0.0222)	0.603 (0.0123)	0.371 (0.0158)
NIQE [26]	0.642 (0.0123)	0.458 (0.0170)	0.217 (0.0251)	0.683 (0.0111)	0.540 (0.0135)
HOSA [51]	0.697 (0.0151)	0.605 (0.0218)	0.398 (0.0301)	0.717 (0.0115)	0.643 (0.0177)
DIIVINE [24]	0.660 (0.0138)	0.556 (0.0208)	0.324 (0.0285)	0.684 (0.0114)	0.600 (0.0167)
BRISQUE [25]	0.691 (0.0140)	0.604 (0.0199)	0.480 (0.0281)	0.713 (0.0114)	0.640 (0.0162)
CaHDC [15]	0.748 (0.0651)	0.675 (0.0778)	0.423 (0.0183)	0.790 (0.0696)	0.714 (0.0584)
DBCNN [32]	0.769 (0.0438)	0.669 (0.0577)	0.408 (0.0440)	0.789 (0.0355)	0.711 (0.0202)
Two-stream [33]	0.770 (0.0193)	0.700 (0.0243)	0.497 (0.0579)	0.793 (0.0193)	0.727 (0.0373)
WaDIQaM [11]	0.765 (0.0284)	0.719 (0.0579)	0.634 (0.1351)	0.875 (0.0211)	0.830 (0.0402)
ARSANet [21]	0.857 (0.0120)	0.812 (0.0195)	0.635 (0.0248)	0.870 (0.0115)	0.818 (0.0138)
BDRNet	**0.878** (**0.0087**)	**0.837** (**0.0110**)	**0.779** (**0.0150**)	**0.897** (**0.0078**)	**0.848** (**0.0090**)

TABLE V
MCCR BASED COMPARISON ON TID2013 DATABASE (APPROACH 2). THE FORMAT OF THE RESULT IS MEAN (STD)

	BLEU-1	BLEU-2	BLEU-3	Rouge-L	Meteor
ILNIQE [27]	0.667 (0.0074)	0.578 (**0.0081**)	0.370 (0.0137)	0.674 (0.0079)	0.618 (0.0075)
NIQE [26]	0.661 (0.0114)	0.570 (0.0081)	0.350 (**0.0082**)	0.679 (**0.0031**)	0.610 (**0.0056**)
HOSA [51]	0.743 (0.0444)	0.646 (0.0422)	0.421 (0.0498)	0.775 (0.0375)	0.688 (0.0424)
DIIVINE [24]	0.771 (0.0337)	0.689 (0.0558)	0.522 (0.0648)	0.810 (0.0275)	0.724 (0.0435)
BRISQUE [25]	0.750 (0.0183)	0.673 (0.0269)	0.530 (0.0404)	0.790 (0.0152)	0.704 (0.0218)
CaHDC [15]	0.800 (**0.0057**)	0.732 (0.0108)	0.571 (0.0111)	0.838 (0.0142)	0.758 (0.0073)
DBCNN [32]	0.758 (0.0626)	0.660 (0.0564)	0.465 (0.0651)	0.825 (0.0937)	0.708 (0.0922)
Two-stream [33]	0.853 (0.0117)	0.772 (0.0125)	0.633 (0.0134)	0.883 (0.0103)	0.807 (0.0120)
WaDIQaM [11]	0.866 (0.0109)	0.782 (0.0148)	0.662 (0.0252)	0.872 (0.0110)	0.800 (0.0125)
ARSANet [21]	0.894 (0.0142)	0.821 (0.0235)	0.765 (0.0377)	0.892 (0.0130)	0.856 (0.0554)
BDRNet	**0.928** (0.0091)	**0.890** (0.0102)	**0.835** (0.0091)	**0.935** (0.0064)	**0.899** (0.0110)

TABLE VI
MCCR BASED COMPARISON ON CSIQ DATABASE (APPROACH 2). THE FORMAT OF THE RESULT IS MEAN (STD)

