Neurocomputing 547 (2023) 126378

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Reduced-reference image deblurring quality assessment based on multi-scale feature enhancement and aggregation

Bo Hu [a,b,c], Shuaijian Wang [a,c], Xinbo Gao [a,c,*], Leida Li [d], Ji Gan [a,c], Xixi Nie [a,c]

[a] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[b] Jiangsu Key Lab of Image and Video Understanding for Social Security, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, PR China
[c] Chongqing Institute for Brain and Intelligence, Guangyang Bay Laboratory, Chongqing 400064, China
[d] School of Artificial Intelligence, Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 30 June 2022
Revised 21 March 2023
Accepted 22 May 2023
Available online 26 May 2023

Keyword:
Image quality assessment
Image deblurring
Vision transformer
Blurry image
Deblurred image

ABSTRACT

Image deblurring is a basic task in the field of computer vision, and has attracted much attention because of its application prospects in traffic monitoring and medical imaging, etc. Due to the inherent weakness of the model, it is difficult to obtain well-pleasing deblurred images for all the visual contents so far. Therefore, how to objectively evaluate the quality of these deblurred results is very important for the rapid development of image deblurring. In recent years, numerous convolutional neural networks based quality assessment methods have been proposed to automatically predict the quality of synthetic and authentic distorted images, producing results that are mildly consistent with subjective perception. However, they are limited in Image Deblurring Quality Assessment (IDQA). For IDQA, it is more meaningful to predict the quality difference of blurry-deblurred image (BDI) pair than to make prediction on single deblurred image. Inspired by this, we propose a novel reduced-reference image deblurring quality assessment method based on multi-scale feature enhancement and aggregation. Firstly, the multi-scale features of BDI pair are generated from a versatile vision Transformer. Secondly, the discrepancy information is exploited to implicitly enhance the initial deep features. Finally, the enhanced features of different scales are aggregated and then mapped to the quality difference of BDI pair. Experimental results on four challenging datasets demonstrate that the proposed method is superior to the state-of-the-art quality assessment methods.

1. Introduction

Image deblurring is a frontier in computer vision and many image deblurring algorithms have been proposed to recover potential high-quality images from blurry images [2–5]. Especially with the prevalence of Convolutional Neural Networks (CNN), image deblurring has made great progress. However, almost all image deblurring algorithms are difficult to recover an undistorted high-quality image due to inherent weakness of the model, diversity of distortion and variation of content, etc. To make matters worse, instead of improving, image quality actually decreased during the deblurring process. Such an example is given in Fig. 1. The second column represents the blurry images, and the first and third columns are the corresponding deblurred images. The quality of the deblurred versions in the first column is worse than that of the blurry images, as can be observed in Fig. 1. Therefore, how to evaluate these deblurred results and algorithms objectively and fairly is particularly important. In literature, researchers usually compare the restoration performance of image deblurring algorithms in two respects, namely subjective evaluation and objective evaluation [6]. It is well known that the former cannot accomplish this task when there is a lot of data needs to be processed in real time. Objective evaluation, whose purpose is to predict the image quality automatically by constructing models, has attracted much attention and made great strides in solving this problem. When the reference image exists, the famous PSNR and SSIM [7] are frequently applied in Image Deblurring Quality Assessment (IDQA). However, the reference image is difficult to obtain or even nonexistent in an open environment. No-reference (NR) image quality metric, which only relies on distorted images for quality modeling, theoretically can handle the situation.

* Corresponding author at: Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.
E-mail address: gaoxb@cqupt.edu.cn (X. Gao).

https://doi.org/10.1016/j.neucom.2023.126378
0925-2312/

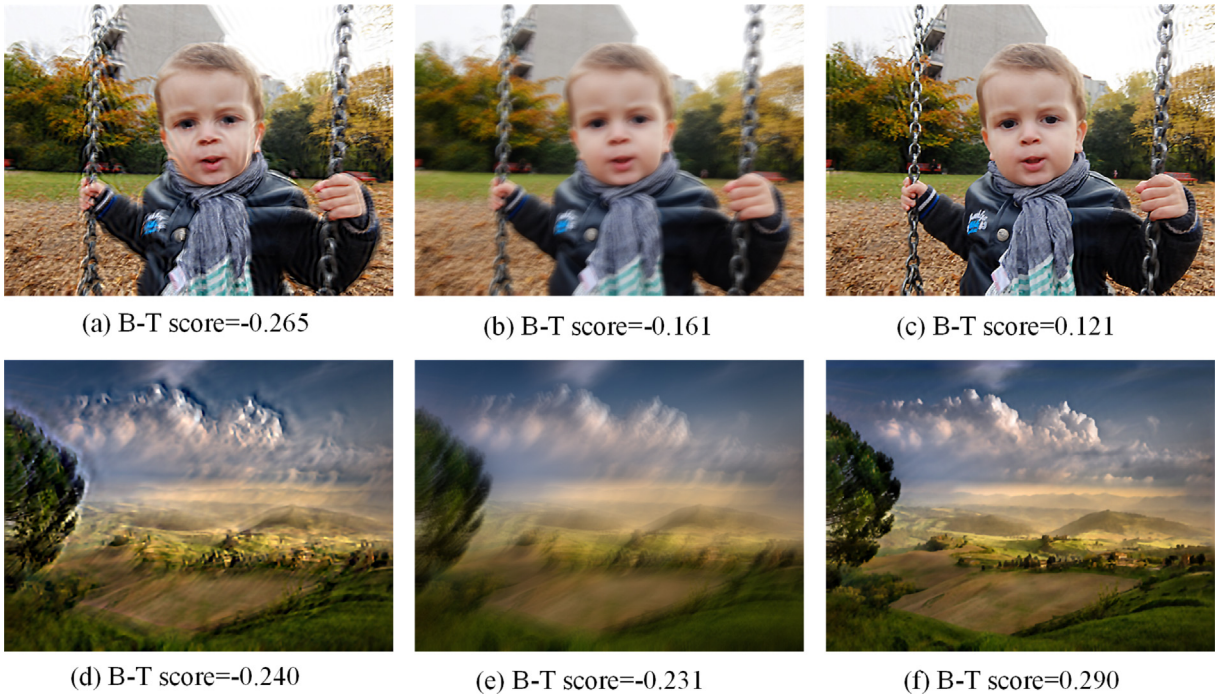


Fig. 1. Image deblurring results in the DNB database [1]. The second column represents the blurry image, and the first and third columns are the corresponding deblurred versions by two different algorithms. The quality of the deblurred versions in the first column is inferior to that of the blurry images. Better quality is indicated by a higher Bradley-Terry (B-T) score.

Over the recent years, it has witnessed tremendous progress in NR quality assessment and it has achieved remarkable results on common databases, such as CSIQ [8] and TID2013 [9]. Generally, blind quality metrics can be roughly divided into hand-crafted and deep learning-based quality metrics [10,11]. For most hand-crafted based metrics, it can be further divided into opinion-aware methods [12–14] and opinion-unaware methods [15,16]. In the process of model building, the former requires the participation of subjective scores, while the latter does not. They show promising performance in the evaluation of synthetic distortions, but there are some problems such as poor representation ability of hand-crafted features and weak regression model. To alleviate these problems, deep learning based metrics have gained wide attention and have won the favor of researchers [17,18]. Studies have shown that CNN-based metrics have more advantages than traditional hand-crafted based metrics in respect to both prediction accuracy and generalization ability. More recently, the success of Transformer network in computer vision and Natural Language Processing (NLP) has attracted the attention of researchers in the field of IQA. Some Transformer-based models have been proposed successively and have achieved relatively satisfactory results [19,20]. However, these quality metrics cannot be directly used to IDQA [21]. The reasons are manifold and are analyzed as follows. Firstly, most traditional NR quality methods are developed and tested on uniform and common distortion types, which evenly distributed throughout the image, such as Gaussian noise and Gaussian blur. However, ringing effect, which tends to be localized, often occurs in deblurred images, especially around the edges. Secondly, most of the NR quality methods are designed to deal with distorted images with a single distortion, while the deblurred images usually contain multiple distortions, such as residual blur, ringing effect, unnatural structure and detail loss, etc. Thirdly, existing NR quality methods can only model and generate a scalar score for each distorted image and cannot accurately evaluate the relative quality difference between two images, which makes more sense in IDQA. Therefore, quality assessment models for image deblurring are urgently, which is also the main work of this paper.

