Blind Image Quality Assessment With Active Inference

Jupo Ma, Jinjian Wu, *Member, IEEE*, Leida Li, *Member, IEEE*, Weisheng Dong, *Member, IEEE*, Xuemei Xie, *Senior Member, IEEE*, Guangming Shi, *Fellow, IEEE*, and Weisi Lin, *Fellow, IEEE*

***Abstract*— Blind image quality assessment (BIQA) is a useful but challenging task. It is a promising idea to design BIQA methods by mimicking the working mechanism of human visual system (HVS). The internal generative mechanism (IGM) indicates that the HVS actively infers the primary content (i.e., meaningful information) of an image for better understanding. Inspired by that, this paper presents a novel BIQA metric by mimicking the active inference process of IGM. Firstly, an active inference module based on the generative adversarial network (GAN) is established to predict the primary content, in which the semantic similarity and the structural dissimilarity (i.e., semantic consistency and structural completeness) are both considered during the optimization. Then, the image quality is measured on the basis of its primary content. Generally, the image quality is highly related to three aspects, i.e., the scene information (content-dependency), the distortion type (distortion-dependency), and the content degradation (degradation-dependency). According to the correlation between the distorted image and its primary content, the three aspects are analyzed and calculated respectively with a multi-stream convolutional neural network (CNN) based quality evaluator. As a result, with the help of the primary content obtained from the active inference and the comprehensive quality degradation measurement from the multi-stream CNN, our method achieves competitive performance on five popular IQA databases. Especially in cross-database evaluations, our method achieves significant improvements.**

***Index Terms*— Blind image quality assessment, internal generative mechanism, generative adversarial network, convolutional neural network.**

I. Introduction

OBJECTIVE image quality assessment (IQA) aims to evaluate the perceptual quality of an image automatically. During the process of image acquisition, transmission, compression and storage, various distortions will be introduced, resulting in the deterioration of image quality. Commonly, as the ultimate receiver of images, subjective IQA is deemed as the most accurate and reliable method. However, subjective quality evaluation is laborious, expensive and cannot be embedded into real-time image processing system. Therefore, it's essential to develop objective IQA algorithm for automatic quality prediction.

Based on the availability of reference image, objective IQA can be generally classified into three categories: full-reference (FR) IQA, reduced-referenced (RR) IQA and no-reference (NR) IQA [1]. With the full reference image or partial information for comparison, many excellent FR-IQA or RR-IQA methods have been proposed in the literature. However, in most real-world scenarios, the reference image is not accessible. Thus, NR-IQA, a.k.a. blind IQA (BIQA), attracts growing attention in recent years.

Generally, traditional BIQA methods involve two phases. First, handcrafted descriptors are designed to extract quality-aware features [2]–[6]. Next, a mapping function is designed to map the features into quality values. However, the representation of handcrafted features is limited in modeling the characteristics of diverse distortions and image contents. Recently, due to the strong feature representation capability of convolutional neural network (CNN), some CNN-based BIQA methods have been developed. Compared to traditional BIQA, dramatic improvement has been achieved by CNN-based methods [7], [8]. However, due to the lack of reference information as guidance, it's still a challenging task for existing CNN-based methods to predict the image quality highly consistent with subjective perception.

As a highly developed system, human visual system (HVS) can easily judge the image quality. It's a natural idea to develop BIQA methods by mimicking the working mechanisms of HVS. In fact, what we "see" is not the lateral translation of input stimuli but the response of interactions between the external stimuli and the internal brain mechanisms. Recently, some researches on neural science, such as free energy principle [9], [10] and Bayesian brain hypothesis [11], indicate that HVS works with an internal generative mechanism (IGM) for perception and recognition. For an input scene, IGM tries to avoid the disorderly information and gives the best explanation in a constructive manner [12]–[14]. In other words, the HVS has an active inference process within IGM. For an input image, IGM first analyzes the correlation among pixels. Then combining with the inherent prior knowledge, IGM infers the corresponding primary content actively to better understand the input.

The primary content comprises the primary scene information (i.e., regular structure that represents meaningful information of the input image for better understanding) and will

Manuscript received May 18, 2020; revised January 8, 2021; accepted February 27, 2021. Date of publication March 11, 2021; date of current version March 17, 2021. This work was supported in part by the NSF of China under Grant 62022063, Grant 61772388, and Grant 61632019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao M. Ascenso. *(Corresponding author: Jinjian Wu.)*

Jupo Ma, Jinjian Wu, Leida Li, Weisheng Dong, Xuemei Xie, and Guangming Shi are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: jupoma@stu.xidian.edu.cn; jinjian.wu@mail.xidian.edu.cn; ldli@xidian.edu.cn; wsdong@mail.xidian.edu.cn; xmxie@mail.xidian.edu.cn; gmshi@xidian.edu.cn).

Weisi Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Digital Object Identifier 10.1109/TIP.2021.3064195

be transported to the high level of HVS for interpretation [15]. The prediction error between the input image and its primary content has little effect on the understanding of image, but it reflects the distortion and may cause uncomfortable feelings of HVS. Besides, the interaction between the primary content and the distortion (i.e., the prediction error) will cause the image content degradation. These properties indicate that we can explore the image perceptual quality from different perspectives.

In this work, inspired by IGM, an active inference module is firstly established to emulate the working process of IGM. Here, the active inference refers to the simulating of IGM to predict the primary content. Considering the excellent performance of generative adversarial network (GAN) in inferring image contents and synthesizing realistic images, the GAN framework is adopted in the active inference module. GAN, however, is usually used to generate high-quality or distortion-free images. Different from that, our proposed GAN aims to predict the primary content of a distorted image, which is not the best-effort restoration of the reference image. For example, when a distorted image is destroyed severely, the HVS can't infer the pristine distortion-free content. In other words, IGM will improve the understanding of the input image, but it will not change the underlying visual information. The main semantics of the distorted image should be highly consistent with that of its primary content. Besides, since the prediction error comprises disorderly information, the structure similarity between the prediction error and the primary content should be as small as possible. Therefore, two new IGM-inspired constraints, i.e., semantics similarity constraint and structure dissimilarity constraint, are proposed in the objective function. As a result, the proposed GAN-based active inference module is effective to simulate the IGM theory to predict the primary content for multifaceted quality analysis.

Next, a multi-stream CNN-based quality evaluator is proposed to measure the image quality from multiple aspects. Generally, the psychovisual quality of an image is highly related to three aspects, i.e., the scene information (content-dependency), the distortion type (distortion-dependency), and the content degradation (degradation-dependency). According to the inherent correlation between the distorted image and its primary content, different prior information can be calculated to model the characteristics of the three aspects. For the content-dependency, the related features are extracted from the primary content directly. For the distortion-dependency, the prediction error is used to model its characteristic. Meanwhile, since the HVS is highly sensitive to structure, the structural features are widely used to measure the content degradation. Thus, the structure similarity [16] between the distorted image and its primary content is used to analyze the effect of the degradation-dependency on image quality. Finally, by considering the different prior information as input, a multi-stream quality evaluator is built, which incorporates the characteristics of the content-/distortion-/degradation-dependency together for quality prediction. Experiments on five popular IQA databases verify the effectiveness of our method. Especially in cross database evaluations, thanks to the prior information obtained from the active inference and the multifaceted quality analysis, our method significantly outperforms existing state-of-the-art methods at most cases.

In summary, the main contributions of this paper can be summarized as below:

1) We propose a novel GAN to emulate the active inference process of IGM. Thanks to the two IGM-inspired constraints, i.e., the semantics similarity constraint and the structure dissimilarity constraint, the proposed GAN-based active inference module can effectively predict the primary content of a distorted image for multifaceted quality analysis. To the best of our knowledge, this paper is the first one which adopts GAN to predict the primary content of a distorted image for BIQA.

2) On the basis of the primary content, we design a multi-stream quality evaluator network that can simultaneously measure the effects of the content-/distortion-/degradation-dependency on image quality. Thanks to the multifaceted quality analysis, the proposed quality evaluator can effectively leverage the properties of IGM to predict the image quality. Experiment results on five benchmark IQA databases verify the superiority of our method.

The source code and pretrained model of the proposed method are available at https://web.xidian.edu.cn/wjj/paper.html.

II. Related Work

In this section, we first review some traditional and CNN-based BIQA methods. Then, we give a brief introduction of GAN and some GAN-based BIQA methods.