	BLEU-1	BLEU-2	BLEU-3	Rouge-L	Meteor
ILNIQE [27]	0.662 (0.0191)	0.618 (0.0190)	0.479 (0.0180)	0.663 (0.0191)	0.632 (0.0187)
NIQE [26]	0.508 (0.0334)	0.405 (0.0305)	0.290 (0.0365)	0.540 (0.0324)	0.453 (0.0319)
HOSA [51]	0.751 (0.0274)	0.671 (0.0344)	0.482 (0.0402)	0.778 (0.0244)	0.706 (0.0300)
DIIVINE [24]	0.777 (0.0464)	0.709 (0.0496)	0.562 (0.0454)	0.816 (0.0441)	0.739 (0.0472)
BRISQUE [25]	0.772 (0.0226)	0.701 (0.0308)	0.580 (0.0392)	0.803 (0.0207)	0.738 (0.0262)
CaHDC [15]	0.787 (0.0353)	0.724 (0.0369)	0.617 (0.0407)	0.814 (0.0301)	0.751 (0.0345)
DBCNN [32]	0.741 (0.0315)	0.662 (0.0371)	0.485 (0.0571)	0.776 (0.0328)	0.695 (0.0336)
Two-stream [33]	0.859 (0.0133)	0.787 (0.0206)	0.681 (0.0179)	0.878 (0.0114)	0.822 (0.0150)
WaDIQaM [11]	0.843 (0.0275)	0.766 (0.0322)	0.632 (0.0483)	0.855 (0.0237)	0.809 (0.0284)
ARSANet [21]	0.910 (**0.0102**)	0.865 (0.0193)	0.800 (0.0181)	0.918 (0.0106)	0.881 (**0.0124**)
BDRNet	**0.937** (0.0123)	**0.899** (**0.0178**)	**0.835** (**0.0178**)	**0.941** (**0.0105**)	**0.908** (0.0126)

better than other metrics in terms of overall performance (average results) on all databases. And the proposed BDRNet can also obtain the highest or second highest performance stability in most of the databases. Moreover, the experimental results also demonstrate that the proposed BDRNet based on the bi-directional relationship is superior to our previous ARSANet [21]. This is mainly because the correlations among hierarchical semantics and semantic degradation dependencies are fully explored.

2) Comparison Based on Approach 2: The performance comparisons are also conducted with the above mentioned Approach 2. The results are given in Tables V-VI. And the best results are in bold and the second best results are underlined. Since the LIVE-CH database is distorted by authentic distortion, it is difficult to distinguish the type of distortion, so there is no comparison based on Approach 2. These tables show that the proposed BDRNet still outperforms other methods on TID2013, LIVE, and CSIQ databases. Another important thing these tables show is the higher performance achieved by the existing BIQA methods through Approach 2 compared to Approach 1. The result is as expected. Since different types of distortion have varying effects on the image degradation, resulting in different quality score changes. That is to say, with the same P, the Δ_i is different for different distortions.

TABLE VII
MCCR BASED COMPARISON ON LIVE DATABASE (APPROACH 2). THE FORMAT OF THE RESULT IS MEAN (STD)

	BLEU-1	BLEU-2	BLEU-3	Rouge-L	Meteor
ILNIQE [27]	0.624 (0.0220)	0.455 (0.0268)	0.318 (0.0315)	0.617 (0.0199)	0.555 (0.0233)
NIQE [26]	0.544 (0.0199)	0.423 (0.0246)	0.288 (0.0255)	0.611 (0.0180)	0.484 (0.0219)
HOSA [51]	0.735 (0.0402)	0.566 (0.0436)	0.453 (0.0569)	0.782 (0.0274)	0.651 (0.0394)
DIIVINE [24]	0.729 (0.0444)	0.547 (0.0368)	0.424 (0.0444)	0.769 (0.0346)	0.642 (0.0414)
BRISQUE [25]	0.754 (0.0295)	0.579 (0.0348)	0.513 (0.0429)	0.781 (0.0223)	0.672 (0.0295)
CaHDC [15]	0.794 (0.0375)	0.623 (0.0291)	0.528 (0.0422)	0.822 (0.0241)	0.709 (0.0314)
DBCNN [32]	0.762 (0.0694)	0.618 (0.0929)	0.502 (0.0264)	0.819 (0.0694)	0.697 (0.0262)
Two-stream [33]	0.829 (0.0193)	0.645 (0.0243)	0.564 (0.0579)	0.870 (0.0193)	0.736 (0.0373)
WaDIQaM [11]	0.838 (0.0211)	0.635 (0.0246)	0.572 (0.0413)	0.881 (0.0132)	0.765 (0.0207)
ARSANet [21]	0.902 (0.0135)	0.853 (0.0213)	0.771 (0.0299)	0.919 (**0.0109**)	0.871 (**0.0157**)
BDRNet	**0.914** (**0.0126**)	**0.880** (**0.0206**)	**0.812** (**0.0164**)	**0.923** (0.0118)	**0.887** (0.0163)