For image deblurring quality evaluation, quantifying the relative quality difference of blurry-deblurred image (BDI) pairs is more valuable than quality prediction based on deblurred images only, because the former not only provides performance ranking, but also intuitively predicts the increase or decrease in quality during the deblurring process. In theory, most existing image quality evaluation algorithms can predict the quality of blurred and deblurred images separately, and then calculate the difference to derive the quality difference of the BDI pairs. However, it may be difficult to accurately predict the difference in quality between the two images because none of these methods fully utilize the difference information between the deblurred and blurred images. Inspired by this, we present a simplified reference image deblurring quality metric based on Siamese network structure with multi-scale feature enhancement and aggregation. The underlying idea of this work is to quantify the relative quality difference of BDI pair in multi-scale space by combining vision Transformer and CNN, so as to achieve an objective model consistent with human visual perception. Firstly, multi-scale deep features of BDI pair are generated from a versatile Transformer network. Then discrepancy features are designed to implicitly guide feature learning and enhancement in multi-scale representation. Finally, the relative quality difference of BDI pair is obtained after multi-scale feature aggregation and regression. Experimental results on four challenging datasets confirm that the proposed method outperforms the compared quality assessment methods. In addition, it has the best generalization performance in the cross-dataset validation experiments.

The main contributions of this work are summarized as follows:

- Based on the fact that discrepancy information is important and rarely studied for IDQA, we propose a reduced-reference image deblurring quality assessment method with versatile Trans-

former and flexible CNN. Different from most existing methods, the proposed method predicts the quality difference of a BDI pair rather than predicting the quality of a single deblurred image.
- An implicit discrepancy information guidance based feature enhancement (DGE) module is proposed to implicitly enhance the initial deep features of BDI pairs, thus making the features more discriminating. Moreover, a multi-scale feature aggregation module is proposed to obtain more comprehensive feature representation.
- Extensive experimental results demonstrate the proposed method outperforms the state-of-the-arts by a large margin on both synthetic and real deblurring datasets. Besides, the proposed method is superior to the compared quality metrics in terms of generalization ability.

2. Related work

This section provides a brief review of the literature related to image deblurring, image deblurring quality assessment and vision Transformer.

2.1. Image deblurring

With the frequent use of cameras, shake is inevitable when using a hand-held camera to take photographs, which can produce an unclear image, known as a blurry image. Blur is mainly manifested as widened edges and loss of detail, which has a serious impact on reflecting real scenes, so research into deblurring algorithms is essential. The goal of image deblurring is to generate a high-quality original version from a blurry image. Over the last decade, a large number of deblurring algorithms have been proposed with remarkable results, and they can be roughly classified into optimization-based and deep learning-based algorithms [22]. This subsection briefly reviews these related algorithms and briefly describes their advantages and disadvantages.

The optimization-based methods aim to recover images by making full utilization of a priori information about the image, which is obtained mainly by counting certain features of the image and calculating their distribution. Some representative prior knowledge includes super-Laplacian prior [23], sparse gradient [24–26], face piece recurrence prior [27], normalized sparse prior [28], L_0-norm prior [29], discriminative learning prior [30], etc. Some existing optimization-based algorithms utilising the above prior knowledge have been able to provide competitive results on general natural images. However, these prior knowledge does not generalise well to some specific scenarios. Therefore, specific prior knowledge has been introduced again for images in specific scenes, for example, content-aware prior for foreground segmentation [31] for images in text scenes and light streak prior knowledge [32] for images in low light scenes. Pan *et al.* [33] found that the sparsity of the dark channel varies with the degree of blurring of the image and proposed a method to enhance the sparsity of the dark channel, which can be well applied to images in various scenes. With regards to photos with a lot of brilliant pixels, this prior is less effective. To solve this problem, Yan *et al.* [34] proposed the bright channel prior knowledge to process images containing a high amount of bright pixels, and this method took advantage of both the bright and dark channel priors and achieved good results in major scenes. Although optimization-based algorithms have achieved good results in image deblurring, these methods suffer from two common problems, namely the time-consuming parameter adjustment process and the simplifying assumptions on the blurring model, which hinder their performance in practical situations.

With the advancement of deep learning in an unprecedented way, a high amount of CNN-based deblurring algorithms have emerged. Deep learning based models concentrate on self-learning using large amounts of data to learn the mapping functions in the image degradation process [35,36]. Nah et al. [37] proposed a multi-scale network for removing motion blur from images and a multi-scale loss function with the traditional coarse-to-fine idea incorporated into it. Zhang *et al.* [38] presented a deep stacked hierarchical multi-patch network which processes blurred images by layering the representation from fine to coarse. The method enables blurred images of different scales to be concerned by the network, and finally extracts multi-scale blurred information. Kupyn *et al.* [39] proposed an image deblurring deep network in an end-to-end manner, namely DeblurGAN. Specifically, the generator consists of two transformed convolution blocks, two-strided convolution blocks, and nine residual blocks to generate clear images corresponding to blurred images. Cai *et al.* [22] proposed DBCPeNet, which incorporates the prior knowledge of the bright and dark channels into the objective function and naturally embeds the prior knowledge into the neural network as a way to achieve effective deblurring. To deal with bokeh blurring of single images, Ma *et al.* [40] proposed a method, namely DID-ANet, which is based on a multi-task framework, and the network uses predicted bokeh maps as subtasks to improve the effectiveness of deblurring. Although the CNN-based algorithms perform well, they are data-driven and require high amounts of data to train the model.

2.2. Image deblurring quality assessment

Compared with the booming image deblurring, the quality evaluation research in this field is still in its infancy, and few works have been reported [41]. In [42], the authors argue that good deblurring results should have a high degree of naturalness and sharpness, and therefore propose a method to extract a group of features to measure the extent to which the deblurring result satisfies these two constraints in order to assess the perceived quality of the deblurred results, and that the method is a NR metric that can be better applied to real scenes. Inspired by [42], Li et al. [43] designed a novel NR quality assessment framework for image deblurring by measuring residual blur, noise and ringing (NRRB). Residual blur, ringing and noise usually co-exist in the deblurred images. Based on this observation, these three aspects were measured by extracting specific features. By combining these three scores, the total quality score of a deblurred result was then determined. The above metrics are designed for a specific restoration scenario, namely image deblurring. In recent years, researchers have made great efforts to evaluate the quality of general image restoration. In [21], the authors proposed a rank learning framework that is based on pairwise comparison. The approach is able to extract low-level features from multiple domains to characterize multiple types of distortions, and eventually combine these features to build a generic image restoration quality model. In [41], a reduced-reference image restoration quality metric was proposed based on the difference information of distorted-resorted image pairs. Although they have advanced the development of the field to a certain extent, they also have obvious weaknesses and much room for improvement. Firstly, most existing metrics are designed based on hand-crafted features and thus cannot comprehensively describe such complex distortions, which leads to unsatisfactory predictions. Secondly, the CNN-based methods only use the highest-level features to learn a mapping function, and lack the rational exploitation of multi-level features, and there are natural limitations to the perceptual field of CNN. Finally, there is a lack of research on relative quality difference of BDI pair, which

is a natural index to evaluate the performance of deblurring algorithms.

2.3. Vision transformer

Transformer was originally a technique applied in NLP with great success [44,45]. Prior to Transformer, NLP used recurrent neural networks (RNNs) to model sequences and thus compute the underlying connections between sequences. However, RNNs are difficult to compute in parallel, a drawback that makes them poor at scaling large datasets as well as complex models. Transformer employs a Multi-head Self-attention (MHSA) mechanism to model long-distance feature sequences. Differs from convolutional layers, the MHSA layer has a global acceptance domain and dynamic weights, and it can be computed in parallel. As a result, the Transformer can easily handle long sequences and exhibit good performance.

The success of Transformer and MHSA in NLP has caught the eyes of researchers in computer vision. To extend the Transformer to computer vision fields, Dosovitkiy et al. [46] designed a novel Transformer framework, namely Vision Transformer (ViT). For ViT, the input image is divided into small sized patches and then each patch is encoded. The linear embedding sequence of these patches is then used as input to the Transformer, which eventually predicts the image class using Multi-layer Perceptron (MLP). Yuan *et al.* [47] argue that the simple token structure in ViT fails to model the local structure (edges and lines) of the image, which leads to a lack of local features and reduces the training efficiency, and that the attentional design in ViT is redundant. Therefore, T2T-ViT gradually structures the image into tokens by merging adjacent tokens by means of a Token-to-Token transformation, so that local features can be modelled and the length of tokens can be reduced. Liu *et al.* [48] proposed a sliding window based mechanism Transformer, namely Swin Transformer, which uses a layered structure similar to CNN to design the network, allowing the model to handle images of different sizes flexibly. The sliding window mechanism allows the computation of self-attention to be restricted within each window, greatly reducing the computational complexity, and also allows cross-window connections. It also allows cross-window connections, ensuring that the individual windows are connected to each other. Among the recent Transformer-based networks, the Pyramid Vision Transformer (PVT) [49] with its pyramid structure is active in several vision tasks and has achieved significant success in image segmentation, image classification, and object detection. In this paper, we designed a new Siamese Transformer-based image deblurring quality evaluation framework based on PVT, which can accurately predict the quality differences between deblurred images and blurry images.