A. Traditional Blind Image Quality Assessment

Many BIQA methods have been proposed during the past few years. Commonly traditional BIQA methods first extract handcrafted quality-aware features and then a regression function (e.g., SVR) is adopted to map the features into quality score. One of the most common approaches is natural scene statistic (NSS) based methods. For example, DIIVINE [17], BLIINDS-II [18], and BRISQUE [19] respectively extract NSS-related features from the DWT, DCT and spatial domain to predict the perceptual quality. These methods are based on the hypothesis that the natural undistorted image possesses certain statistical properties which are altered in the presence of distortion. Another most common approach is HVS-based methods. Such as NRSL [20], LPSI [21], M3 [22] and [23] are proposed based on the assumption that HVS is adapted to the structural information, in which the features about gradient, luminance contrast or local binary pattern are extracted. In RISE [24], multiscale features are extracted for quality prediction based on the multiscale characteristic of HVS. Inspired by the free energy principle in HVS, NFEQM [12] proposes to use the free energy to predict the image quality on the premise of knowing the distortion type. In NRFRM [14], the NSS-related features, structure-related features and free-energy-related features are combined together to predict the image quality. However, due to the complexities of distortions and image contents, the representation ability of handcrafted features is still limited.

B. CNN-Based Blind Image Quality Assessment

Recently, CNN has achieved great success in various computer vision tasks, such as image classification, object recognition and semantic segmentation. Due to the powerful feature representation ability, many CNN-based BIQA methods have been proposed.

In [25], a shallow CNN network which consists of one convolutional layer, two fully connected layers and an output node is proposed for IQA. WaDIQaM [26] proposes a deep neural network for IQA, which comprises ten convolutional layers and five pooling layers for feature extraction, and two fully connected layers for regression. It takes the image patches as input and the weighted average score of image patches is computed as the final image quality. In BIECON [27], FR-IQA method is utilized to compute a proxy quality score for each image patch to train the network. Usually, large-scale databases are required to train a robust deep network. Existing benchmark IQA databases are hard to meet this requirement.

To cope with the problem of limited size of IQA database, RankIQA [28] proposes to pretrain the network on a large-scale ranked dataset in which the ranked images are automatically generated by applying distortion to reference image. In MEON [29], a large-scale synthetic dataset with known distortion types is collected to pretrain a multi-task network. In DB-CNN [30], both the popular ImageNet dataset and the synthetic collected dataset with known distortion types and levels are adopted. Pretraining on large-scale datasets related to IQA task effectively improves the performance of these CNN-based methods.

To leverage more complicated features, BLINDER [31] proposes to extract hierarchical features from multilevel of deep CNN model for BIQA, rather than just using features from the last convolutional layer. Considering different subjects may have divergent subjective perception for an image, a scalar quality score may not be adequate to represent the divergence. PQR [32] and DeepRN [33] propose to use the distribution of scores to represent the image quality. To learn more effective feature representations for BIQA, the gradient map is also fed into the network in Two-stream [34] to capture different level information and ease the difficulty of extracting features from only the input distorted image. However, due to the lack of reference information as guidance, it's difficult to train a robust CNN model for BIQA. In NAR-DCNN [35], the authors show that non-aligned image with similar scene could be well used for reference. But the measurement for selecting images with similar scene still needs to be explored.

C. Generative Adversarial Network

Generally, GAN [36] consists of two subnetworks: a generator G to generate samples and a discriminator D to distinguish the real samples from the generated samples. The training of G and D is a minmax game with objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}}\left[\log\left(D\left(x\right)\right)\right] + \mathbb{E}_{z\sim P_z}\left[\log\left(1-D\left(G\left(z\right)\right)\right)\right] \quad (1)$$

where P_{data} is the real data distribution, P_z is the distribution of the input noise. D is optimized to assign the correct label for both the real samples and the generated samples. G is optimized to minimize $\log\left(1-D\left(G\left(z\right)\right)\right)$. Through the adversarial training, the generator is expected to generate more 'realistic' samples to fool the discriminator. Various GAN models have been developed and achieved great success in many image generation tasks, such as image synthesis, image super-resolution, image style transfer and image enhancement, which demonstrate the standout performance of GAN in generating realistic semantics and high-quality details.

However, It's known that the training of GAN suffers from instability, such as vanishing gradients, mode collapse etc [37]. To ease the instability of the adversarial training process, WGAN [38] proposes to use the Earth-Mover (EM) distance instead of the Jensen-Shannon divergence to measure how close the real data distribution and the generated data distribution is. Under mild assumption, the EM distance is continuous and differentiable almost everywhere. Unlike traditional 0-1 classification, the discriminator D in WGAN solves a regression problem. Weight clipping is additionally used on D to enforce the Lipschitz constraint. However, weight clipping may lead to undesired behaviors, such as generating only poor samples or failing to converge. Therefore, instead of clipping weights, WGAN-GP [39] proposes to penalize the norm of gradient of D's output with respect to its input directly to enforce the Lipschitz constraint. In our work, the objective function of WGAN-GP is adopted to stabilize the adversarial training.

Some GAN-based IQA methods [40]–[45] have also been developed in the last few years. In GADA [40], the GAN is adopted to generated distorted images to augment the size of the training dataset. To address the absence of reference image, H-IQA [41] proposes to use a GAN model to restore a hallucinated reference from the distorted image. Because it's hard to design a universal GAN to restore the high-quality reference image for all distortion types, H-IQA introduces a novel modified discriminator to constrain the influence of bad restoration. Then the distorted image and the discrepancy map (i.e., the error between the distorted image and its hallucinated reference) are forward into a regression network to predict the image quality. Motivated by the free energy principle, RAN4IQA [45] supposes that HVS will unveil the mask of distortion and add details to figure out the pristine content. So a GAN model is established to restore the distorted image. Then an evaluator is built to measure the perceptual discrepancy between the distorted image and its restored counterpart. Our approach is different from them in the following ways. (1) First, our proposed IGM-guided GAN is not intended to restore the distorted image to a quality-perfect or distortion-free image, but to predict the primary content which may still contains degraded information. Because when an image is distorted severely, IGM couldn't infer the pristine or distortion-free content effectively. (2) In our method, by utilizing the output of the GAN-based active inference module, different prior information can be obtained to measure the image quality from multiple aspects simultaneously, rather than just measuring the discrepancy between the distorted image and its restored counterpart. By integrating the multiple

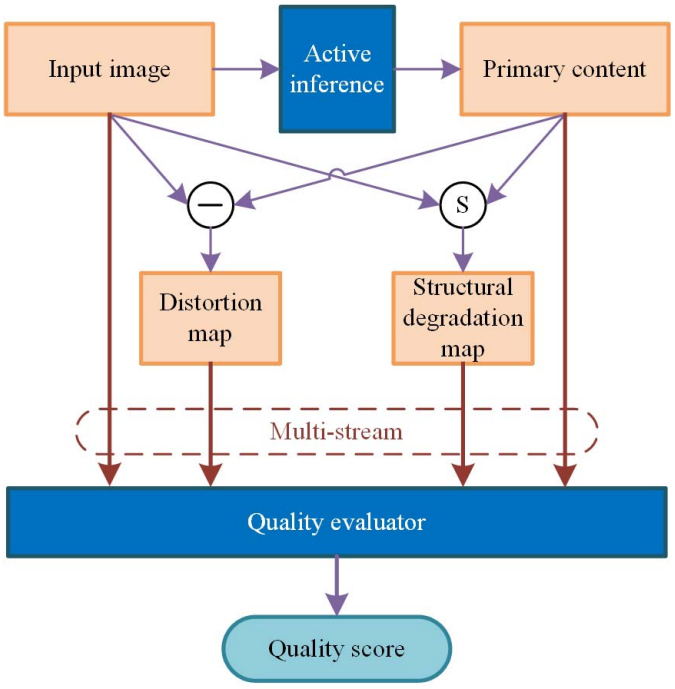


Fig. 1. Flowchart of the proposed AIGQA. AIGQA is mainly composed of two parts: the GAN-based active inference module and the multi-stream CNN-based quality evaluator.

information as input, the proposed quality evaluator could better leverage the properties of IGM for BIQA.