TABLE VIII
COMPARISON OF BDRNET AND OTHER BIQA MODELS ON QUALITY CAPTION

	TID2013		LIVE		CSIQ		LIVE-CH	
	Rouge-L	Meteor	Rouge-L	Meteor	Rouge-L	Meteor	Rouge-L	Meteor
CaHDC [15]	0.705	0. 620	0.647	0.510	0.788	0.720	0.625	0.469
DBCNN [32]	0.570	0. 415	0.558	0.469	0.720	0.652	0.615	0.459
Two-stream [33]	0.680	0. 535	0.611	0.525	0.797	0.738	0.634	0.494
WaDIQaM [11]	0.672	0. 625	0.655	0.632	0.723	0.692	0.740	0.699
BDRNet	**0.935**	**0.899**	**0.923**	**0.887**	**0.941**	**0.908**	**0.897**	**0.848**

Approach 2 assists to remove distortion type dependency of the objective BIQA methods, leading to maximum performance.

D. Comparison of BDRNet and Other BIQA Models on Quality Caption

To verify the effectiveness of our model structure, the performances of some DCNN-based BIQA models (CaHDC [15], DBCNN [32], Two-stream [33] and WaDIQaM [11]) trained on the newly labeled databases for quality caption P are compared. Concretely, the output of these BIQA models (i.e., a regression quality value) is replaced with the same as ours (i.e., a binary activation pattern P). As mentioned above, each database is divided into the training set (80% distorted images) and the testing set (20% distorted images). For all compared algorithms, the details of the experiment are consistent with our approach. Table VIII shows the experimental results on four datasets (TID2013, LIVE, CSIQ, and LIVE-CH). As can be seen, the proposed BDRNet performs significantly better than other BIQA models on all four databases. The DBCNN [32], Two-stream [33] and WaDIQaM [11] only extract the single level of feature (i.e., the last layer of the DCNN) for image quality description, which does not consider hierarchical features. Although the hierarchical features are extracted by CaHDC [15] to predict quality caption, the dependencies among hierarchical semantics degradation are not considered. Benefiting from the BRDA module, degradation dependencies between different levels of semantics are considered in our work for generating quality caption P.

E. Ablation Experiments

In the BDRNet, we propose a BHSI module to explore the correlations among hierarchical semantics for semantics extraction and a BRDA module to learn the degradation dependencies between different levels of semantics for semantics degradation prediction. In this subsection, the effectiveness of the proposed BHSI and BRDA is verified with ablation experiment. In this ablation study, with four different models, the ablation experiments are conducted on four databases (i.e., TID2013, LIVE, CSIQ, and LIVE-CH). A *Baseline* version is first experimented, which neglects semantics correlations and degradation dependencies. The second is *Baseline+BHSI*, which regularizes hierarchical semantics in a bi-directional interaction manner via BHSI. The third is *Baseline+BRDA*, which considers semantics degradation dependencies with a bi-directional degradation attractor strategy. The fourth is the proposed *Baseline+BHSI+BRDA*, which considers both semantics correlations and degradation dependencies.