3. Proposed method

Most existing methods make quality prediction on a single distorted image. This results in them not being able to accurately provide the discrepancy information of BDI pairs, which directly indicates whether the quality of the deblurred image is improved in the deblurring process. Based on the fact that the discrepancy information of BDI pair is very important and rarely studied for IDQA, we propose to exploit discrepancy information to implicitly guide multi-scale feature enhancement and aggregation in a reduced-reference manner. The structure diagram of the proposed method is shown in Fig. 2. Firstly, a versatile vision Transformer network is used to generate hierarchical feature maps of BDI pairs. Secondly, a discrepancy information guidance based feature enhancement module is proposed to conduct feature learning and enhancement for deep features at multi-scale space. After that, the enhanced features are fed into the multi-scale feature aggregation module to account for complex distortions. Finally, the feature vector is mapped to the relative quality differences of BDI pairs through MLP. The following subsections describe the proposed method in detail.

3.1. Vision transformer-based multi-scale feature extraction

Transformer has been a significant success in NLP and has naturally attracted extensive attention from researchers in computer vision. Lately, numerous vision-Transformer networks are active in several vision tasks (image classification, object detection and image segmentation, etc.) and encouraging progresses have been shown [46]. Compared to CNN, one advantage of Vision Transformer is that it can freely extract the global information of an image, which is also crucial for IDQA.

Inspired by this, the feature extraction module is designed in a vision Transformer framework. Specifically, the versatile pyramid vision Transformer V2 (PVTv2) [50] is exploited for feature extraction due to the significant performance on pixel-level prediction tasks. PVTv2 is separated into four stages, including a Transformer encoder and a patch embedding layer in each level. The overlapping patch embedding layer is used to label the image and change the resolution of the feature map by varying the stride of the convolution, which compensates to some extent for the local informa-

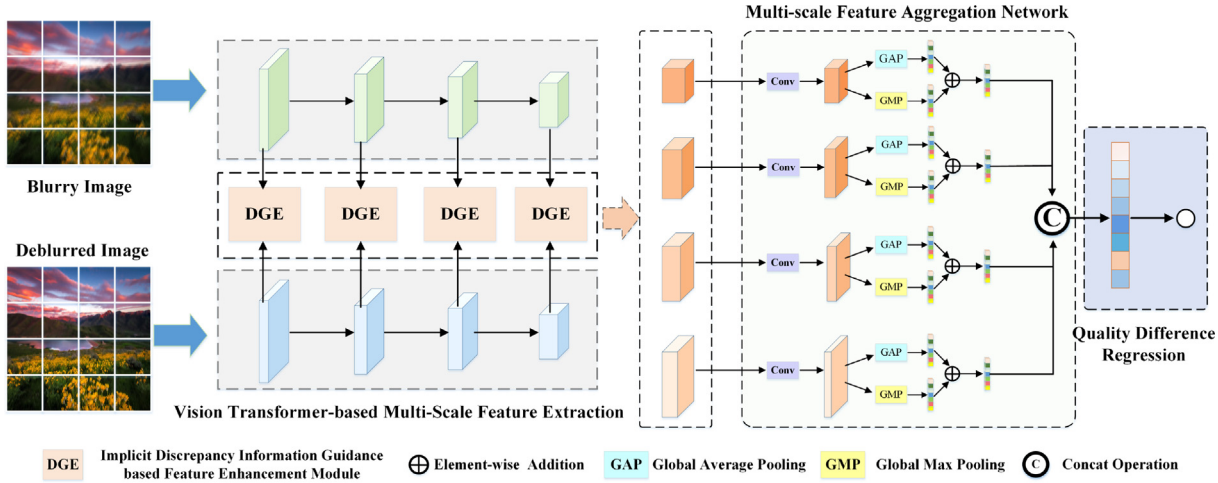


Fig. 2. Framework of the proposed reduced-reference image deblurring quality assessment method.

tion continuity of the image as the adjacent windows are overlapping. For example, given an input of size $m \times m \times c$, the stride is set to S, the padding size is set to $S-1$, the convolution kernel size is set to $2S-1$, and the number of convolution kernel of c'. Ultimately, the output size should be that of $\frac{m}{S} \times \frac{m}{S} \times c'$. By incorporating the pyramid structure from CNN, PVTv2 produces deep feature maps of four different scales, namely $\{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$, and these maps progressively shrink from high (4-stride) to low (32-stride).

Unlike the traditional Transformer module, the Transformer block of PVTv2 is implemented based on the Spatial-Reduction Attention (SRA) module, which significantly reduces the computational cost. The details of the SRA in stage i can be expressed as:

$$SRA(Q,K,V) = Concat(head_0, \ldots, head_{N_i})W^o, \tag{1}$$

$$head_j = Attention(QW_j^Q, SR(K)W_j^K, SR(V)W_j^V), \tag{2}$$

where Q, K, V denote the query matrix, key matrix and value matrix mapped from the input. $Concat(\cdot)$ is the concatenation operation. The attention layer of stage i contains N_i heads. $W_j^Q \in \mathscr{R}^{C_i \times d_h}, W_j^K \in \mathscr{R}^{C_i \times d_h}, W_j^V \in \mathscr{R}^{C_i \times d_h}$, and $W^O \in \mathscr{R}^{C_i \times C_i}$ are learnable parameters. d_h denote the dimension of each head, which is equal to $\frac{C_i}{N_i}$. $SR(\cdot)$ is the calculation used to decrease the spatial dimension of the K and V sequence, which is formulated as:

$$SR(x) = Norm(Reshape(x, R_i)W^S), \tag{3}$$

where $x \in \mathscr{R}^{(H_i W_i) \times C_i}$ is an input sequence, and R_i denote the rate of space reduction in stage i. The $Reshape(x, R_i)(\cdot)$ operation reshapes the shape of the input sequence x to $\frac{H_i W_i}{R_i^2} \times (R_i^2 C_i)$. $W_S \in \mathscr{R}^{R_i^2 C_i \times C_i}$ is a learnable linear projection parameter that maps the dimension of the input sequence to C_i. $Norm(\cdot)$ is a layer normalization operation. The attention operation is formulated as:

$$Attention(\mathbf{q}, \mathbf{k}, \mathbf{v}) = Softmax(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_h}})\mathbf{v}. \tag{4}$$

PVTv2 repeats above operations at different stages, eventually extracting features at four different scales and with global information, which is very important for IDQA.

3.2. Implicit discrepancy information guidance based feature enhancement network

Features from the Vision Transformer network have preliminary quality perception ability, which can deal with slight distortion to a certain extent. However, they are difficult to handle complex distortions in image deblurring, which leads to performance degradation of the model. In [51], the difference information between the blurred and reference maps is introduced to motivate the model to accurately perceive the quality of the blurred image.

Inspired by this, in IDQA, we introduce the difference information between blurred and deblurred images to guide the model in feature learning. Specifically, we first calculate the difference information between the deblurred image and the blurred image in the deep feature space. Then, the difference feature is fed into the convolutional layer to generate the spatial guidance map and the channel guidance map. After that, we use the spatial guidance map and the channel guidance map to guide the model to further enhance the features of BDI pair in the spatial dimension and the channel dimension respectively. Finally, the features of these two parts are added with their original features, and then concated together to obtain the final enhanced features through a convolution layer. It is well known that humans perceive objects following a hierarchical perception mechanism (from local structure to global semantics). Therefore, we embed the DGE module at each stage of feature extraction so that the model can extract information from local structure to global semantics, which is very important for IQA. The diagram of this module is shown in Fig. 3. The objective of the proposed method is to calculate the quality difference between the deblurred image and the blurry image, which is a difference value. Therefore, intuitively, we use the difference of their feature maps to represent the difference information. Therefore, in each stage, the discrepancy information of BDI pair is obtained by:

$$\mathbf{F}_s^i = \mathbf{F}_d^i - \mathbf{F}_b^i, \tag{5}$$

where $\mathbf{F}_d^i$ and $\mathbf{F}_b^i$ are the i^{th} deep feature maps of BDI pairs. Then the discrepancy information $\mathbf{F}_s$ is translated into two types of guidance weights, which are formulated as:

$$\mathbf{P} = Sigmoid(P(\mathbf{F}_s)), \tag{6}$$

$$\mathbf{c} = Sigmoid(C(\mathbf{F}_s)), \tag{7}$$

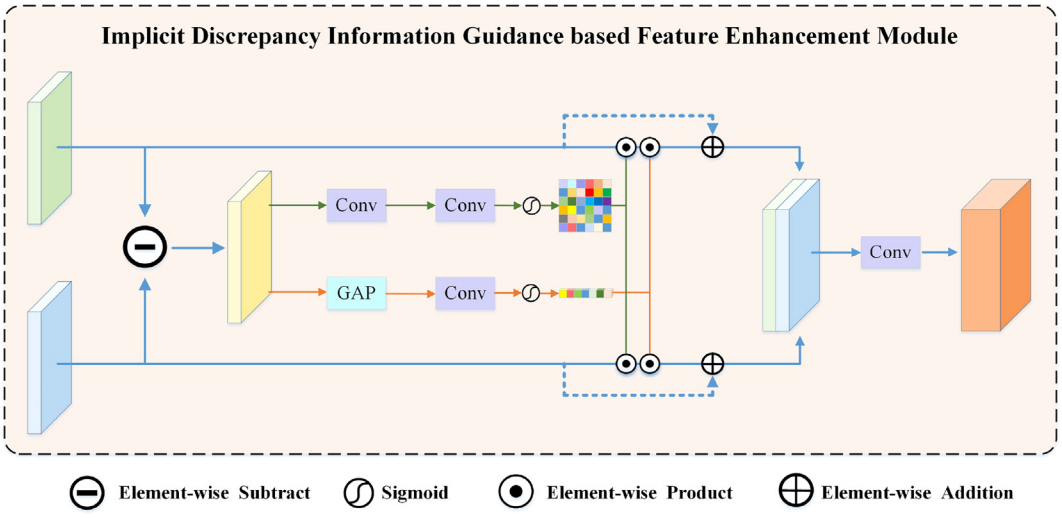


Fig. 3. Diagram of the implicit discrepancy information guidance based feature enhancement module.

where function P consists of two-layer convolution. Function C goes through global average pooling (GAP) first, and then a convolutional layer.

The feature maps of both deblurred and blurry images are guided by the implicit discrepancy information, which is defined as:

$$\mathbf{F}\prime_i = \mathbf{F}_i \oplus (\mathbf{F}_i \odot \mathbf{P} \odot \mathbf{c}), \tag{8}$$

where $i = \{d, b\}$, "$\oplus$" denotes element-wise addition, "$\odot$" denotes element-wise product. Finally, the output feature of this module is formulated as:

$$\mathbf{F} = Conv(\mathbf{F}\prime_d © \mathbf{F}\prime_b), \tag{9}$$

where "©" denotes concat operation.

3.3. Multi-scale feature aggregation network

After the feature enhancement process mentioned above, this module will further process the features of each stage to obtain a more robust and high-expressive representation. To be specific, after passing through a convolution layer, the GAP and global max pooling are carried out first and then a simple addition operation is used to combine the two parts, which is defined as:

$$\mathbf{f} = GAP(Conv(\mathbf{F})) + GMP(Conv(\mathbf{F})). \tag{10}$$

Finally, concat operation is performed on the feature vectors obtained in the four stages, which is formulated as:

$$\mathbf{f}\prime = \mathbf{f}_1 © \mathbf{f}_2 © \mathbf{f}_3 © \mathbf{f}_4. \tag{11}$$

3.4. Relative quality difference regression

To obtain the relative quality difference of BDI pair, we introduce a three-layer MLP based regression network, which is formulated as:

$$q = MLP(\mathbf{f}\prime). \tag{12}$$

Then, the MSE loss function is applied to optimize the proposed model [18], which is formulated as:

$$L = \frac{1}{N}\sum_{i=1}^{N}(\hat{q}_i - q_i)^2, \tag{13}$$

where N is the number of BDI pairs for training, $\hat{q}_i$ is the relative quality difference of i-th BDI pair predicted by the proposed method, q_i denotes the relative quality difference of i-th BDI pair, and it is defined as:

$$q = m_d - m_b, \tag{14}$$

where m_d and m_b are the subjective scores of deblurring and blurry images in a BDI pair.

4. Experiments

4.1. Experimental setting

Implementation Details. PVTv2 can be divided into seven versions depending on the settings of the hyper-parameters [50]. In this work, PVTv2-B1 with parameters pre-trained on ImageNet is used as encoder to balance performance and computation costs. We resize the BDI pairs to 512 × 512 to feed into the encoder. The batch size is set to be 16. The initial learning rates of the PVTv2-B1 encoder and the added layer are set as 1e-5 and 1e-4, respectively. Moreover, the Adam Weight Decay Optimizer (AdamW) is used to optimize the model. The remaining hyper-parameters are set as follows: weight decay of 1e-2, and total epoch of 100. Pytorch [52] is used to implement the proposed method. The training and testing processes are implemented on a server with Intel 10-core I9-10900X CPU, 32 GB RAM and NVIDIA GeForce RTX 3090 GPU.

Databases. Four challenging image deblurring quality datasets are employed to evaluate the performance of the proposed method, including motion deblurring quality database (MDD) [42], image deblurring for uniform blur (DUB) database, image deblurring for non-uniform blur (DNB) database and image deblurring for real blur (DRB) database [1]. For MDD database, it includes 240 blurred images with different levels of distortion, and each blurred image is processed by five representative deblurring algorithms, resulting in 1199 deblurred images. Each image in the database has a B-T score. For DUB and DNB databases, 25 high-quality images were contaminated with four different degrees of simulated uniform and non-uniform blur, resulting in 100 uniform and non-uniform blur images, respectively. For DRB database, 100 real blur images were collected from different camera devices. Then, 13 image deblurring algorithms were used to restore these three types of blurry images, producing 1300 deblurring images for each distortion type. Finally, human subject study was carried out to obtain B-T scores of these images using Amazon Mechanical Turk.

Criteria. In general, the Pearson linear correlation coefficient (PLCC) and root mean-squared error (RMSE) are employed to evaluate the predication accuracy [53,54]. Spearman rank order correlation coefficient (S_r) and Kendalls rank order correlation coefficient (K_r) are employed to evaluate the predication monotonicity [55]. The proposed method predicts the relative quality difference between the deblurred image and the blurred image, and then the relative quality difference is compared to obtain the performance ranking of the deblurring algorithms, which means that the proposed method is concerned with the monotonicity between the predicted results and the subjective values. Therefore, we use S_r and K_r to evaluate the predication monotonicity of the proposed method, instead of PLCC and RMSE to evaluate the predication accuracy. The S_r is defined as:

$$S_r = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2, \tag{15}$$

where d_i is the difference between the i-th BDI pair's ranks in objective and subjective evaluations. The second evaluation metric is K_r, which is formulated as:

$$K_r = \frac{n_a - n_d}{1/2n(n-1)}, \tag{16}$$

where n_a denotes the number of accordant pairs in the testing dataset, n_d denotes the number of dissonant pairs in the testing data set. In general, better quality evaluation metrics lead to better S_r and K_r values. In the implementation, we need to calculate the S_r and K_r values for the deblurred images in groups, which are derived from the same blurred image and processed by different image deblurring methods. Higher S_r and K_r values represent better performance.

4.2. Performance evaluation

To assess the effectiveness of the proposed method, we compare it with four kinds of state-of-the-art quality assessment methods including: 1) six traditional hand-crafted based quality metrics [12–16,56]; 2) two representative deep learning-based methods [17,18]; 3) two quality metrics designed for image deblurring [42,43] and 4) two general-purpose quality metrics for image restoration [21,41]. For fair comparison, all learning-based metrics are re-trained on the corresponding dataset. More specifically, we divide the database into two parts, as in [17,21], namely training

subset (80% sample data) and test subset (20% sample data) based on visual content. Then, the training and testing processes were repeated 20 times, and the average performance was reported.

The performance of different metrics on the MDD, DUB, DNB and DRB databases is summarized in Table 1, where the best outcomes are denoted in bold in the table. From these results, we can draw the following conclusions: 1) it can be observed that the proposed method is obviously superior to the compared quality metrics on these four databases. The possible reason for this result is that the proposed method adopts a more reasonable framework and feature processing mode, namely multi-scale feature enhancement and aggregation. 2) Among existing metrics, Hyper [17] has achieved encouraging results, with its S_r being the only one above 0.68 on the challenging DRB database. However, it is much lower than the S_r of the proposed method, namely 0.768. 3) Ref. [21] and Ref. [41], which are specifically designed for image deblurring, achieved similar results and only moderate performance. One potential reason is that they do not make proper use of multi-scale features. 4) The performances of the hand-crafted based metrics are lower than that of the CNN-based metrics. Generally, the metric for image deblurring outperforms the general-purpose quality metric. 5) The DRB database is more challenging than the other three databases. Overall, these results demonstrate the superiority of the proposed method. Thus, the significance of discrepancy information is confirmed.