III. The Proposed Method

By mimicking the active inference of IGM with GAN and measuring the image quality from multiple aspects with CNN, a novel BIQA model (named as AIGQA) is built in this work. As shown in Fig.1, AIGQA mainly consists of two parts: the GAN-based active inference module and the multi-stream CNN-based quality evaluator. The active inference module is designed to emulate the active inference process of IGM to predict the primary content. The quality evaluator aims to integrate multi-stream prior information together to predict the perceptual quality. Following we will discuss the active inference module and the quality evaluator in detail.

A. Active Inference Module

Inspired by IGM, an active inference module is firstly proposed to predict the primary content of a distorted image. During the past few years, GAN has shown prefect performance in image generation tasks. It can effectively understand the representation of image data and synthesize realistic samples. On the other hand, IGM can be viewed as a process of "analysis by synthesis" [12]. So the GAN framework is adopted to construct the active inference module, which comprises two components: the generator G and the discriminator D. G takes the distorted image I_d as input and aims to predict its primary content I_g, i.e., $I_g = G(I_d)$. D aims to discriminate the real primary content I_r which is inferred by IGM from the predicted version I_g. Through the adversarial training between G and D, the predicted primary content I_g is expected to be indistinguishable from the real primary content I_r.

By adopting the WGAN-GP [39] framework, the objective function of D is defined as:

$$D^* = \arg\min\left(-L_{adv} + L_{GP}\right) \tag{2}$$

where L_{adv} is the adversarial loss, L_{GP} is the gradient penalty term. L_{adv} is formulated as:

$$L_{adv} = \mathbb{E}_{I_r \sim P_r}\left[D\left(I_r\right)\right] - \mathbb{E}_{I_g \sim P_g}\left[D\left(I_g\right)\right] \tag{3}$$

where P_r is the distribution of the real primary content, P_g is the distribution of the predicted primary content generated by G. L_{GP} is formulated as:

$$L_{GP} = \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{x}}}\left[\left(\left\|\nabla_{\hat{x}} D\left(\hat{x}\right)\right\|_2 - 1\right)^2\right] \tag{4}$$

where $P_{\hat{x}}$ represents the sampling distribution which samples uniformly along straight lines between P_r and P_g. $\left\|\nabla_{\hat{x}} D\left(\hat{x}\right)\right\|_2$ is the gradient norm of D's output with respect to its input. λ is the penalty coefficient, and as in WGAN-GP [39], $\lambda = 10$ in our experiments. Under the constraint of the adversarial loss L_{adv}, the generator G is optimized to generate realistic images.

Except for the adversarial loss, in order to optimize G to generate more realistic samples, the pixel loss L_{pix} and the content loss $L_{content}$ [46] are also added into the loss function of G. L_{pix} calculates the discrepancy between the predicted primary content I_g and the real primary content I_r at the pixel space, which is formulated as:

$$L_{pix} = MSE\left(I_g, I_r\right) \tag{5}$$

where $MSE(\cdot)$ calculates the mean square error between the two inputs. The content loss $L_{content}$, a.k.a. perceptual loss, is defined as the discrepancy between I_g and I_r at the feature space:

$$L_{content} = MSE\left(\phi_k\left(I_g\right) - \phi_k\left(I_r\right)\right) \tag{6}$$

where $\phi_k(\cdot)$ is the feature maps at the k-th convolution layer of a pretrained network. In this work, the feature maps of $VGG_{3,3}$ from VGG19 network pretrained on ImageNet are defined as the feature space to calculate the content loss. The pixel loss L_{pix} ensures that G could capture correct low frequencies. The content loss $L_{content}$ will enhance the ability of G to learn perceptual representations. In addition, note that since we have no access to get the real primary content inferred by IGM, the reference image of a distorted image is adopted as the substitute.

The adversarial loss, pixel loss and content loss are widely used to generate high-quality or distortion-free images in recent GAN models, such as SRGAN [47], DeblurGAN [46]. However, it's hard to train a universal GAN to restore the reference image for all distortion types. Different from that, our proposed GAN aims to mimic the active inference process of IGM to predict the primary content, rather than restore a pristine or distortion-free image. Thus, as shown in Fig.2, two IGM-inspired constraints are proposed in the objective function of G to make the predicted primary content more consistent with IGM.

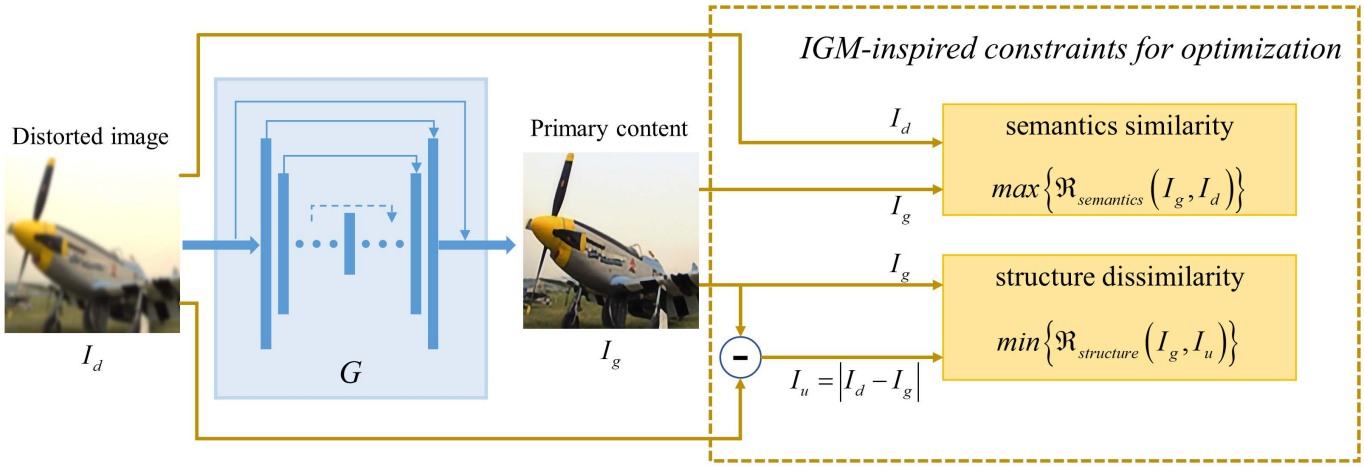


Fig. 2. The proposed generator G with two IGM-inspired constraints during the optimization of GAN. $\Re_{semantics}(I_g, I_d)$ refers to the semantics similarity defined in Eq.7, which should be as large as possible. $\Re_{structure}(I_g, I_u)$ refers to the structure similarity defined in Eq.11, which should be as small as possible. Taking the distorted image as input, G aims to predict the primary content.

1) Semantics Similarity Constraint: IGM aims to infer the primary content to help human brain better understand the input image, but it will not change the underlying visual information of the input. For example, for a distorted image which is severely destroyed by Gaussian blur, IGM couldn't infer the pristine content. As show in Fig.6(b), IGM can't infer the text message as clear as in Fig.6(a). The main semantics of the input image I_d and its primary content I_g should be highly consistent. Thus, in order to maintain consistency in understanding and interpreting images, the semantics similarity constraint proposed here is defined by maximizing the following formula:

$$\Re_{semantics}(I_g, I_d) = -MSE\left(\phi_k(I_g) - \phi_k(I_d)\right) \tag{7}$$

where $\phi_k(\cdot)$ is the same as in Eq.6, because CNN naturally learns hierarchical semantic features with the depth of layers from shallow to deep.

2) Structure Dissimilarity Constraint: To give the best explanation of an input image, IGM also tries to avoid the disorderly information which is represented by the prediction error I_u between the input distorted image I_d and its primary content I_g,

$$I_u = |I_d - I_g| \tag{8}$$

The primary content I_g contains the main structure information about the input scene. While the prediction error I_u comprises disorderly information. Thus the structure similarity between the primary content I_g and the prediction error I_u should be as small as possible. In this work, we utilize the classical SSIM [16] to calculate structure similarity, which is defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{9}$$

where x and y are two signals to compare, μ_x and μ_y are the mean intensity, σ_x and σ_y are the standard deviation, the constant C_1 and C_2 are included to avoid instability when denominator is very small. By applying Eq.9 pixel-by-pixel over the entire image within a local 8×8 square window, the structure similarity map I_s between the primary content I_g and the prediction error I_u is obtained. For convenience, the process to calculate structure similarity map is formulated as:

$$I_s = SSIM(I_g, I_u) \tag{10}$$

Note that according to the definition in SSIM [16], I_g and I_u are first converted to grayscale image to calculate I_s. As a result, in order to maintain the structure completeness of the primary content, the structure dissimilarity constraint is defined by minimizing the following formula:

$$\Re_{structure}(I_g, I_u) = \frac{1}{WH}\left\|SSIM(I_g, I_u)\right\|_2^2 \tag{11}$$

where W and H represent the width and height of I_g.