These four models are trained and evaluated through the same procedure as depicted above, and the experimental results are given in Table IX, from which we can make several useful conclusions. Firstly, the bi-directional semantics refinement strategy is effective for quality caption prediction. The BHSI composed of the channel and spatial attention is applied to regularize hierarchical semantics, which can focus on relevant information meanwhile suppress irrelevant information during refinement, resulting in higher performance. Secondly, the *Baseline+BRDA* obtains significantly higher performance than the *Baseline*, which demonstrates that considering hierarchical semantics degradation dependencies is a crucial part of the BDRNet. Thirdly, the results are improved when we consider both semantics correlations and degradation dependencies (*Baseline+BHSI+BRDA*).

F. Generalization Performance Evaluations

In general, an effective BIQA method should not only work well on the training database but can be well generalized

TABLE IX

ABLATION EXPERIMENTS

	TID2013		LIVE		CSIQ		LIVE-CH	
	Rouge-L	Meteor	Rouge-L	Meteor	Rouge-L	Meteor	Rouge-L	Meteor
Baseline	0.861	0.801	0.864	0.792	0.870	0.821	0.826	0.751
Baseline+BHSI	0.893	0.844	0.894	0.827	0.916	0.875	0.851	0.789
Baseline+BRDA	0.907	0.855	0.916	0.859	0.925	0.890	0.861	0.795
Baseline+BHSI+BRDA	**0.935**	**0.899**	**0.923**	**0.887**	**0.941**	**0.908**	**0.897**	**0.848**

TABLE X

CROSS DATABASE EVALUATION WHEN TRAINED ON TID2013 DATABASE

Train	TID2013					
Test	LIVE		CSIQ		LIVE-CH	
	Rouge-L	Meteor	Rouge-L	Meteor	Rouge-L	Meteor
CaHDC [15]	0.667	0. 591	0.642	0.546	0.729	0.664
DBCNN [32]	0.546	0.496	0.521	0.462	0.729	0.668
Two-stream [33]	0.660	0.555	0.657	0.541	0.747	0.677
WaDIQaM [11]	0.626	0.547	0.613	0.497	0.717	0.627
ARSANet [21]	0.747	0.625	0.754	0.670	0.755	0.696
BDRNet	**0.805**	**0.694**	**0.797**	**0.738**	**0.761**	**0.703**

TABLE XI

CROSS DATABASE EVALUATION WHEN TRAINED ON LIVE DATABASE

Train	LIVE					
Test	TID2013		CSIQ		LIVE-CH	
	Rouge-L	Meteor	Rouge-L	Meteor	Rouge-L	Meteor
CaHDC [15]	0.695	0.625	0.785	0.718	0.746	0.691
DBCNN [32]	0.706	0.642	0.644	0.614	0.721	0.656
Two-stream [33]	0.705	0.629	0.730	0.649	0.738	0.692
WaDIQaM [11]	0.681	0.621	0.810	0.744	0.750	0.690
ARSANet [21]	0.770	0.687	0.821	0.753	0.776	0.705
BDRNet	**0.792**	**0.706**	**0.849**	**0.799**	**0.790**	**0.717**

TABLE XII

CROSS DATABASE EVALUATION WHEN TRAINED ON CSIQ DATABASE

Train	CSIQ					
Test	TID2013		LIVE		LIVE-CH	
	Rouge-L	Meteor	Rouge-L	Meteor	Rouge-L	Meteor
CaHDC [15]	0.697	0.627	0.722	0.581	0738.	0.673
DBCNN [32]	0.681	0.611	0. 575	0.499	0.701	0.625
Two-stream [33]	0.708	0.632	0.762	0.634	0.735	0.686
WaDIQaM [11]	0.683	0.624	0. 670	0.521	0.748	0.661
ARSANet [21]	0.742	0.662	0.775	0.682	0.767	0.699
BDRNet	**0.803**	**0.714**	**0.818**	**0.735**	**0.780**	**0.707**

