Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07874-2

S.I.: EFFICIENT ARTIFICIAL INTELLIGENT ALGORITHMS FOR MEDICAL IMAGE ANALYSIS BASED ON HIGH-PERFORMANCE COMPUTING

CSPP-IQA: a multi-scale spatial pyramid pooling-based approach for blind image quality assessment

Jingjing Chen[1,2] · Feng Qin[3] · Fangfang Lu[3,4] · Lingling Guo[5] · Chao Li[6] · Ke Yan[7] · Xiaokang Zhou[8,9]

Received: 24 February 2022 / Accepted: 21 September 2022

Abstract
The traditional image quality assessment (IQA) methods are usually based on convolutional neural networks (CNNs). For these IQA methods using CNNs, limited by the feature size of the fully connected layer, the input image needs be tailored to a pre-defined size, which usually results in destroying the original structure and content of the input image and thus reduces the accuracy of the quality assessment. In this paper, a blind image quality assessment method (named CSPP-IQA), which is based on multi-scale spatial pyramid pooling, is proposed. CSPP-IQA allows inputting the original image when assessing the image quality without any image adjustment. Moreover, by facilitating the convolutional block attention module and image understanding module, CSPP-IQA achieved better accuracy, generalization and efficiency than traditional IQA methods. The result of experiments running on real-scene IQA datasets in this study verified the effectiveness and efficiency of CSPP-IQA.

Keywords Image quality assessment · Spatial pyramid pooling · Semantic feature extraction · Attention mechanism · Quality score

Jingjing Chen and Feng Qin contributed equally to this article.

✉ Fangfang Lu
lufangfang@shiep.edu.cn

✉ Lingling Guo
lguo@zjut.edu.cn

[1] Zhejiang University City College, Hangzhou, China

[2] School of Economics, Fudan University, Shanghai, China

[3] College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, China

[4] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China

[5] College of Chemical Engineering, Zhejiang University of Technology, Hangzhou, China

[6] Zhijiang College, Zhejiang University of Technology, Shaoxing, China

[7] Department of the Built Environment, National University of Singapore, Singapore, Singapore

[8] Faculty of Data Science, Shiga University, Hikone 522-8522, Japan

[9] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

Published online: 17 October 2022

1 Introduction

Due to the wide adoption of smart phones, the generation of digital images has shown explosive growth. Quality assessment is an essential for enhancing the quality of visual contents sent to the terminal users in a visual communication systems (VCSs). [1]. Video and images also play an important role in maintaining social stability and prosperity. For example, in the medical field, with the sudden outbreak of the novel coronavirus pneumonia (COVID-19), images play an extremely important role in fighting the epidemic. Experts across the country can conduct online medical analysis and diagnosis through images [2–7]. It is an essential to assess the quality of real-scene images to optimize the parameters and performance of the image-oriented systems, so as to ensure the quality of visual content delivered to the terminal users.

Methods for IQA could be grouped to objective ones and subjective ones. Objective methods are trained by using subjective assessment data. It can automatically predict content quality and is suitable for real-time scenarios. Human visual systems (HVS) are the ultimate receiver of visual signals in most of VCSs; hence, HVS-based subjective methods are reliable and accurate. However, it is time-consuming and usually cannot be directly integrated to other applications as an optimized factor.

Generally, objective IQA methods can be grouped to: (1) No-Reference (NR), which is also called Blind IQA (BIQA); (2) Reduced-Reference (RR) and (3) Full-Reference (FR), which is usually used as a metric to assess algorithms for image processing. RR-IQA methods are usually integrated into the other systems to optimize the algorithm performance. Usually, there is no possibility to acquire a reference image without any distortions. NR-IQA is the promising method in practice.