Finally, the objective function of G is formulated as

$$G^* = \arg\min\left(\mu_1 L_{adv} + \mu_2 L_{pix} + \mu_3 L_{content} + \mu_4 L_{ss} + \mu_5 L_{sd}\right) \tag{12}$$

where $L_{ss} = -\Re_{semantics}(I_g, I_d)$, $L_{sd} = \Re_{structure}(I_g, I_u)$. In our experiments, we set $\mu_2 = 1.0$, $\mu_3 = 0.01$, $\mu_4 = 0.01$ and $\mu_5 = 1.0$ to balance the scale of each loss. For μ_1, we set $\mu_1 = \mu_2$ for simplicity. Benefiting from the two IGM-inspired constraints, the proposed GAN is endowed with properties of IGM to predict the primary content.

B. Quality Evaluator

On the basis of the primary content, a multi-stream CNN-based quality evaluator which can measure image quality from multiple aspects is proposed. Existing CNN-based methods commonly only take the distorted image as input, which makes it difficult to learn effective features for multifaceted quality analysis. As discussed above, the psychovisual quality of image is highly related to three aspects (i.e., the content-dependency, the distortion-dependency and the degradation-dependency) and different prior information can be calculated to model the characteristics of the three aspects. Specially, the primary content I_g is used to analyze the effect of the content-dependency on image quality. The characteristic of

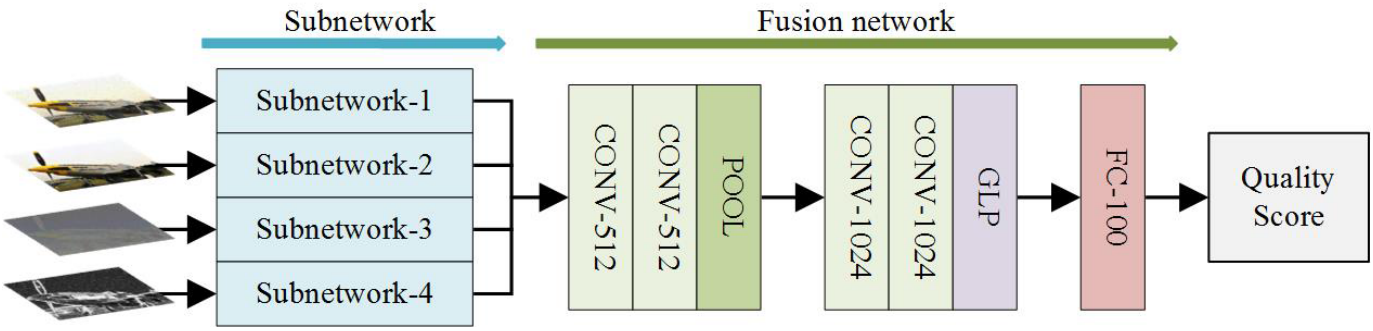


Fig. 3. Configurations of the multi-stream quality evaluator. The subnetwork-1 to subnetwork-4 respectively take the distorted image, the primary content, the distortion map and the structural degradation map as input. Each subnetwork is configured as a series of stacked layers which include {CONV-64, CONV-64, POOL, CONV-128, CONV-128, POOL, CONV-256, CONV-256, POOL, CONV-512, CONV-512, POOL}. CONV-m denotes convolution layer with 3×3 kernel, 1×1 stride and m output channels. POOL denotes maxpooling layer with 2×2 kernel and 2×2 stride. GLP denotes the global maxpooling layer.

the distortion-dependency is represented by the distortion map I_{dm} which is defined as the prediction error, i.e., $I_{dm} = I_u$. Furthermore, by applying SSIM [16] again, the effect of the degradation-dependency is measured from the structural degradation map I_{sm}, which is defined as the structure similarity map between the distorted image I_d and its primary content I_g.

Thus, by incorporating the characteristics of the three aspects, a multi-stream quality evaluator is built as show in Fig.3. The primary content I_g, the distortion map I_{dm} and the structural degradation map I_{sm} are fed into the subnetwork-2, subnetwork-3, subnetwork-4 respectively to extract features from different aspects. Besides, the distorted image I_d is also fed into the subnetwork-1 to extract information about the original input scene. Then the features from the four subnetworks are concatenated together and transported into the fusion network to predict the quality score. The whole process is formulated as:

$$\begin{aligned} q &= Q\left(I_d, I_g, I_{dm}, I_{sm}\right) \\ &= F\left(sub_1\left(I_d\right), sub_2\left(I_g\right), sub_3\left(I_{dm}\right), sub_4\left(I_{sm}\right)\right) \end{aligned} \quad (13)$$

where Q, F, sub_i denote the processes of the whole quality evaluator, the fusion network and the subnetwork-i respectively. Note that I_{sm} is a grayscale image, so the input channel of the first convolution layer of subnetwork-4 is set to 1. By taking the subjective quality score as target, MSE loss is employed to train Q. Through end-to-end optimization, the quality evaluator can effectively analyze the image quality from different aspects and predict the image quality highly consistent with subjective perception.

IV. Experimental Results

In this section, we first describe the experimental setups, including datasets, evaluation criteria, and network architecture details. Then we compare the performance of AIGQA with other BIQA methods. We next conduct a series of ablation studies to identify the contribution of the key components of AIGQA. Finally, we also present some visualization samples obtained from the active inference module.

A. Experimental Setups

1) Training and Datasets: Both the active inference module and the quality evaluator adopt the Adam optimization algorithm for training. The training process is divided into two steps: pretraining on the collected synthetic images and finetuning on the standard IQA databases.

At the pretraining step, the collected synthetic images derive from the Waterloo Exploration Database (WED) [56]. WED contains 94 880 distorted images, which are generated from 4744 high-quality pristine images with 4 distortion types at 5 levels, i.e., white Gaussian noise, Gaussian blur, JPEG compression and JPEG2000 compression. As in [28]–[30], we additionally add 13 more distortion types which come from TID2013 [57] (i.e., #2, #5, #6, #7, #9, #14, #15, #16, #17, #18, #19, #22, #23) to the pristine images. Since FR-IQA has achieved high consistency with subjective perception. Thus, similar to [27], the state-of-the-art FR-IQA method, VSI, is utilized to label a quality score for each distorted image. As a result, a large scale training set is collected. Both the active inference module and the quality evaluator are pretrained on the collected training dataset. More specifically, during the pretraining process, we first train the GAN-based active inference module. Then we freeze the weights of the GAN, and pretrain the quality evaluator.

At the finetuning step, only the quality evaluator is finetuned on the standard IQA databases, including LIVE [58], CSIQ [59], TID2013 [57], LIVE-MD [60], LIVE-CH [61] and KADID-10K [62]. The LIVE consists 779 distorted images generated form 29 reference images by adding 5 distortion types. The CSIQ contains 886 distorted images created from 30 reference images, 6 distortion types. The TID2013 possesses 3000 distorted images generated from 25 reference images, 24 distortion types, and five levels for each distortion type. The LIVE-MD focuses on multiply distorted images, which consists 450 multiple distorted images created form 15 reference images. The LIVE-CH focuses on authentic distortion, which contains 1162 images captured by a large amount of camera devices in read world. And there is no reference image in LIVE-CH. The KADID-10k is a recently published large-scale synthetic database which contains 81 pristine images, each degraded by 25 distortion types in 5 levels. The subjective quality values (i.e., MOS or DMOS) are available in all the six standard IQA databases.