TABLE XIII

CROSS DATABASE EVALUATION WHEN TRAINED ON LIVE-CH DATABASE

Train	LIVE-CH					
Test	TID2013		LIVE		CSIQ	
	Rouge-L	Meteor	Rouge-L	Meteor	Rouge-L	Meteor
CaHDC [15]	0.685	0. 612	0.654	0.501	0.659	0.598
DBCNN [32]	0.691	0626496	0.634	0.492	0.658	0.588
Two-stream [33]	0.697	0.628	0.614	0.484	0.648	0.550
WaDIQaM [11]	0.697	0.615	0.679	0.535	0.668	0.612
ARSANet [21]	0.709	0.649	0.690	0.553	0.678	0.626
BDRNet	**0.731**	**0.654**	**0.711**	**0.593**	**0.697**	**0.647**

to other databases. To prove the generalization ability of the proposed BDRNet, cross-database experiments are conducted. Concretely, the model is trained on one database and tested on other databases. Since generalizability is a significant problem for deep-learning-based algorithms, some DCNN-based BIQA models (CaHDC [15], DBCNN [32], Two-stream [33], WaDIQaM [11] and our previous ARSANet [21]) are compared, and the MCCR based performances are obtained. And the best results are in bold and the second best results are underlined.

We first compare the generalizability of BDRNet with other BIQA methods when training on the TID2013 and testing on the other three databases. As can be seen in Table X, the proposed BDRNet performs better than other BIQA methods on LIVE, CSIQ, and LIVE-CH databases. Table XI shows the results of training on LIVE database and testing on TID2013, CSIQ, and LIVE-CH databases. We can observe that the proposed method achieves the best performance on all test databases compared to other BIQA methods. The results of training on CSIQ database and testing on the other three databases are presented in Table XII, from which we can see that the proposed BDRNet achieves the best results.

Further, the model is trained on authentically distorted database and tested on synthetically distorted database to illustrate the generalization capability. The data distribution of a synthetic database is different from that of an authentic database. Thus, it is a challenge to obtain good results when training on LIVE-CH and testing on the TID2013, CSIQ, and LIVE. The performances trained on LIVE-CH are listed in Table XIII. As we can see, the BDRNet gets improved performance on all test databases compared to other BIQA methods.

In summary, the proposed BDRNet shows good generalization to predict the quality caption on many unseen distortion types not in the training set compared to other BIQA methods, since the model learns some features shared by the invisible distortion types and the ones in the training set. In particular, the proposed BDRNet is significantly superior to ARSANet [21]. The reason may be that the size of input images in BDRNet is larger than the ARSANet [21], and our BDRNet with SPP [23] module can greatly improve the evaluation performance by inputting images with original size.

G. Analysis of Needs-Oriented Cognition

Cognition is a needs-oriented task. Further, different cognition needs require diverse image quality. The current BIQA methods can only represent coarse image quality with mediocre quality scores, and can not reflect degradation on hierarchical semantics, which fails to characterize the diverse quality demands of different cognition needs. In our previous ARSANet [21], the qualitative analyses are presented to

TABLE XIV
QUANTITATIVE ANALYSIS OF NEEDS-ORIENTED COGNITION

	GB		WGN		JP2K		JPEG	
	TOP-1.Acc	TOP-5.Acc	TOP-1.Acc	TOP-5.Acc	TOP-1.Acc	TOP-5.Acc	TOP-1.Acc	TOP-5.Acc
level-0	83.66	97.17	83.75	97.16	83.30	96.96	82.66	97.00
level-1	83.64	97.16	83.74	97.14	83.29	96.95	82.64	96.98
level-2	82.46	96.10	81.72	96.00	81.60	95.10	80.90	95.52
level-3	70.58	90.80	72.82	91.88	70.28	91.40	73.92	92.44

TABLE XV
COMPLEXITY ANALYSIS

Method	Rouge-L	Parameters (M)	FLOPs (G)
CaHDC [15]	0.838	0.90	2.296
DBCNN [32]	0.753	15.46	33.017
Two-stream [33]	0.883	1.38	5.128
WaDIQaM [11]	0.872	5.38	6.632
BDRNet	0.935	7.73	4.526

demonstrate that the proposed quality caption model can characterize quality demands for needs-oriented cognition. In this paper, taking the image classification task as an example, we conduct a quantitative analysis for needs-oriented cognition.