4.3. Visualization analysis

This paper focuses on the discrepancy information of BDI pairs, which plays a crucial role in the performance of the proposed method. Therefore, it is necessary to visualize and analyze the extracted discrepancy information. To this end, visualization analysis of discrepancy information is explored in this subsection. For the proposed method, the discrepancy information was extracted at all four stages. For simplicity, only the deep features of the first stage were shown. The visualization of the discrepancy information of the proposed method is shown in Fig. 4. The first column shows the blurry images, the second column shows the corresponding feature maps of (a) and (f), the third column shows the deblurred images, the fourth column shows the corresponding feature maps of (c) and (h), the discrepancy information is shown in the last column.

It can be seen from the figure that the discrepancy information mainly contain the structure information, which has been proved to be crucial for image quality assessment [13]. Therefore, it is feasible and effective to use the discrepancy information to guide the feature learning of blurred image and deblurred image, which will be further demonstrated by ablation experiments in the next subsection.

4.4. Impact of different modules

To verify the effectiveness of the two proposed modules, namely discrepancy information guidance based feature enhancement (DGE) module and multi-scale feature aggregation (MFA) module, we further conduct ablation experiments. To be specific, we removed the two modules separately and then tested the performance of these methods. The experimental results are given in Fig. 5.

As can be seen from the Fig. 5, the performance of the method decreases regardless of which module is removed. The metric with these two modules, namely the proposed method, achieved the best results. These results indicate that both the modules play an important role in IDQA. This also shows the necessity of combining these two modules.

4.5. Impact of different pooling strategies

In the proposed model, two types of features are extracted from multi-scale feature maps, i.e. maximum pooling features and average pooling features. To assess the relative importance of the two features for the proposed method, we performed ablation experiments, to be specific, we kept one of the features and kept two of the features, respectively, and then tested the performance of these metrics. Fig. 6 illustrates the S_r and K_r values of these three metrics on DUB and DNB datasets.

Our analysis of the data in Fig. 6 leads to the following three conclusions. First, both S_r and K_r values of global maximum pooling features are larger than those of global average pooling features, which indicates that global maximum pooling feature has a greater contribution to the performance of the proposed method. Second, the proposed model achieves the best performance in all databases when we keep both features. These results demonstrate that the combination of the two features enables the model to perform better. Moreover, it can be observed that in both databases, for the version that retains only one kind of feature, the performance is still better than the existing quality metrics, even if it is not as good as the combination version. These results demonstrate the effectiveness of the extracted features for characterizing the distortions in both the blurry and deblurred images.

Table 1
Performance comparison of different quality models on the MDD, DUB, DNB and DRB databases. "TH": the traditional hand-crafted based method, "MD": the metric for image deblurring, "GR": the general-purpose image restoration quality metric, and "DL": deep learning-based method.

Method	Type	MDD		DUB		DNB		DRB	
		S_r	K_r	S_r	K_r	S_r	K_r	S_r	K_r
BIQI [12]	TH	0.755	0.679	0.479	0.364	0.379	0.301	0.284	0.218
BRISQUE [13]	TH	0.792	0.711	0.447	0.346	0.359	0.268	0.251	0.192
GMLOG [14]	TH	–	–	0.541	0.423	0.198	0.145	0.293	0.219
QAC [15]	TH	–	–	–0.158	–0.124	–0.239	–0.195	0.005	–0.009
ILNIQE [16]	TH	–	–	–0.372	–0.282	–0.127	–0.092	–0.293	–0.219
CORNIA [56]	TH	–	–	0.479	0.352	0.119	0.086	0.159	0.116
NRRB [43]	MD	0.870	0.803	0.274	0.223	0.194	0.167	0.083	0.096
MMD [42]	MD	–	–	0.599	0.479	0.439	0.332	0.212	0.156
Ref. [21]	GR	–	–	0.759	0.614	0.579	0.467	0.522	0.416
Ref. [41]	GR	–	–	0.762	0.622	0.591	0.471	0.538	0.425
DBCNN [18]	DL	0.881	0.816	0.815	0.674	0.710	0.568	0.604	0.478
VCRNet [57]	DL	0.852	0.779	0.806	0.664	0.696	0.554	0.583	0.455
Hyper [17]	DL	0.862	0.785	0.855	0.718	0.749	0.603	0.685	0.545
Proposed	MD	**0.907**	**0.847**	**0.877**	**0.748**	**0.772**	**0.627**	**0.768**	**0.617**

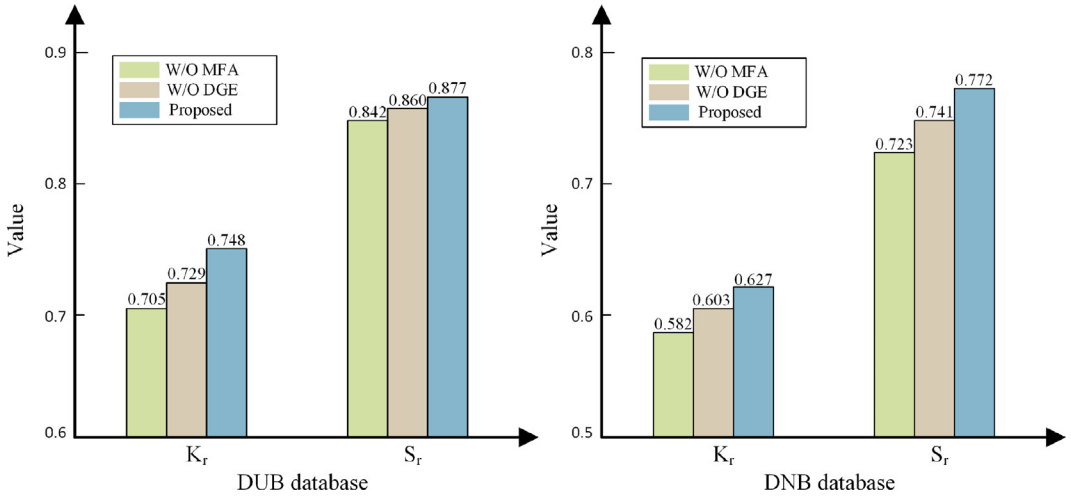


Fig. 4. Visualization of the discrepancy information of the proposed method. The first column shows the blurry images, the second column shows the corresponding feature maps of (a) and (f), the third column shows the deblurred images, the fourth column shows the corresponding feature maps of (c) and (h), the discrepancy information is shown in the last column.

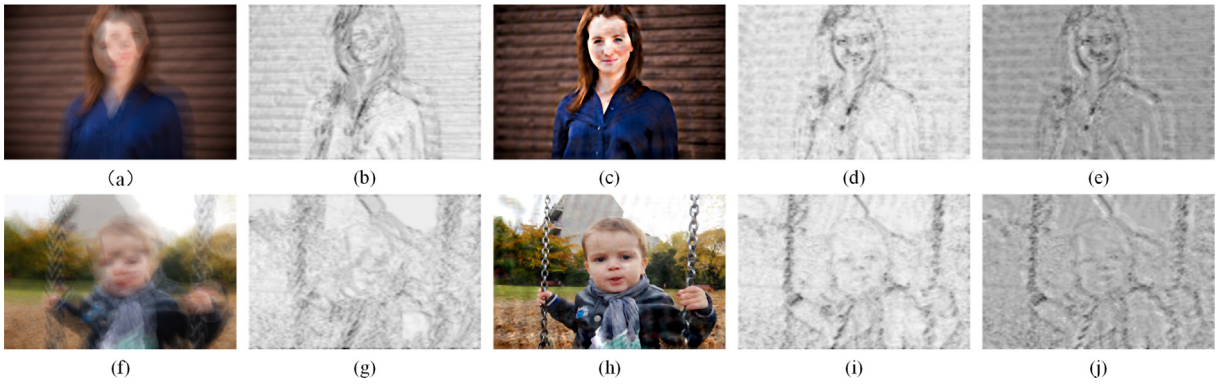


Fig. 5. Experimental results of the impact of different modules.