2) Evaluation Criteria: The Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Order Correlation Coefficient (SROCC) are adopted to measure the performance. PLCC is to measure the linear correlation between

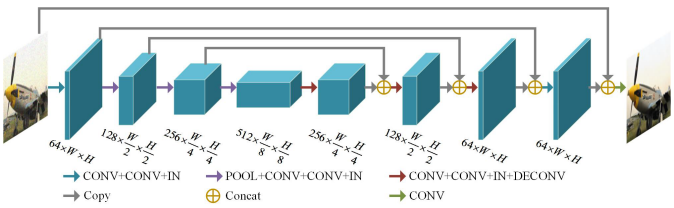


Fig. 4. Architecture details of the generator G. CONV denotes the convolution layer with 3×3 kernel, 1×1 stride. POOL denotes the maxpooling layer with 2×2 kernel, 2×2 stride. DECONV denotes the deconvolution layer with 3×3 kernel, 2×2 stride. IN denotes the Instance Normalization layer. The dimensions of the feature maps at each layer are listed below, formatted as $channel \times width \times height$. W and H are the width and height of the input image.

the predicted score and the ground truth, which is formulated as:

$$PLCC = \frac{\sum_i (p_i - p_m)(\hat{p}_i - \hat{p}_m)}{\sqrt{\sum_i (p_i - p_m)^2}\sqrt{\sum_i (\hat{p}_i - \hat{p}_m)^2}} \tag{14}$$

where p_i and $\hat{p}_i$ are the predicted score and the subjective quality score, p_m and $\hat{p}_m$ are the average of each. SROCC is to measure the monotonicity between the predicted score and the ground truth, which is defined as:

$$SROCC = 1 - \frac{6\sum_{i=1}^{L}(m_i - n_i)^2}{L(L^2-1)} \tag{15}$$

where L is the number of images, m_i is the rank of p_i in the predicted scores, n_i is the rank of $\hat{p}_i$ in the subjective quality values. For both criteria, a higher value indicates higher performance of the algorithm.

3) Network Architecture Details: In our experiments, the architecture of the discriminator D is identical to PatchGAN [63]. It takes $N \times N$ cropped patches as input and discriminates whether each patch is real or fake, i.e., the real primary content or the predicted primary content. The discriminator D is operated across the full input image convolutionally, and the average response of all patches is set as the ultimate output of D. N can be much smaller than the full size of the image, and we set $N = 70$ as in [46], [63]. By restricting attention at the scale of local patches, high-frequencies can be better modeled. Besides, such a patch-level discriminator has fewer parameters and can be applied to arbitrary large images.

In the generator G, U-shaped network is adopted. As shown in Fig.4, there are mainly two reverse phases in G. The first is downsampling process, in which the number of feature channels are increased and the spatial size of feature maps are downsampled progressively through a series of stacked layers. The second phase is the upsampling process, in which the feature channels are decreased and the spatial size are upsampled progressively. Through the skip connection [63], feature maps at mirrored layers are concatenated together to share hierarchical representations. Besides, similar to [46], [64], the input image is concatenated with the last feature maps of the U-shape net directly to provide more original information about the input scene. In the generator, LeakyReLU is used as the activation function for all deconvolutional layers and convolutional layers except for the last one which adopts Tanh instead.

B. Performance Comparison Within Individual Databases

In this section, experiments within individual standard IQA databases are conducted to validate the effectiveness of AIGQA. Following the experimental protocol in [30], [41], [65], one database is randomly divided into 80% for training and 20% for testing. To ensure that there is no overlapping image content between the training set and the testing set, the database is divided according to the reference image. For LIVE-CH, there is no reference image, so we divide the database straight forward by distorted images. All the experiments are repeated 100 sessions, and the median SROCC and PLCC are reported.

We first compare the proposed AIGQA with 8 traditional BIQA methods, 9 CNN-based methods and 1 GAN-based method on 5 popular benchmark IQA databases. The results are listed in Tab.I, the best two SROCC and PLCC are highlighted in bold. For the 8 traditional BIQA methods, all the results are reproduced by the source codes released by their authors. For the 9 CNN-based and 1 GAN-based methods, the results come from the original papers. The detailed comparison is as follows:

1) Compared with the 8 traditional BIQA methods, AIGQA achieves the best performance on all databases.
2) When compared with the 9 CNN-based methods, AIGQA also achieves competitive results on 3 databases, i.e., CSIQ, TID2013 and LIVE-MD. On LIVE, AIGQA gets a slightly lower performance, but it still achieves acceptable results, about 0.96 SROCC and 0.957 PLCC. As for LIVE-CH, AIGQA achieves the second best results. However, due to the huge difference between synthetic distortion and authentic distortion, the performance of AIGQA is relatively poor when compared to DB-CNN [30] which adopts the ImageNet database for pretraining. This also naturally motivates us a promising future direction for AIGQA, i.e., AIGQA could leverage the ImageNet to improve the performance on authentic distortion. Although AIGQA can't achieve the best performance on all databases in Tab.I, AIGQA shows higher robustness and makes significant improvement in cross database evaluations which can be seen in Section IV-D.
3) When compared to the GAN-based method, H-IQA [41] achieves better performance on LIVE, about 2.3% higher on SROCC and 2.6% higher on PLCC. AIGQA achieves better on CSIQ, about 4.7% higher on SROCC and 4.6% higher on PLCC. On TID2013, H-IQA is better at SORCC (about 0.9% higher), while AIGQA is better at PLCC (about 1.5% higher).

Next, we also evaluate the proposed AIGQA on KADID-10k which is the largest synthetic IQA database published recently. Tab.III lists the SROCC and PLCC results. All the compared results are taken from [66]. We can see

TABLE I
PERFORMANCE COMPARISON ON FIVE BENCHMARK IQA DATABASES

Methods	LIVE		CSIQ		TID2013		LIVE-MD		LIVE-CH	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
BLIINDS-II [18]	0.919	0.920	0.570	0.534	0.536	0.628	0.827	0.845	0.405	0.450
DIIVINE [17]	0.925	0.923	0.784	0.836	0.654	0.549	0.874	0.894	0.546	0.568
BRISQUE [19]	0.939	0.942	0.750	0.829	0.573	0.651	0.897	0.921	0.607	0.585
NIQE [48]	0.915	0.919	0.630	0.718	0.299	0.415	0.745	0.815	0.430	0.480
CORNIA [49]	0.942	0.943	0.714	0.781	0.549	0.613	0.900	0.915	0.618	0.662
HOSA [50]	0.948	0.949	0.781	0.842	0.688	0.764	0.902	0.926	0.660	0.680
ILNIQE [51]	0.902	0.865	0.807	0.808	0.519	0.640	0.878	0.892	0.430	0.510
FRIQUEE [52]	0.948	0.962	0.839	0.863	0.669	0.704	0.925	0.940	0.720	0.720
MEON [29]	-	-	-	-	0.808	-	-	-	-	-
DIQaM [26]	0.960	0.972	-	-	0.835	0.855	-	-	0.606	0.601
RANK [28]	**0.981**	**0.982**	-	-	0.780	0.799	0.921	0.936	-	-
VIDGIQA [53]	0.969	0.973	-	-	-	-	-	-	0.701	-
BIECON [27]	0.958	0.960	0.815	0.823	0.717	0.762	0.909	0.933	0.595	0.613
DIQA [54]	0.975	0.977	0.884	0.915	0.825	0.850	**0.939**	**0.942**	0.703	0.704
BPSQM [55]	0.973	0.963	0.874	0.915	0.862	**0.885**	-	-	-	-
Two-stream [34]	0.969	0.978	-	-	-	-	-	-	-	-
DB-CNN [30]	0.968	0.971	**0.946**	**0.959**	0.816	0.865	0.927	0.934	**0.851**	**0.869**
H-IQA [41]	**0.982**	**0.982**	0.885	0.910	**0.879**	0.880	-	-	-	-
AIGQA	0.960	0.957	**0.927**	**0.952**	**0.871**	**0.893**	**0.933**	**0.947**	**0.751**	**0.761**

TABLE II
PERFORMANCE COMPARISON (SROCC) ON INDIVIDUAL DISTORTIONS OF TID2013. THE NUMBER OF TIMES (N.O.T) EACH METHOD ACHIEVES THE BEST PERFORMANCE IS LISTED ON THE LAST ROW