Recently, various kinds of IQA methods are proposed for the natural image quality assessment [8–17]. These features extracted by manual operations are not enough to perfectly achieve the IQA. Besides, the low-level features are not enough for representing the complex distortion in practice. Although these exist a number of NR-IQA methods [18–33] for distortion achieved, most of existing deep neural networks are not specially proposed for IQA tasks; therefore, these methods could only work with the global features rather than the local features. However, the image distortion mostly occurs in the local areas. Moreover, local distortions are sensitive for HVS; hence, both the global and local quality while designing an IQA algorithm must be taken into consideration. Moreover, these IQA methods usually introduce CNN to extract semantic features of the original images (as illustrated in Fig. 1b). Constrained by the image size, the original image must be tailored prior to inputting to the convolutional network. As illustrated in Fig. 1a, the overexposed frame missed due to the operation of cropping on the original image; the latter image is stretched, which indirectly introduces distortions and destroys the original content of the original image. In consequence, the caused distortion deviation weakens the performance of these methods.

In this paper, a novel method for IQA, which is based on multi-scale spatial pyramid pooling, is proposed. In practices, such as capturing images with mobile phones, the proposed method could provide reliable and efficient image quality assessment for terminal users. The contributions in this work are as below:

- We proposed a convolutional spatial pyramidal pooling structure to address the issue of size limitation of semantic features, and the structure is capable of assessing the image quality by directly inputting the original image without the operations tailoring image.

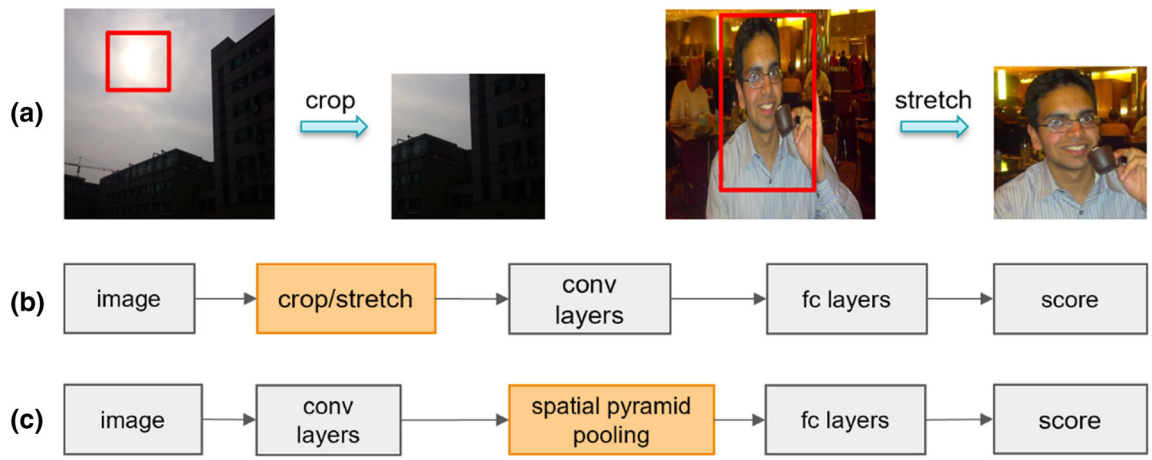

Fig. 1 Comparison of the workflow between traditional CNN-based methods and the proposed method

- The ResNet50 is adopted as the backbone network to integrate the multi-scales semantic information of the image to efficiently capture the local distortion. In the proposed method, the Convolutional Block Attention Module is integrated to highlight important features while suppressing unimportant ones.
- In the proposed method, SmoothL1 loss is introduced in the process of quality scoring to address the issue that the original model is not smooth near the zero point.

The remainder of this paper is organized as follows. Section 2 introduces the related work. The proposed model is introduced in Sect. 3. Section 4 presents the experiments of the proposed method. We conclude the paper and discuss the future directions in Sect. 5.