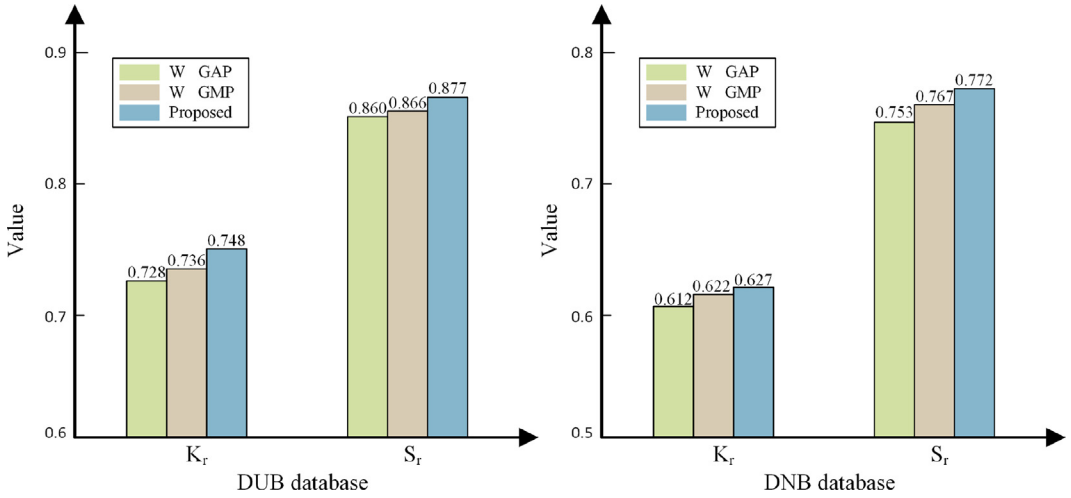


Fig. 6. Experimental results of the impact of different pooling strategies.

4.6. Impact of different image sizes

In this part, to investigate the effect of different image sizes on the performance of the proposed method, we conducted image size comparison experiments on the DUB, DNB and DRB databases. Specifically, we preprocessed the inputs by resizing them to five sizes of 224, 256, 384, 512 and 600, and then performed training and testing on the DUB, DNB and DRB databases, respectively. The results are summarized in Table 2.

It can be observed from Table 2 that the performance of the proposed method is highest when the image size is (512, 512). Therefore, we adopted (512, 512) as the image size in all experiments.

Table 2
Comparison of different image sizes on the DUB, DNB and DRB databases.

Size	DUB		DNB		DRB	
	S_r	K_r	S_r	K_r	S_r	K_r
224	0.869	0.738	0.750	0.606	0.622	0.485
256	0.868	0.734	0.763	0.615	0.692	0.547
384	0.866	0.735	0.763	0.619	0.756	0.604
512	**0.877**	**0.748**	**0.772**	**0.627**	**0.768**	**0.617**
600	0.868	0.738	0.765	0.619	0.753	0.604

4.7. Impact of different PVT configuration

In this section, we explore the impact of different versions in PVTv2 on the performance of the proposed method. Specifically, we conducted comparative experiments using versions B0, B1, B2 and B3 of PVTv2 as feature extractors on the DRB, DUB and DNB datasets respectively. Table 3 summarizes the experimental results, from which it can be seen that the PVTv2-B1-based metric, namely the proposed method, has the best performance. Therefore, we adopt PVTv2-B1 as the feature extractor in all experiments.

4.8. Cross-content experiments

As shown in Fig. 7, in the DUB and DNB databases, images can be divided into five attributes: saturated, natural, human, artificial and text. To investigate the cross-content ability of the proposed method, we further conducted cross-content experiments on the DUB and DNB databases. Specifically, we sequentially use one of the attributes as the test set and the other four attributes as the training set for a total of five experiments, with each experiment yielding the corresponding experimental results, and finally averaging the five results as the final experimental results. The results are summarized in Tables 4, 5 and the last two columns are the average of the results of the five experiments.

It can be observed from Tables 4, 5 that: 1) the cross-content experimental results of the proposed method have a significant advantage over the compared methods, especially on the DNB dataset. The proposed metric yields a S_r value that is 7 percentage points higher than the second place, which is further evidence that the proposed method has a good cross-content ability. 2) When the test sets are People and Natural, the results of the cross-content experiments of the proposed method achieve competitive results compared to previous methods. On Saturated, Manmade and Text, the proposed method achieves the best results. Even though the content distributions of the training and test datasets differ significantly, the inherent discrepancy information of BDI pair is always present, and the proposed model can learn this implicit discrepancy information. Thus, for this case, the proposed model is still able to achieve good cross-content performance.

4.9. Generalization ability

Generalization ability is important for learning-based quality assessment method and desirable in real-world scenarios. To this end, cross-database validation and User Generated Content (UGC) test are carried out in this subsection.

For cross-database validation, four quality metrics with better performance in the *Performance Evaluation* subsection are also tested for comparison. Experimental results are listed in Table 6, with the best results in bold. It can be seen from Table 6 that the proposed method is superior to the compared quality metrics by a large margin on all cross-database validation experiments. This further confirms the effectiveness and significance of discrepancy information in IDQA.

Although extensive experiments have demonstrated the superiority of the proposed method, all the above experiments were

Table 3
Comparison of different PVT configurations.

Backbone	DUB		DNB		DRB	
	S_r	K_r	S_r	K_r	S_r	K_r
PVTv2-B0	0.866	0.735	0.765	0.622	0.720	0.578
PVTv2-B1	**0.877**	**0.748**	**0.772**	**0.627**	**0.768**	**0.617**
PVTv2-B2	0.872	0.742	0.762	0.618	0.738	0.591
PVTv2-B3	0.875	**0.748**	0.755	0.609	0.763	0.606

(a) saturated (b) natural (c) people (d) man-made (e) text

Fig. 7. (a)-(e) are example images of each of the five image contents.

Table 4
Comparison of cross-content performance analysis on the DUB dataset.

Method	Saturated		People		Natural		Manmade		Text		DUB(Average)	
	S_r	K_r	S_r	K_r	S_r	K_r	S_r	K_r	S_r	K_r	S_r	K_r
DBCNN [18]	0.500	0.405	0.850	0.694	0.825	0.687	0.836	0.704	0.891	0.751	0.780	0.648
VCRNet [57]	0.525	0.399	0.861	0.709	0.800	0.636	0.839	0.705	0.905	0.779	0.786	0.646
Hyper [17]	0.577	0.450	**0.882**	**0.740**	**0.863**	**0.723**	0.869	0.736	0.878	0.746	0.814	0.679
Proposed	**0.605**	**0.484**	0.860	0.721	0.862	0.718	**0.903**	**0.792**	**0.924**	**0.799**	**0.831**	**0.703**

Table 5
Comparison of cross-content performance analysis on the DNB dataset.

Method	Saturated		People		Natural		Manmade		Text		DNB(Average)	
	S_r	K_r	S_r	K_r	S_r	K_r	S_r	K_r	S_r	K_r	S_r	K_r
DBCNN [18]	0.674	0.542	0.453	0.359	**0.741**	**0.597**	0.681	0.537	0.726	0.575	0.655	0.522
VCRNet [57]	0.743	0.590	**0.652**	**0.518**	0.636	0.491	0.647	0.509	0.571	0.464	0.650	0.514
Hyper [17]	0.688	0.562	0.354	0.275	0.699	0.557	0.660	0.510	0.608	0.491	0.602	0.479
Proposed	**0.788**	**0.646**	0.614	0.485	0.695	0.541	**0.794**	**0.653**	**0.734**	**0.598**	**0.725**	**0.585**

Table 6
Generalization ability results of the proposed model and four state-of-the-art quality models for image deblurring.

Training dataset	Testing dataset	Method	S_r	K_r
DRB	DNB	BRISQUE [13]	−0.063	−0.045
		Hyper [17]	0.646	0.511
		DBCNN [18]	0.579	0.455
		Ref. [41]	0.571	0.452
		Proposed	**0.687**	**0.553**
DRB	DUB	BRISQUE [13]	0.107	0.091
		Hyper [17]	0.737	0.599
		DBCNN [18]	0.715	0.572
		Ref. [41]	0.762	0.624
		Proposed	**0.858**	**0.722**
DNB	DRB	BRISQUE [13]	0.054	0.040
		Hyper [17]	0.544	0.431
		DBCNN [18]	0.512	0.403
		Ref. [41]	0.335	0.252
		Proposed	**0.651**	**0.517**
DUB	DRB	BRISQUE [13]	0.081	0.062
		Hyper [17]	0.533	0.418
		DBCNN [18]	0.419	0.327
		Ref. [41]	0.366	0.288
		Proposed	**0.633**	**0.498**