Type	BLIINDS-II [18]	DIIVINE [17]	BRISQUE [19]	M3 [22]	MEON [29]	DB-CNN [30]	DIQA [54]	H-IQA [41]	AIGQA
#1	0.714	0.756	0.711	0.766	0.813	0.790	0.915	0.923	**0.932**
#2	0.728	0.464	0.432	0.560	0.722	0.700	0.755	0.880	**0.916**
#3	0.825	0.869	0.746	0.782	0.926	0.826	0.878	**0.945**	0.944
#4	0.358	0.374	0.252	0.577	0.728	0.646	**0.734**	0.673	0.662
#5	0.852	0.794	0.842	0.900	0.911	0.879	0.939	**0.955**	0.953
#6	0.664	0.704	0.765	0.738	0.901	0.708	0.843	0.810	**0.911**
#7	0.780	0.650	0.662	0.832	0.888	0.825	0.858	0.855	**0.908**
#8	0.852	0.900	0.871	0.896	0.887	0.859	**0.920**	0.832	0.917
#9	0.754	0.814	0.612	0.709	0.797	0.865	0.788	**0.957**	0.914
#10	0.808	0.795	0.764	0.844	0.850	0.894	0.892	0.914	**0.945**
#11	0.862	0.804	0.745	0.855	0.891	0.916	0.912	0.624	**0.932**
#12	0.251	0.514	0.301	0.375	0.746	0.772	**0.861**	0.460	0.858
#13	0.755	0.892	0.748	0.718	0.716	0.773	0.812	0.782	**0.898**
#14	0.081	0.215	0.269	0.173	0.116	0.270	0.659	**0.664**	0.130
#15	0.371	0.389	0.207	0.379	0.500	0.444	0.407	0.122	**0.723**
#16	0.159	0.124	0.219	0.119	0.177	-0.009	0.299	0.182	**0.554**
#17	-0.082	0.189	-0.001	0.155	0.252	0.548	0.687	0.376	**0.830**
#18	0.109	0.280	0.003	-0.199	0.684	0.631	-0.151	0.156	**0.689**
#19	0.699	0.691	0.717	0.738	0.849	0.711	0.904	0.850	**0.948**
#20	0.222	0.340	0.196	0.353	0.406	0.752	0.655	0.614	**0.886**
#21	0.451	0.690	0.609	0.692	0.772	0.860	**0.930**	0.852	0.897
#22	0.815	0.769	0.831	0.908	0.857	0.833	**0.936**	0.911	0.908
#23	0.568	0.700	0.615	0.570	0.779	0.732	0.756	0.381	**0.889**
#24	0.856	0.795	0.807	0.893	0.855	0.902	**0.909**	0.616	0.908
N.o.T	0	0	0	0	0	0	6	4	**14**

that AIGQA still achieves the best performance in terms of both SROCC and PLCC. In general, AIGQA works well on all databases, which verifies the effectiveness of the proposed method.

C. Performance Comparison on Individual Distortions

Performance on individual distortions is compared in this subsection to investigate the stability of AIGQA. Tab.II lists the SROCC on individual distortions in TID2013 and the best results are highlighted. AIGQA achieves the best performance on 14 out of 24 distortion types. Especially on the #15 (local block-wise distortions) and the #17 (contrast change) distortion types, most previous methods failed to predict the image quality consistent with human perception. While in AIGQA, different prior information (i.e., the primary content, the distortion map, the structural degradation map) is obtained to measure the image quality from multiple aspects. Benefiting from the predicted primary content and the multifaceted quality analysis, AIGQA achieves significant improvements on them. For the #3, #5, #8, #12, #24 types, AIGQA achieves

TABLE III
PERFORMANCE COMPARISON ON KADID-10K

Methods	SROCC	PLCC
BLIINDS-II [18]	0.530	0.548
DIIVINE [17]	0.428	0.423
BRISQUE [19]	0.386	0.383
NIQE [48]	0.309	0.273
ILNIQE [51]	0.211	0.230
SCORER [67]	0.856	0.855
MultiGAP-GPR [68]	0.814	0.820
AIGQA	**0.864**	**0.863**

TABLE IV
PERFORMANCE COMPARISON (SROCC) WHEN TRAINING ON LIVE AND TESTING ON THE FULL SET OF CSIQ, TID2013 AND LIVE-MD

Methods	CSIQ	TID2013	LIVE-MD
BLIINDS-II [18]	0.654	0.405	0.456
DIIVINE [17]	0.553	0.487	0.662
BRISQUE [19]	0.549	0.466	0.550
HOSA [50]	0.631	0.465	0.616
FRIQUEE [52]	0.688	0.468	0.502
WaDIQaM [26]	0.704	0.462	-
VIDGIQA [53]	0.641	0.415	-
DB-CNN [30]	0.758	0.524	-
Two-stream [34]	0.614	0.461	-
AIGQA	**0.847**	**0.698**	**0.833**

TABLE V
PERFORMANCE COMPARISON (SROCC) WHEN TRAINING ON TID2013 AND TESTING ON THE FULL SET OF LIVE, CSIQ AND LIVE-MD

Methods	LIVE	CSIQ	LIVE-MD
BLIINDS-II [18]	0.836	0.568	0.509
DIIVINE [17]	0.687	0.590	0.479
BRISQUE [19]	0.681	0.491	0.314
HOSA [50]	0.842	0.622	0.469
FRIQUEE [52]	0.847	0.637	0.421
WaDIQaM [26]	-	0.733	-
DB-CNN [30]	**0.891**	0.807	-
AIGQA	0.886	**0.823**	**0.799**

almost the same performance as the best method. On the rest 5 types, AIGQA also achieves competitive performance except for the #14 (non eccentricity pattern noise). It's probably because HVS is insensitive to the #14 distortion. Thus, most BIQA methods fails to model the characteristics of the #14 distortion type.

D. Cross Database Evaluations

In this section, we compare the generalization ability of AIGQA in cross database evaluations.

In Tab. IV and Tab.V, 5 traditional BIQA methods and 4 CNN-based methods are adopted for comparison. Specifically, the results of the 5 traditional BIQA methods are reproduced by the source codes released by their authors. The results of the 4 CNN-based methods are from their original papers.

Tab.IV lists the SROCC results when training on LIVE and testing on the full set of CSIQ, TID2013 and LIVE-MD. In CSIQ, the distortion type is similar with that in LIVE. Most methods could achieve good performance. For instance, the state-of-the-art BIQA method DB-CNN [30] achieves 0.758 SROCC. Furthermore, AIGQA obtains a better 0.847 SROCC, about 12% improvement compared to DB-CNN. For TID2013, it has more distortion types than LIVE. It's a challenging task to test on it. AIGQA obtains obvious advantage and gets 0.698 SROCC, about 33% improvement compared to DB-CNN. LIVE-MD focuses on complex multiply distortion, AIGQA still achieves 0.833 SROCC on it which is relatively consistent with subjective perception.

TABLE VI
SROCC RESULTS OF THE CROSS-DATABASE EVALUATIONS. ALL METHODS ARE CONDUCTED USING IDENTICAL DATASETS AND TEST PROTOCOL AS THE PROPOSED AIGQA

Methods	LIVE	CSIQ	LIVE-MD	KADID-10K
WaDIQaM [26]	0.861	0.731	0.655	0.446
MEON [29]	0.861	0.735	0.374	0.476
Two-stream [34]	0.853	0.784	0.181	0.461
RAN4IQA [45]	0.871	0.808	0.748	0.507
AIGQA	**0.886**	**0.823**	**0.799**	**0.567**

TABLE VII
STATICAL SIGNIFICANCE TEST RESULTS

	WaDI-QaM	MEON	Two-stream	RAN4-IQA	AIGQA
WaDIQaM	0	0	-1	-1	-1
MEON	0	0	-1	-1	-1
Two-stream	1	1	0	-1	-1
RAN4IQA	1	1	1	0	-1
AIGQA	1	1	1	1	0

In Tab.V, we list the results when training on TID2013 and testing on the other databases. On LIVE, the proposed AIGQA achieves the second best performance, almost the same to the best method DB-CNN. Expect for LIVE, AIGQA achieves the highest SROCC on the other two databases. Especially on LIVE-MD, AIGQA makes apparent improvement.

To give a fairer comparison with existing deep-learning-based methods, we reproduce the experiments of some BIQA methods using the identical datasets and test protocol as our method. Tab.VI list the SROCC when training on TID2013 and testing on the other databases. WaDIQaM [26], MEON [29] and Two-stream [34] are reimplemented according to the source codes released by their authors. RAN4IQA [45] is reimplemented by our own version. From Tab.VI we can see that AIGQA achieves the best performance on all the cases. Besides, the statistical significant test is also conducted to examine the significance of the performances between each two methods. The SROCC values of each method when training on TID2013 and testing on CSIQ are used as input for t-test. The t-test results are listed in Tab.VII. '1/-1/0' respectively represents the model in row is statistically better than/worse than/indistinguishable with the model in column with 95% confidence level. From this table, we can see AIGQA is statistically better than the other 4 deep-learning-based methods.