2 Related work

2.1 NR-IQA-based synthetic distortion

NR-IQA can be grouped to synthetic distortion based on NR-IQA and authentic distortion based on NR-IQA. The synthetic distortion based on NR-IQA can be grouped to: (1) natural scene statistical-based models (NSS), (2) manual feature extraction-based models and (3) deep learning-based models. There are a number of NSS-based NR-IQA methods such as BRISQUE [8], NIQE [9], BIQI [10], DIIVINE [11], BLIINDS [12], ShearletIQM [13] and SPNSS [14]. Moreover, CORINA [15] is NR-IQA-based model. ILNIQE is a BIQA-based method [16]. Xu et al. [17] proposed a High Order Statistics Aggregation hybrid-based model. Rajevenceltha [34] proposed a rotation-invariant and efficient NR-IQA method, which extracts features using the modified local binary patterns and statistical methods and then employs support vector regression to measure image quality. Azam et al. concluded the fusion quality assessments metric of fused images with different imaging modalities [35].

The deep learning-based IQA methods adopt deep neural networks to extract the visual features of an image and then calculate the functional expression of the distorted image to get its quality score. Kang et al [18] introduced CNN to construct an IQA method. Kim [19] used an image block-based method to increase the amount of training data. Bosse et al. [20] performed random sampling operations to image dataset without the operation of normalization, so these global features (e.g., luminance, etc.) can be taken into considerations.

2.2 NR-IQA-based authentic distortion

The authentic distortion is usually randomly and mixed distributed in an image; therefore, traditional synthetic distortion-based NR-IQA methods cannot perfectly achieve the quality score of authentic distorted images. Ghadiyaram et al. [21] proposed a feature-map-based image quality assessment method. Bianco et al. [22] introduced a pre-trained CNN to build a function mapping CNN features to subjective quality scores. Li et al. [23] proposed a NR-IQA based method, which is only valid for authentic blurred. Zeng et al. [24] proposed a probabilistic model to predict the distribution of five different quality scores instead of only one score. HyperIQA [26] is proposed to predict the image quality of with an adaptation to scenarios. Zhang et al. [31] and Sun et al. [32] proposed a BIQA for in-the-wild images via hierarchical feature fusion and iterative mixed database training.

To address the cross-distortion-scenario challenges, [25] proposed a novel model consisting of two networks, in which one network is used for synthetic distortion and other network is used for authentic distortion. The change of image content and types of image distortions hinders the definition of image quality while assessing the blind quality of distorted images [36, 37].

3 The proposed model

As shown in Fig. 2, the proposed model consists of (1) multi-scale semantic feature extraction, (2) quality prediction module and (3) adaptive content understanding. Firstly, four different scales of semantic features are extracted from the input image by ResNet50, and then four different scales of semantic feature vectors are generated, respectively, by spatial pyramid pooling. Then, these semantic feature vectors are integrated as the input of the quality prediction module. By introducing attention mechanism, the 4th scale semantic features are transferred to high-level semantic features, which will then be fed to the adaptive content understanding module. The image content understanding module is designed to generate the weights and biases for the quality prediction module. The quality prediction module is designed to calculate these weights and biases with a multi-scale semantic feature vector to get the image's quality score.

3.1 Multi-scale semantic feature extraction

The ResNet50 [38] is adopted as the backbone network. The fully connected layer and average pooling layer are removed from the classical ResNet50 model and the output