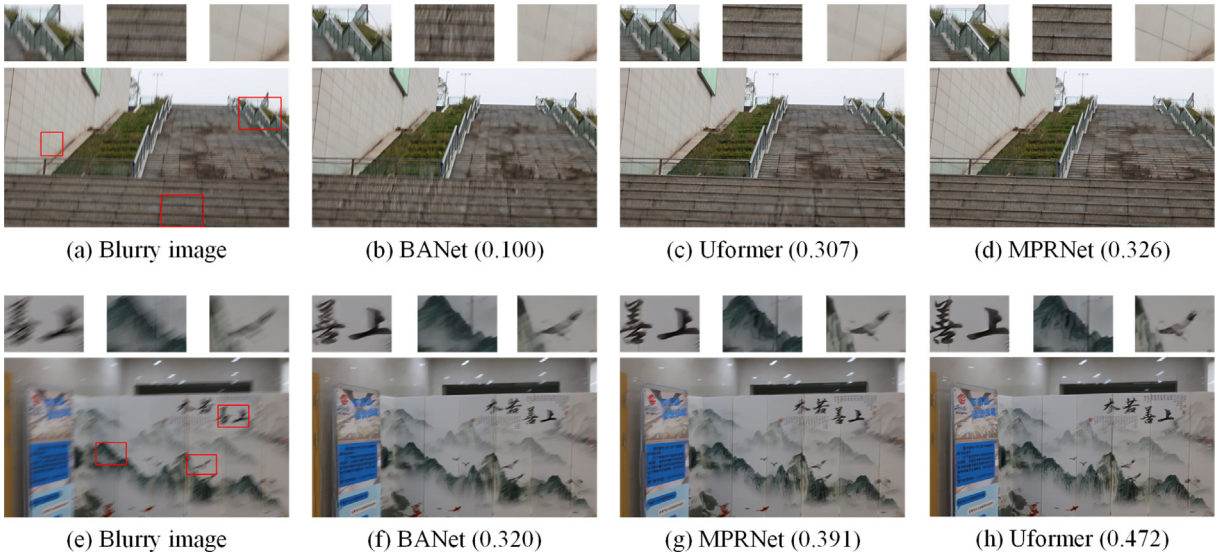


Fig. 8. The predicted results of the proposed method on the deblurring images of UGC.

conducted on standard datasets. So the question naturally arises, how does it perform when applied to real data or UGC. To solve this problem, we explore the effectiveness of the proposed method in processing real data or UGC here. Specifically, we first shot several groups of real blurry images with a Canon camera, and then processed them with three SOAT deblurring algorithms (namely, BANet [58], Uformer [59] and MPRNet [60]), producing the deblurring images. Finally, the relative quality differences of these BDI pairs are predicted by the proposed method. It is worth noting that these algorithms and blurry images were not used to build the standard datasets. Therefore, the generalization ability of the proposed method will be further verified. The predicted results of the proposed method on the deblurring images of UGC are shown in Fig. 8. Fig. 8 (b)-(d) (or (f)-(h)) are the deblurred images of (a) (or (e)) generated by the three algorithms with the predicted relative quality differences.

It can be seen from Fig. 8 that the visual quality of these blurry images has been improved to some extent after processing by these algorithms. Further observation found that in each row, the visual quality increases monotonously from the left, which is clearly demonstrated by the enlarged image blocks. Meanwhile, the proposed method accurately predicts this trend, that is, the predicted relative differences are monotonically increasing. These results fully prove that the proposed method is effective for real data or UGC and has good generalization ability.

5. Conclusion

Compared with the rapid development of image deblurring, its quality evaluation is lagging behind. In this paper, we have proposed a novel reduced-reference image deblurring quality assessment method based on multi-scale feature enhancement and aggregation. The discrepancy information of BDI pair was exploited to implicitly guide the enhancement of multi-scale features in a versatile vision-Transformer framework. Multi-scale features were integrated organically in a flexible feature aggregation module. Finally, the relative quality difference of BDI pair was obtained from a MLP-based regression network. Experimental results on four public databases have demonstrated that our method is superior to the state-of-the-arts in terms of both prediction performance and generalization ability.

In future work, we will further explore the interaction and fusion of features between different scales, and this method will also be extended as a general-purpose quality metric for a variety of different restoration scenarios. More measure forms of discrepancy information, such as mutual information and dependence, will also be explored in future work. In addition, environmental monitoring plays an important role in industrial safety warning system and ecological protection. In recent years, researchers have carried out a series of in-depth studies on $PM_{2.5}$ forecast [61], $PM_{2.5}$ monitoring [62], smoke detection [63], soot density recognition [64], and made great progress. Image quality factor may play a positive role in the image-based environmental monitoring algorithms, so how to use it to improve the performance of these algorithms is one of the directions we will focus on in the future.

CRediT authorship contribution statement

Bo Hu: Conceptualization, Methodology, Software. **Shuaijian Wang:** Data curation, Writing - original draft. **Xinbo Gao:** Visualization, Investigation, Supervision. **Leida Li:** Software, Validation. **Ji Gan:** Writing - review & editing. **Xixi Nie:** Writing - original draft.

Data availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62101084, 62036007 and 62176195, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grants KJQN202100628 and KJQN202200638, in part by the Special Project on Technological Innovation and Application Development under Grant No. cstc2020jscx-dxwtB0032, in part by Chongqing Excellent Scientist Project under Grant No. cstc2021ycjh-bgzxm0339, in part by Chongqing University of Posts and Telecommunications Ph.D. Innovative Talents Project under Grant BYJS202112, in part by the National Nature Science Foundation of China under Grant No. 62206035 , and the Surface Project of Natural Science Foundation of Chongqing under Grant No. CSTB2022NSCQ-MSX0547.

References

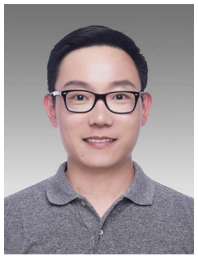
[1] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, M.-H. Yang, A comparative study for single image blind deblurring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 1701–1709.
[2] Y. Bai, H. Jia, M. Jiang, X. Liu, X. Xie, W. Gao, Single-image blind deblurring using multi-scale latent structure prior, IEEE Trans. Circuits Syst. Video Technol. 30 (7) (2019) 2033–2045.
[3] S. Cheng, R. Liu, Y. He, X. Fan, Z. Luo, Blind image deblurring via hybrid deep priors modeling, Neurocomputing 387 (2020) 334–345.
[4] H. Zhang, Y. Wu, L. Zhang, Z. Zhang, Y. Li, Image deblurring using tri-segment intensity prior, Neurocomputing 398 (2020) 265–279.
[5] W.-Z. Shao, Y.-Z. Lin, Y.-Y. Liu, L.-Q. Wang, Q. Ge, B.-K. Bao, H.-B. Li, Gradient-based discriminative modeling for blind image deblurring, Neurocomputing 413 (2020) 305–327.
[6] J. Pan, W. Ren, Z. Hu, M.-H. Yang, Learning to deblur images with exemplars, IEEE Trans. Pattern Anal. Mach. Intell. 41 (6) (2019) 1412–1425.
[7] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[8] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (1) (2010).
[9] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al., Image database tid2013: Peculiarities, results and perspectives, Signal Process.: Image Commun. 30 (2015) 57–77.
[10] D. Liang, X. Gao, W. Lu, J. Li, Deep blind image quality assessment based on multiple instance regression, Neurocomputing 431 (2021) 78–89.
[11] A. Li, J. Wu, S. Tian, L. Li, W. Dong, G. Shi, Blind image quality assessment based on progressive multi-task learning, Neurocomputing 500 (2022) 307–318.
[12] A.K. Moorthy, A.C. Bovik, A two-step framework for constructing blind image quality indices, IEEE Signal Process. Lett. 17 (5) (2010) 513–516.
[13] A. Mittal, A.K. Moorthy, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process. 21 (12) (2012) 4695–4708.
[14] W. Xue, X. Mou, L. Zhang, A.C. Bovik, X. Feng, Blind image quality assessment using joint statistics of gradient magnitude and laplacian features, IEEE Trans. Image Process. 23 (11) (2014) 4850–4862.
[15] W. Xue, L. Zhang, X. Mou, Learning without human scores for blind image quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2013, pp. 995–1002.
[16] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (2015) 2579–2591.
[17] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, Y. Zhang, Blindly assess image quality in the wild guided by a self-adaptive hyper network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3667–3676.
[18] W. Zhang, K. Ma, J. Yan, D. Deng, Z. Wang, Blind image quality assessment using a deep bilinear convolutional neural network, IEEE Trans. Circuits Syst. Video Technol. 30 (1) (2020) 36–47.