TABLE VIII
ABLATION EXPERIMENTS ABOUT THE MULTIFACETED QUALITY ANALYSIS

	SROCC	PLCC
BL	0.627	0.675
BL+CD	0.641	0.697
BL+CD+DD	0.673	0.722
BL+DD+SD	0.644	0.693
AIGQA (BL+CD+DD+SD)	**0.698**	**0.728**

In conclusion, a series of cross-databases evaluations show the superior generalization ability of the proposed IGM-inspired method. By simulating the active inference process of IGM and integrating the multiple information together, AIGQA can better leverage the IGM theory to measure the image quality. When dealing with new distortion types or new image contents or more complex distortion, AIGQA achieves better generalization ability.

E. Ablation Experiments

In this subsection, we conduct a series of ablation experiments to identify the contribution of key components of the proposed method. In the ablation experiments, all models are trained on LIVE and tested on the full set of TID2013. Except for the specified statement, all experimental setups are the same as above.

We firstly analyze the gains of multifaceted quality analysis by changing input to the quality evaluator. Experimental results are listed in Tab.VIII. The modified model which only takes the distorted image as input of the quality evaluator (i.e., the subnetwork-2, subnetwork-3 and subnetwork-4 are removed) is set to be the baseline network (**BL**), in which the effects of different aspects on image quality are not explicitly analyzed. Model **BL+CD** takes the distorted image and the primary content as input to measure the effect of the content-dependency explicitly, which achieves about 2.2% improvement on SROCC and 3.6% improvement on PLCC compared to **BL**. On the basis of **BL+CD**, the model **BL+CD+DD** additionally takes the distortion map as input to measure the effects of the content-dependency and the distortion-dependency simultaneously, which obtains further about 5.0% improvement on SROCC and 3.6% improvement on PLCC. Model **BL+DD+SD** doesn't consider the primary content as input, and only reaches 0.644 SROCC and 0.693 PLCC. Our proposed AIGQA (i.e., **BL+CD+DD+SD**) measures the image quality by integrating the effects of the content-dependency, the distortion-dependency and the degradation-dependency together, and achieves the highest performance, about 0.698 SROCC and 0.728 PLCC.

Then we analyze the effect of the two IGM-inspired constraints on BIQA performance. The experimental results are listed in Tab.IX. In **AIGQA w/o SS+SD**, both of the two constraints are removed and it gets the poorest results on both SROCC and PLCC. After adding the semantics similarity (or the structure dissimilarity) constraint into the objective function, a better performance is obtained by the model **AIGQA w/o SD** (or **AIGQA w/o SS**). And when considering both of the two constraints, the proposed AIGQA achieves the best results.

TABLE IX
ABLATION EXPERIMENTS ABOUT THE IGM-INSPIRED CONSTRAINTS

	SROCC	PLCC
AIGQA w/o SS+SD	0.637	0.686
AIGQA w/o SD	0.642	0.696
AIGQA w/o SS	0.678	0.688
AIGQA	0.698	0.728

Besides, to analyze the effect of the two IGM-inspired constraints more comprehensively, the semantics similarity (i.e., Eq.7) and structure dissimilarity (i.e., Eq.11) obtained by different models on real samples are compared directly. Taking the Fig.5(e) as an example (i.e., as the input distorted image of the active inference module). For **AIGQA**, the $\log\left[-\Re_{semantics}\left(I_g, I_d\right)\right]$ equals 9.531, $\Re_{structure}\left(I_g, I_u\right)$ equals 0.137. For **AGQA w/o SS+SD**, the $\log\left[-\Re_{semantics}\left(I_g, I_d\right)\right]$ equals 10.852, $\Re_{structure}\left(I_g, I_u\right)$ equals 0.156. **AIGQA** achieves higher semantics consistency and lower structure similarity, which verifies that the two IGM-inspired constraints can effectively endow the GAN-based active inference module with properties of IGM.

From the above series of ablation experiments, we can make some insightful conclusions. First, multifaceted quality analysis can effectively improve the accuracy and generalization of the proposed method. By integrating the multiple information as input, the proposed multi-stream quality evaluator could better leverage the characteristics of IGM for BIQA. Second, the predicted primary content also affects the performance of quality prediction. Because it determines the validity of the multiple prior information. Benefiting from the IGM-inspired constraints, the proposed active inference module is effective to predict the primary content of a distorted image and consequently improves the BIQA performance.

F. Visualization of the Prior Information

In this section, we will present some samples about the prior information obtained from the active inference module (i.e., the multi-stream inputs to the quality evaluator) to get an intuition of the proposed method.

In Fig.5, the first column is the white Gaussian noise (WN) distorted images at different distortion levels, from the second column to the last column are the corresponding primary content, distortion map (i.e., prediction error) and structural degradation map (i.e., structure similarity map). Since WN belongs to additive noise, it has little effect on the main semantics. IGM is effective to avoid the noise and infer the primary content with high quality, as show in Fig.5(b) and (f). Meanwhile, as can be seen from Fig.5(c) and (g), most noise is filtered into the distortion map, which will cause uncomfortable perception. Besides, as show in Fig.5(d) and (h), when distorted by different levels of WN, the structural degradation map has diverse patterns. Specifically, Fig.5(a) and (b) have higher structure similarity, which means a lower content degradation in Fig.5(a).

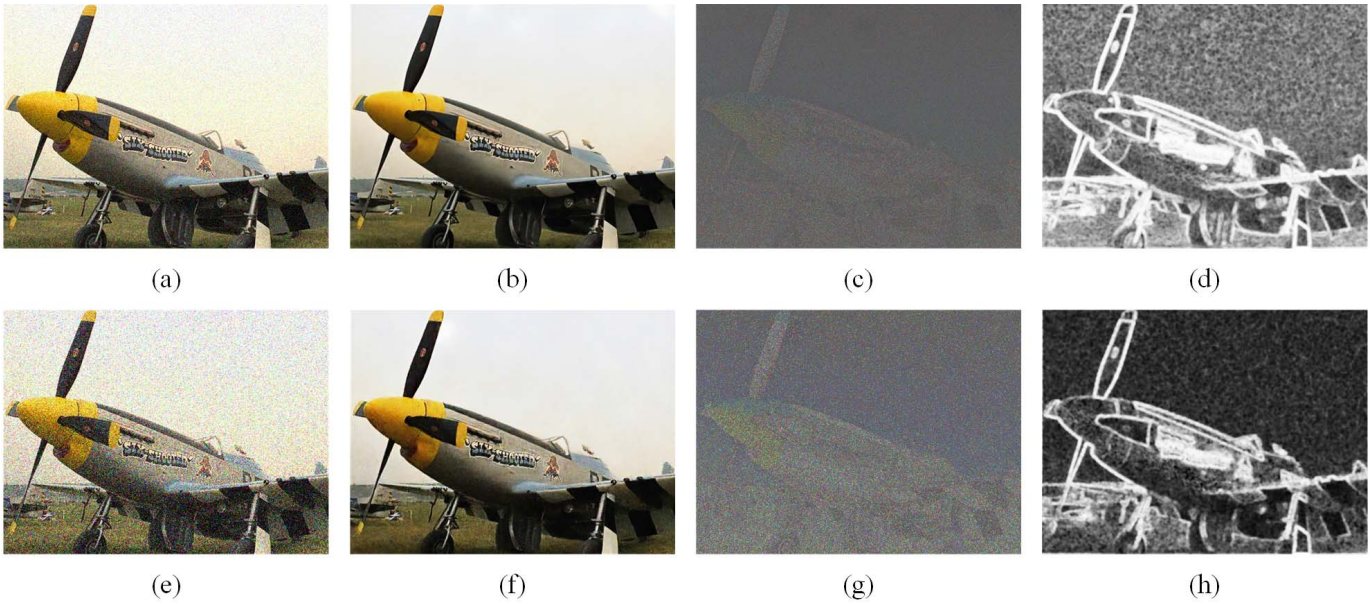


Fig. 5. Visualization of the calculated prior information when the input image is distorted by white Gaussian noise. (a) and (e) are the input distorted images at different distortion levels, which MOS values are 5.5 and 3.8 respectively. The lower the MOS, the higher the distortion level and the worse the perceptual quality. (b) and (f) are the primary content generated by the active inference module. (c) and (g) are the distortion map. (d) and (h) are the structural degradation map.