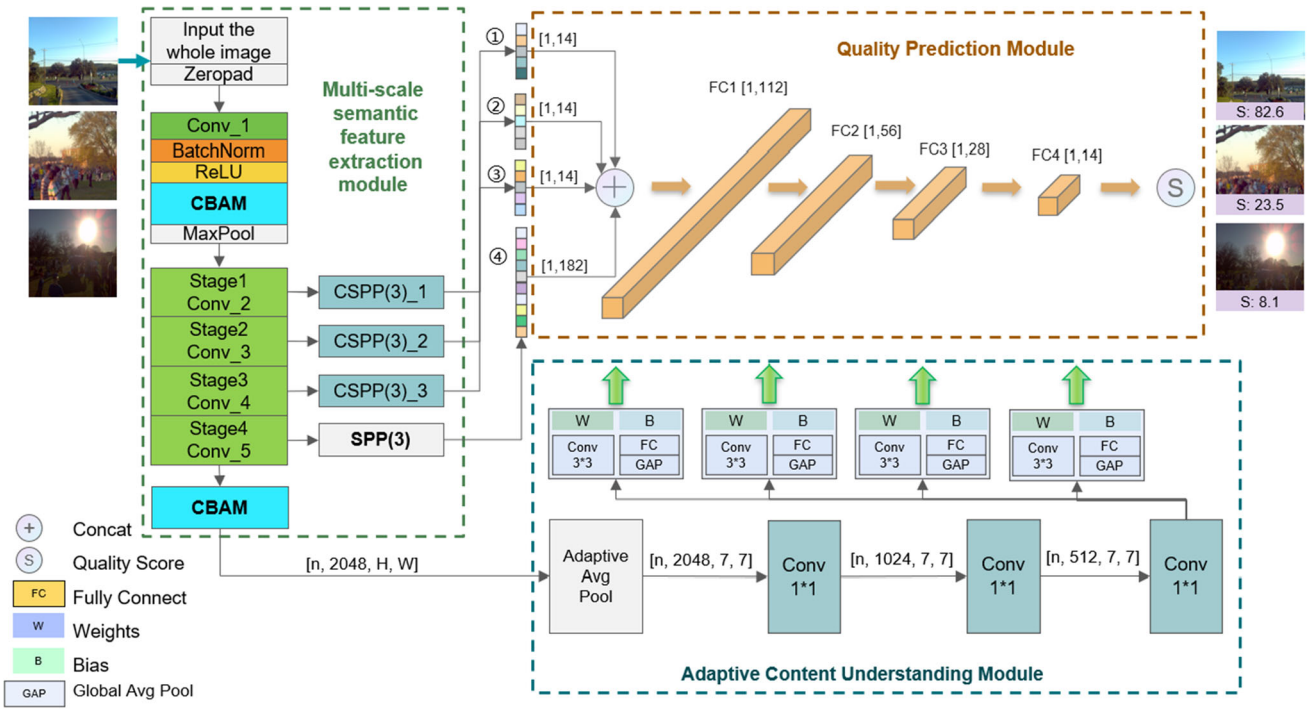

Fig. 2 The structure of the proposed image quality assessment model

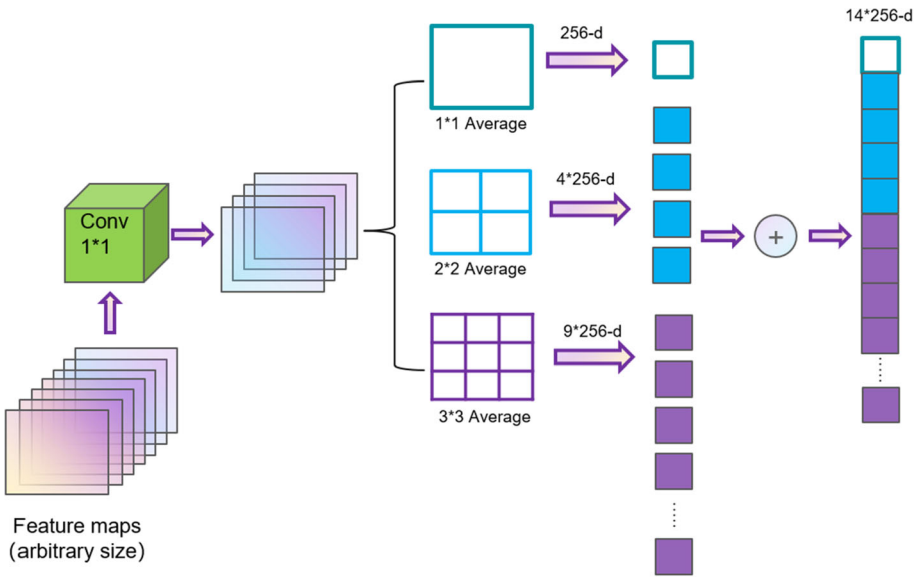

Fig. 3 The internal structure of the CSPP

of this model is designated to the high-level semantic feature stream.