[19] M. Cheon, S.-J. Yoon, B. Kang, J. Lee, Perceptual image quality assessment with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 433–442.
[20] L. Li, T. Song, J. Wu, W. Dong, J. Qian, G. Shi, Blind image quality index for authentic distortions with local and global deep feature aggregation, IEEE Trans. Circuits Syst. Video Technol. PP (99) (2021) 1–1.
[21] B. Hu, L. Li, H. Liu, W. Lin, J. Qian, Pairwise-comparison-based rank learning for benchmarking image restoration algorithms, IEEE Trans. Multimedia 21 (8) (2019) 2042–2056.
[22] J. Cai, W. Zuo, L. Zhang, Dark and bright channel prior embedded network for dynamic scene deblurring, IEEE Trans. Image Process. 29 (2020) 6885–6897.
[23] D. Krishnan, R. Fergus, Fast image deconvolution using hyper-laplacian priors, Adv. Neural Inform. Process. Syst. 22 (2009) 1033–1041.
[24] Q. Shan, J. Jia, A. Agarwala, High-quality motion deblurring from a single image, ACM Trans. Graphics (tog) 27 (3) (2008) 1–10.
[25] A. Levin, Y. Weiss, F. Durand, W.T. Freeman, Understanding and evaluating blind deconvolution algorithms, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1964–1971.
[26] L. Xu, J. Jia, Two-phase kernel estimation for robust motion deblurring, in: European Conference on Computer Vision, Springer, 2010, pp. 157–170.
[27] T. Michaeli, M. Irani, Blind deblurring using internal patch recurrence, in: European Conference on Computer Vision, Springer, 2014, pp. 783–798.
[28] D. Krishnan, T. Tay, R. Fergus, Blind deconvolution using a normalized sparsity measure, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 233–240.
[29] L. Xu, S. Zheng, J. Jia, Unnatural l0 sparse representation for natural image deblurring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2013, pp. 1107–1114.
[30] W. Zuo, D. Ren, S. Gu, L. Lin, L. Zhang, Discriminative learning of iteration-wise priors for blind deconvolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015, pp. 3232–3240.
[31] X. Chen, X. He, J. Yang, Q. Wu, An effective document image deblurring algorithm, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 369–376.
[32] Z. Hu, S. Cho, J. Wang, M.-H. Yang, Deblurring low-light images with light streaks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014, pp. 3382–3389.
[33] J. Pan, D. Sun, H. Pfister, M.-H. Yang, Blind image deblurring using dark channel prior, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 1628–1636.
[34] Y. Yan, W. Ren, Y. Guo, R. Wang, X. Cao, Image deblurring via extreme channels prior, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 4003–4011.
[35] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
[36] J. Cai, S. Gu, L. Zhang, Learning a deep single image contrast enhancer from multi-exposure images, IEEE Trans. Image Process. 27 (4) (2018) 2049–2062.
[37] S. Nah, T. Hyun Kim, K. Mu Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 3883–3891.
[38] H. Zhang, Y. Dai, H. Li, P. Koniusz, Deep stacked hierarchical multi-patch network for image deblurring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5978–5986.
[39] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, J. Matas, Deblurgan: Blind motion deblurring using conditional adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8183–8192.
[40] H. Ma, S. Liu, Q. Liao, J. Zhang, J.-H. Xue, Defocus image deblurring network with defocus map estimation as auxiliary task, IEEE Trans. Image Process. 31 (2022) 216–226.
[41] L. Li, B. Hu, Y. Huang, H. Zhu, Reduced-reference perceptual discrepancy learning for image restoration quality assessment, in: CAAI International Conference on Artificial Intelligence, Springer, 2021, pp. 359–370.
[42] Y. Liu, J. Wang, S. Cho, A. Finkelstein, S. Rusinkiewicz, A no-reference metric for evaluating the quality of motion deblurring, ACM Trans. Graphics 32 (6) (2013) 1–12.
[43] B. Hu, L. Li, J. Qian, Perceptual quality evaluation for motion deblurring, IET Comput. Vision 12 (6) (2018) 796–805.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inform. Process. Syst. 30 (2017) 5998–6008.
[45] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929.
[47] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.
[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[49] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.
[50] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, Computat. Visual Media 8 (3) (2022) 415–424.
[51] L. Li, Y. Zhou, K. Gu, Y. Yang, Y. Fang, Blind realistic blur assessment based on discrepancy learning, IEEE Trans. Circuits Syst. Video Technol. 30 (11) (2019) 3859–3869.
[52] A. Paszke, S. Gross, C. al, Automatic differentiation in pytorch, International Conference on Neural Information Processing Systems Workshop.
[53] K. Gu, G. Zhai, X. Yang, W. Zhang, Hybrid no-reference quality metric for singly and multiply distorted images, IEEE Trans. Broadcast. 60 (3) (2014) 555–567.
[54] K. Gu, X. Xu, J. Qiao, Q. Jiang, W. Lin, D. Thalmann, Learning a unified blind image quality metric via on-line and off-line big training instances, IEEE Trans. Big Data 6 (4) (2019) 780–791.
[55] Y. Liu, K. Gu, Y. Zhang, X. Li, G. Zhai, D. Zhao, W. Gao, Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception, IEEE Trans. Circuits Syst. Video Technol. 30 (4) (2019) 929–943.
[56] P. Ye, J. Kumar, L. Kang, D. Doermann, Unsupervised feature learning framework for no-reference image quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1098–1105.
[57] Z. Pan, F. Yuan, J. Lei, Y. Fang, X. Shao, S. Kwong, Vcrnet: Visual compensation restoration network for no-reference image quality assessment, IEEE Trans. Image Process. 31 (2022) 1613–1627.
[58] F.J. Tsai, Y.T. Peng, Y.Y. Lin, C.C. Tsai, C.W. Lin, Banet: A blur-aware attention network for dynamic scene deblurring, IEEE Trans. Image Process. 31 (2022) 6789–6799.
[59] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, H. Li, Uformer: A general u-shaped transformer for image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17683–17693.
[60] S.W. Zamir, A. Arora, S. Khan, M. Hayat, F.S. Khan, M.H. Yang, L. Shao, Multi-stage progressive image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14821–14831.
[61] K. Gu, Z. Xia, J. Qiao, Stacked selective ensemble for pm forecast, IEEE Trans. Instrum. Meas. 69 (3) (2019) 660–671.
[62] K. Gu, H. Liu, Z. Xia, J. Qiao, W. Lin, D. Thalmann, Pm monitoring: Use information abundance measurement and wide and deep learning, IEEE Trans. Neural Networks Learn. Syst. 32 (10) (2021) 4278–4290.
[63] K. Gu, Z. Xia, J. Qiao, W. Lin, Deep dual-channel neural network for image-based smoke detection, IEEE Trans. Multimedia 22 (2) (2019) 311–323.
[64] K. Gu, Y. Zhang, J. Qiao, Ensemble meta-learning for few-shot soot density recognition, IEEE Trans. Industr. Inf. 17 (3) (2020) 2261–2270.

Bo Hu received the B.S. and Ph.D. degrees from China University of Mining and Technology, Xuzhou, China, in 2014 and 2020, respectively. Currently, he is currently with the School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, China. His research interests include multimedia quality assessment, image aesthetic assessment and perceptual image restoration.

Shuaijian Wang received the B.S. degree from Suzhou University of Science and Technology, Suzhou, China, in 2020. He is currently pursuing a Master's degree from Chongqing University of Posts and Telecommunications. His research interests include multimedia quality assessment and image aesthetic assessment.

Xinbo Gao received the B.Eng., M.Sc. and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a post-doctoral research fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is a Cheung Kong Professor of Ministry of Education of P. R. China, a Professor of Pattern Recognition and Intelligent System of Xidian University. Since 2020, he has been also a Professor of Computer Science and Technology of Chongqing University of Posts and Telecommunications. His current research interests include image processing, computer vision, multimedia analysis, machine learning and pattern recognition. He has published seven books and around 300 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/CoChair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology, a Fellow of the Chinese Institute of Electronics, a Fellow of the China Computer Federation, and Fellow of the Chinese Association for Artificial Intelligence.



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, 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) Lab, 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 an SPC for IJCAI from 2019 to 2021, the Session Chair for ICMR in 2019 and PCM in 2015, and a TPC Member for CVPR in 2021, ICCV in 2021, AAAI from 2019 to 2021, ACM MM from 2019 to 2020, ACM MM-Asia in 2019, ACII in 2019, and PCM in 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.

Ji Gan received the B.E. degree in software engineering from Huazhong University of Science and Technology, Wuhan, China, in 2015, and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2021. He is currently with the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His current research interests include pattern recognition, human-computer interaction, handwriting recognition and generation.

Xixi Nie received the B.Eng. and M.Eng degrees in computer science and technology from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2015 and 2020, respectively. She is currently pursuing the Ph.D degree in computer science and technology of Chongqing University of Posts and Telecommunications, Chongqing, China. Her research interests include computer vision and image processing.