Fig. 6. Visualization of the primary content of a Gaussian blur distorted image. (a) Is the reference image. (b) Is the Gaussian blur distorted image. (c) Is the primary content generated by the active inference module.

In Fig.6, the primary content of a Gaussian blur (GB) distorted image is presented. GB will destroy the image content, such as blurring the edge or contour, and result in the loss of effective information. It's hard to infer the pristine content to give the perfect explanation for a severely GB distorted image, i.e., the primary content may still contain degraded information. For example, in the primary content (i.e., Fig.6(c)) of Fig.6(b), we still can't recognize the text message on the fuselage of the plane. While in the reference image Fig.6(a), the text message is clear.

From Fig.5 and Fig.6, we can draw the following conclusions. First, for different distortion types, the predicted primary content is consistent with the properties of IGM, which verifies the effectiveness of the proposed GAN in simulating the active inference process of IGM. Second, for different distorted images, their corresponding primary content, distortion map and structural degradation map show different characteristics. It's helpful to model the effects of the content-dependency, the distortion-dependency and the degradation-dependency together to evaluate the image quality.

V. Conclusion

In this paper, inspired by the IGM, we propose a novel IGM-inspired BIQA model for image quality prediction. Benefiting from the two proposed IGM-inspired constraints, the GAN-based active inference module is effective to simulate the IGM theory to predict the primary content of a distorted image. By integrating the multiple information obtained from the primary content, the multi-stream quality evaluator is effective to leverage properties of IGM for BIQA. A series of experiments demonstrate the effectiveness and superiority of the proposed method.

REFERENCES

[1] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, May 2011.

[2] 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.

[3] 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.

[4] 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.

[5] G. Zhai, "On blind quality assessment of JPEG images," in *Proc. 7th Int. Conf. Cloud Comput. Big Data (CCBD)*, Nov. 2016, pp. 322–328.

[6] 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, Mar. 2018.

[7] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[8] X. Yang, F. Li, and H. Liu, "A survey of DNN methods for blind image quality assessment," *IEEE Access*, vol. 7, pp. 123788–123806, 2019.

[9] D. C. Knill and A. Pouget, "The Bayesian brain: The role of uncertainty in neural coding and computation," *Trends Neurosci.*, vol. 27, no. 12, pp. 712–719, Dec. 2004.

[10] K. Friston, J. Kilner, and L. Harrison, "A free energy principle for the brain," *J. Physiol.*, vol. 100, nos. 1–3, pp. 70–87, Jul. 2006.

[11] K. Friston, "The free-energy principle: A unified brain theory?" *Nature Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, Feb. 2010.

[12] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41–52, Jan. 2012.

[13] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 43–54, Jan. 2013.

[14] 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.

[15] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Process.*, vol. 70, no. 3, pp. 177–200, Nov. 1998.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[17] 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.

[18] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, "No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1030–1040, May 2017.

[25] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1733–1740.

[26] 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. 2018.

[27] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[28] 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.

[29] 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. 2018.

[30] 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.

[31] 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.

[32] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 609–613.

[33] D. Varga, D. Saupe, and T. Sziranyi, "Deeprn: A content preserving deep architecture for blind image quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[34] 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.

[35] Y. Liang, J. Wang, X. Wan, Y. Gong, and N. Zheng, "Image quality assessment using similar scene as reference," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 3–18.

[36] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[37] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–17.

[38] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875

[39] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[40] P. Bongini, R. Del Chiaro, A. D. Bagdanov, and A. Del Bimbo, "GADA: Generative adversarial data augmentation for image quality assessment," in *Proc. Int. Conf. Image Anal. Process.*, 2019, pp. 214–224.

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

[42] H. Yang, P. Shi, D. Zhong, D. Pan, and Z. Ying, "Blind image quality assessment of natural distorted image based on generative adversarial networks," *IEEE Access*, vol. 7, pp. 179290–179303, 2019.

[43] S. Ling, J. Li, J. Wang, and P. Le Callet, "GANs-NQM: A generative adversarial networks based no reference quality assessment metric for RGB-D synthesized views," 2019, *arXiv:1903.12088*. [Online]. Available: http://arxiv.org/abs/1903.12088

[44] Y. Ma, X. Cai, F. Sun, and S. Hao, "No-reference image quality assessment based on multi-task generative adversarial network," *IEEE Access*, vol. 7, pp. 146893–146902, 2019.

[45] H. Ren, D. Chen, and Y. Wang, "RAN4IQA: Restorative adversarial nets for no-reference image quality assessment," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–7.

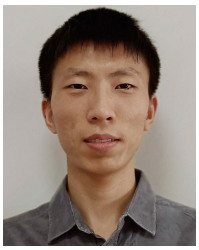
[46] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8183–8192.

[47] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[48] 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. 2013.

[49] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

[50] 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.
[51] 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.
[52] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, Jan. 2017.
[53] J. Guan, S. Yi, X. Zeng, W.-K. Cham, and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2505–2520, Nov. 2017.
[54] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.
[55] D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, and Y. Zhang, "Blind predicting similar quality map for image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6373–6382.
[56] K. Ma *et al.*, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.
[57] 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.
[58] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[59] 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, pp. 19–21, 2010.
[60] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, 2013, pp. 1693–1697.
[61] 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.
[62] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.
[63] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[64] S. Ramakrishnan, S. Pachori, A. Gangopadhyay, and S. Raman, "Deep generative filter for motion deblurring," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2993–3000.
[65] P. Gastaldo, R. Zunino, and J. Redi, "Supporting visual quality assessment with machine learning," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, pp. 1–15, Dec. 2013.
[66] D. Varga, "Comprehensive evaluation of no-reference image quality assessment algorithms on KADID-10k database," 2020, *arXiv:2010.09414*. [Online]. Available: http://arxiv.org/abs/2010.09414
[67] M. Oszust, "Local feature descriptor and derivative filters for blind image quality assessment," *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 322–326, Feb. 2019.
[68] D. Varga, "Multi-pooled inception features for no-reference image quality assessment," *Appl. Sci.*, vol. 10, no. 6, p. 2186, Mar. 2020.



Jupo Ma received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Xidian University, Xi'an, China. His research interests include image quality assessment, deep learning, and dynamic vision sensor (DVS).

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 from 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 Quality Assessment.

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 City, 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, a 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 also 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, where he 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 also serving as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and *SIAM Journal of Imaging Sciences*.

Xuemei Xie (Senior Member, IEEE) received the M.S. degree in electronic engineering from Xidian University in 1994 and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong in 2004. She is currently a Professor with the School of Artificial Intelligence, Xidian University. She has published over 100 academic articles in international journals and conferences. Her research interests include human action recognition, object detection, scene understanding, video analysis, deep learning, and feature representation.

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 is currently the Academic Leader of Circuits and Systems with Xidian University. He has authored or coauthored over 200 articles 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 implementation of algorithms for intelligent signal processing. He awarded the Cheung Kong Scholar Chair Professor by Ministry of Education in 2012. He has served as the Chair for the 90th MPEG and 50th JPEG of the International Standards Organization (ISO) and a Technical Program Chair for FSKD06, VSPC 2009, IEEE PCM 2009, SPIE VCIP 2010, and IEEE ISCAS 2013.

Weisi Lin (Fellow, IEEE) received the Ph.D. degree from the King's College, London University, U.K. He has served as the Lab Head of Visual Processing for the Institute for Infocomm Research, Singapore. He is currently an Associate Professor with the School of Computer Engineering. His research interests include image processing, perceptual signal modeling, video compression, and multimedia communication, in which he has published 170 journal articles, more than 230 conference papers, filed seven patents, and authored two books. He is also a Fellow of IET, and an Honorary Fellow of the Singapore Institute of Engineering Technologists. He has been a Technical Program Chair of IEEE ICME 2013, PCM 2012, QoMEX 2014, and IEEE VCIP 2017. He is also an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He had been an invited/panelist/keynote/tutorial speaker in more than 20 international conferences, as well as a Distinguished Lecturer of IEEE Circuits and Systems Society from 2016 to 2017, and the Asia-Pacific Signal and Information Processing Association (APSIPA) from 2012 to 2013.