Constrained by the size of feature of the fully connected layer in the traditional IQA methods, a test image is randomly cropped into several blocks in the same size, and then these image blocks are subsequently fed into a network for semantic feature extraction. As presented in the SPPnet [39], spatial pyramid pooling structure could improve the model performance and speed up the model training; hence, in the proposed method, the convolutional spatial pyramidal pooling (CSPP) is adopted as the pooling structure in the process of semantic feature extraction.

As shown in Fig. 3, the CSPP is introduced for reducing dimensionalities by spatial pyramidal pooling and 1×1 convolution. There is no constraint to the size of the input feature map, and the size of the output channels is set to 256. A three-level spatial pyramid structure is used for the operation of pooling. By merging the results from the three sources, the final feature size of $14 * 256$ is obtained. So that, it is transformed to a one-dimensional matrix prior to

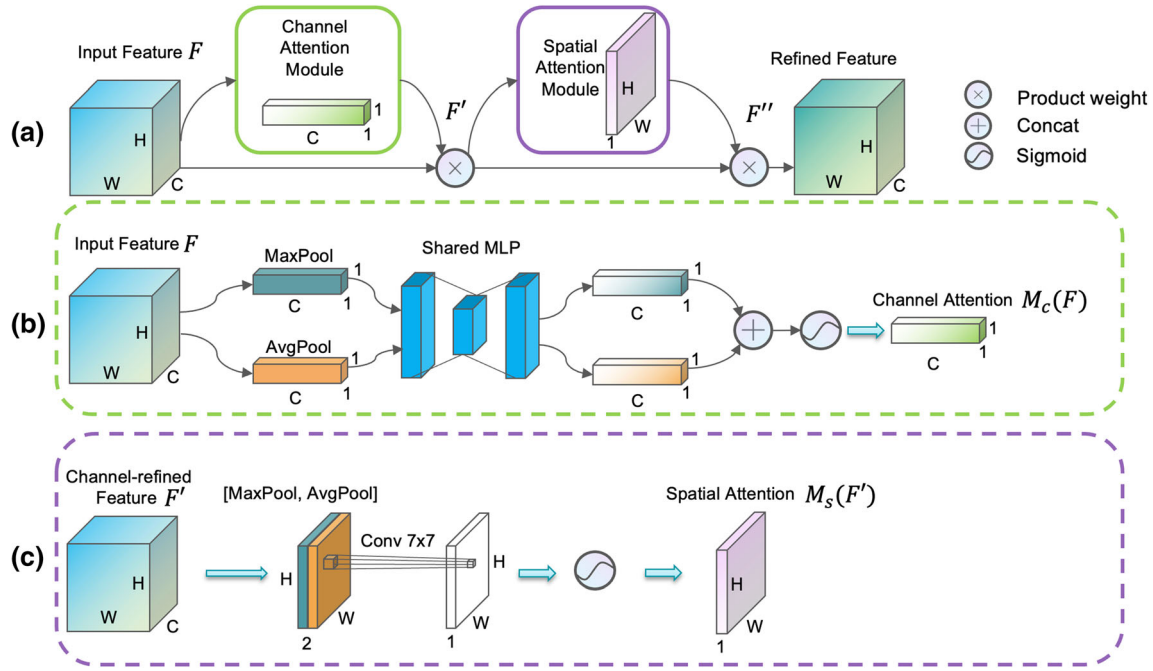

Fig. 4 The internal structure of the convolutional block attention module

being passed to the fully connected layer. Whatever the size of the input images, and with the operation of spatial pyramid pooling, the size will be finally set to $1 * 3584$, which is the size of input feature in the fully connected layer.