As mentioned above, intact-global concepts are required in the image classification task, permitting details and contours to be somewhat distorted. That is, as long as the global-concept is basically free of degradation or no degradation, the classification accuracy will not be affected even if there is some degradation in local-details and regional-contours.

We first classify different degradation levels of the images: **level-0**-the semantics of all levels are not decayed; **level-1**-the global-concepts are basically not decayed, allowing some degradation on details and contours; **level-2**-the global-concepts are slightly decayed; **level-3**-the global-concepts are observably decayed. Then, we select 100 categories of images from ImageNet [48] as distortion-free images, including the training set and validation set. Next, each distortion-free image is degraded by 4 types of distortion (i.e., Gaussian blur (GB), white Gaussian noise (WGN), JPEG compression (JPEG), JP2K compression (JP2K)) as done in CaHDC [15]. Following, for each distortion type, we use the proposed BDRNet to select the distorted images corresponding to **level-0** to **level-3**. Note that the distortion-free images are directly considered as **level-0**. Further, to ensure a fair comparison, for the same distortion type, we make the number of images in **level-0** to **level-3** equal. Ultimately, we train the ResNet101 [54] on each **level-** and validate the performance with both top-1 and top-5 accuracy (%). The details of implementation can be referred to [54]. The experimental results are given in Table XIV, from which we can see some interesting things.

Firstly, the difference between **level-0** and **level-1** is extremely small (within 0.02%), which indicates that although details and contours are somewhat distorted, the global-concepts are basically not decayed, which does not affect the classification accuracy. Secondly, the decrease in classification accuracy occurs when the global-concept starts to be slightly decayed (see **level-2**). Thirdly, when the global-concepts are observably decayed, there is a significant decrease in classification accuracy (see **level-3**). In summary, the proposed quality caption model can characterize quality demands for needs-oriented cognition.

H. Analysis on Complexity

We now calculate the parameters and floating point operations (FLOPs) to compare the complexity between different DCNN-based BIQA methods and the proposed BDRNet. Table XV shows the parameters and FLOPs of representative DCNN-based BIQA methods, where the test results (Rouge-L) on TID2013 are adopted to illustrate the performances. And we select the size of the image as 224×224 for the complexity analysis. Among the compared methods, the CaHDC [15] has the lowest computational complexity (smallest parameters and FLOPs), but the performance is not desired. For DBCNN [32], it has the highest computational complexity, but unfortunately, its performance does not have superiority. Compared to the Two-stream [33] and WaDIQaM [11], although the number of parameters of our proposed BDRNet is not superior, the FLOPs are the lowest, and at the same time, the performance is the best among all the compared methods.

V. CONCLUSION AND DISCUSSION

In this paper, a novel quality caption model is inventively proposed to measure fine-grained image quality with hierarchical semantics degradation. The existing BIQA methods evaluate image quality with continuous quality scores, which is inconsistent with human cognition custom and can not reflect degradation on hierarchical semantics for needs-oriented cognition. To comply with human cognition, we represent the image quality as patterns of activation distributed across the diverse degradation on hierarchical semantics. A bi-directional hierarchical semantics interaction module is firstly proposed to explore the hierarchy and reverse hierarchy correlations between hierarchical semantics for semantics extraction. Then, a bi-directional recurrent degradation attractor module is designed to learn the degradation dependencies among hierarchical semantics for semantics degradation prediction. Finally, based on these two modules, a novel bi-directional relationship-based network (BDRNet) is proposed for semantics degradation prediction. Experiments demonstrate that our method achieves superior performance and is highly compliant with human cognition.

In the future, there are two aspects that need to continue to be strengthened. For one hand, we need to further improve the model's ability to handle authentic distortion types, which is important for practical applications. In other hand, the generalization ability (cross-database performance) of the model needs to be further enhanced, hoping to achieve the performance of intra-database evaluations, so as to better deal with the unseen distortion types.

REFERENCES

[1] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1969–1977.

[2] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[3] S. Golestaneh and L. J. Karam, "Reduced-reference quality assessment based on the entropy of DWT coefficients of locally weighted gradient magnitudes," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5293–5303, Nov. 2016.

[4] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, Feb. 2012.

[5] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14131–14140.

[6] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

[7] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.

[8] Q. Wu, H. Li, K. N. Ngan, and K. Ma, "Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2078–2089, Sep. 2018.

[9] K. Gu, W. Lin, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "No-reference quality metric of contrast-distorted images based on information maximization," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4559–4565, Dec. 2017.

[10] Y. Fang, K. Ma, Z. Wang, W. Lin, Z. Fang, and G. Zhai, "No-reference quality assessment of contrast-distorted images based on natural scene statistics," *IEEE Signal Process. Lett.*, vol. 22, no. 7, pp. 838–842, Jul. 2015.

[11] 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 Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2016.

[12] X. Liu, J. van de Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1040–1049.

[13] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2017.

[14] J. Ma *et al.*, "Blind image quality assessment with active inference," *IEEE Trans. Image Process.*, vol. 30, pp. 3650–3663, 2021.

[15] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Trans. Image Process.*, vol. 29, pp. 7414–7426, 2020.

[16] G. Kulkarni *et al.*, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[17] C. D. Aitkin, *Discretization Continuous Features by Human Learners*. New Brunswick, NJ, USA: Rutgers State Univ. New Jersey-New Brunswick, 2009.

[18] X. Zhang, W. Lin, and Q. Huang, "Fine-grained image quality assessment: A revisit and further thinking," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jul. 12, 2021, doi: 10.1109/TCSVT.2021.3096528.

[19] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, and W. Gao, "Fine-grained quality assessment for compressed images," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1163–1175, Mar. 2019.

[20] A. Clarke, K. I. Taylor, B. Devereux, B. Randall, and L. K. Tyler, "From perception to conception: how meaningful objects are processed over time," *Cerebral Cortex*, vol. 23, no. 1, pp. 187–197, 2012.

[21] W. Yang, J. Wu, L. Li, W. Dong, and G. Shi, "Image quality caption with attentive and recurrent semantic attractor network," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4501–4509.

[22] S. Hochstein and M. Ahissar, "View from the top: hierarchies and reverse hierarchies in the visual system," *Neuron*, vol. 36, no. 5, pp. 791–804, 2002.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2014.

[24] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[25] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[26] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2012.

[27] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[28] Q. Li, W. Lin, J. Xu, and Y. Fang, "Blind image quality assessment using statistical structural and luminance features," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2457–2469, Dec. 2016.

[29] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 339–343.

[30] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[31] T. Dai, K. Gu, L. Niu, Y.-B. Zhang, W. Lu, and S.-T. Xia, "Referenceless quality metric of multiply-distorted images based on structural degradation," *Neurocomputing*, vol. 290, pp. 185–195, May 2018.

[32] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[33] Q. Yan, D. Gong, and Y. Zhang, "Two-stream convolutional networks for blind image quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2200–2211, May 2019.

[34] K. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 732–741.

[35] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognit.*, vol. 81, pp. 432–442, Sep. 2018.

[36] J. Wu, J. Zeng, Y. Liu, G. Shi, and W. Lin, "Hierarchical feature degradation based blind image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 510–517.

[37] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Oct. 2006.

[38] N. N. Ponomarenko *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *Proc. Eur. Workshop Vis. Inf. Process. (EUVIP)*, Jun. 2013, pp. 106–111.

[39] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, 2010, Art. no. 011006.

[40] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[41] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10k: Towards an ecologically valid and large-scale IQA database," 2018, *arXiv:1803.08489*.

[42] A. Ciancio, A. L. N. T. da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, Jan. 2011.

[43] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document P. ITU-T Recommendation, International Telecommunication Union, 1999.

[44] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[45] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[47] S. Hochreiter and M. C. Mozer, "A discrete probabilistic memory model for discovering dependencies in time," in *Proc. Int. Conf. Artif. Neural Netw.*, Cham, Switzerland: Springer, 2001, pp. 661–668.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[50] X. Chen *et al.*, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[51] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.

[52] *Method for Specifying Accuracy and Cross-Calibration of Video Quality Metrics (vqm)*, document ITU-T Recommendation, I. T. Union, 2004, p. 913.

[53] P. Hanhart, L. Krasula, P. L. Callet, and T. Ebrahimi, "How to benchmark objective quality metrics from paired comparison data?" in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

Wen Yang received the B.S. degree from Xidian University, Xi'an, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Artificial Intelligence. His research interests include image processing, image quality assessment, and event camera-based vision.

Jinjian Wu (Member, IEEE) received the B.Sc. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. From 2011 to 2013, he was a Research Assistant with Nanyang Technological University, Singapore, where he was a Postdoctoral Research Fellow from 2013 to 2014. From 2015 to 2019, he was an Associate Professor with Xidian University, where he has been a Professor since 2019. His research interests include visual perceptual modeling, biomimetic imaging, quality evaluation, and object detection. He received the Best Student Paper Award at the ISCAS 2013. He has served as an Associate Editor for the journal of *Circuits, Systems and Signal Processing* (CSSP), the Special Section Chair for the IEEE Visual Communications and Image Processing (VCIP) 2017, and the Section Chair/Organizer/TPC Member for the ICME 2014–2015, PCM 2015–2016, ICIP 2015, VCIP 2018, and AAAI 2019.

Shiwei Tian received the B.S. degree in electronic information engineering from Xidian University, Xi'an, in 2008, and the M.S. and Ph.D. degrees in communication and information system from the Army Engineering University of PLA, Nanjing, in 2011 and 2015, respectively. Since 2015, he has been an Assistant Professor with the College of Communications Engineering, Army Engineering University, and since 2021, he has been working with the National Innovation Institute of Defense Technology. His main research interests are satellite navigation, cooperative positioning, and machine learning.

Leida Li (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. In 2008, he was a Research Assistant with the Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search (ROSE) Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. He is currently a Professor with the School of Artificial Intelligence, Xidian University. His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics. He has served as a SPC for IJCAI 2019–2021, the Session Chair for ICMR 2019 and PCM 2015, and a TPC for CVPR 2021, ICCV 2021, AAAI 2019–2021, ACM MM 2019–2020, ACM MM-Asia 2019, ACII 2019, and PCM 2016. He is currently an Associate Editor of the *Journal of Visual Communication and Image Representation* and the *EURASIP Journal on Image and Video Processing*.

Weisheng Dong (Member, IEEE) received the B.S. degree in electronic engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2004, and the Ph.D. degree in circuits and system from Xidian University, Xi'an, China, in 2010. He was a Visiting Student with Microsoft Research Asia, Beijing, China, in 2006. From 2009 to 2010, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. In 2010, he joined Xidian University, as a Lecturer, and has been a Professor since 2016. He is currently with the School of Artificial Intelligence, Xidian University. His research interests include inverse problems in image processing, sparse signal representation, and image compression. He was a recipient of the Best Paper Award at the SPIE Visual Communication and Image Processing (VCIP) in 2010. He is currently serving as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and *SIAM Journal on Imaging Sciences*.

Guangming Shi (Fellow, IEEE) received the B.S. degree in automatic control, the M.S. degree in computer control, and the Ph.D. degree in electronic information technology from Xidian University, Xi'an, China, in 1985, 1988, and 2002, respectively. He had studied at the University of Illinois and The University of Hong Kong. Since 2003, he has been a Professor with the School of Electronic Engineering, Xidian University. He awarded the Cheung Kong Scholar Chair Professor by the Ministry of Education in 2012. He is currently the Academic Leader of Circuits and Systems at Xidian University. He has authored or coauthored over 200 papers in journals and conferences. His research interests include compressed sensing, brain cognition theory, multirate filter banks, image denoising, low-bitrate image and video coding, and the implementation of algorithms for intelligent signal processing. He served as the Chair for the 90th MPEG and 50th JPEG of the International Standards Organization (ISO) and the Technical Program Chair for FSKD 2006, VSPC 2009, IEEE PCM 2009, SPIE VCIP 2010, and IEEE ISCAS 2013.