The CBAM [40] is introduced to highlight the target features of channel and spatial axis as shown in Fig. 4a. Besides, the channel attention module and the spatial attention module in turn, whose functions are adopted to learn the content in the channel axis and the position in the spatial axis, respectively. Given an intermediate feature map $F \in R^{C\times H\times W}$ as input (C for channel, H for height, W for width), CBAM calculates a 1D channel attention map $M_c \in R^{C\times 1\times 1}$ and a 2D spatial attention map $M_s \in R^{1\times H\times W}$ in turn, as shown in Fig. 4. The whole attention process can be summarized as follows:

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

where $\otimes$ denotes the element-wise multiplication, $M_c(\cdot)$ represents the calculation of channel attention module, and $M_s(\cdot)$ represents the calculation of spatial attention module. The execution process of CBAM are as follows: first, the input F is multiplied by channel attention module to get the result F'; then, F' is used as the input of spatial attention module, and the refined feature F'' is calculated by the spatial attention module.

As shown in Fig. 4b, the channel attention module receives the average-pooled and max-pooled features and then feeds these features to a weight-sharing Multilayer Perceptron (MLP). The element-wise summation combining the output feature vectors is carried out. The core idea of channel attention module is to make up for the deficiency of channel attention. The channel attention module is formularized as:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{3}$$

where σ represents sigmoid activation function, MLP represents shared full connection layer, $AvgPool$ represents average pooling, $MaxPool$ represents maximum pooling, and '+' represents element-wise addition.

The structure of spatial attention module is shown in Fig. 4c, a feature map is obtained through max pooling and average pooling, then they are spliced into a 2D feature map and then sent to the standard 7×7 convolution for parameter learning, and a 1D weight feature map is obtained. The spatial attention module is formularized as:

$$M_s(F') = \sigma\left(conv^{7\times 7}([AvgPool(F'); MaxPool(F')])\right) \tag{4}$$

where σ is the sigmoid activation function and $conv^{7\times 7}$ is the convolution kernel of 7×7.

The full workflow of the feature extraction module is as shown in Fig. 2. The original image is inputted to ResNet50 and then passed through the first convolutional layer and CBAM. Three strings of semantic feature streams in different scales are generated from convolutional layers. Next, these three feature streams are inputted into CSPP(3) to generate 3 semantic feature vectors. The high-level

semantic feature streams generated in Stage 4 are further handled by SPP to generate the 4th semantic feature vector. By calculating these four semantic feature vectors, the $C(I)$, which is the input of the quality prediction module, is obtained.

3.2 Adaptive content understanding

The quality scoring in traditional deep learning-based IQA methods is defined in Eq. (5).

$$s = P(I, \theta) \tag{5}$$

where P is a function mapping the original image I to the quality score s. θ denotes the weight of the network. When the training process completes, the weight parameter θ is obtained for all test images.

With regard to the characteristics of HVS, we decode the content of the image and then customize different rules according to the content, and thus the calculation of the quality score is defined as follows:

$$S = P(I, \theta_I) \tag{6}$$

where the parameter θ_I is determined by the content of the test image.

The calculation of parameter θ_I is defined as follows:

$$\theta_I = U(C(I), \lambda) \tag{7}$$

where U is a function mapping the high-level semantic features $C(I)$ to network parameters θ_I, λ is the parameter in the adaptive content understanding module, and $C(I)$ is from the input image.

The module is designed to learn the content of the image and generate the weights and biases for the fully connected layer. It consists of an adaptive average pooling layer, three 1×1 convolutional layers and four weight generation branches. The weights of the fully connected layer are initiated by the operation of convolution.

An adaptive pooling operation is carried out prior to inputting the high-level semantic features into the adaptive content understanding module to ensure that the size of feature maps is subjected to the requirement. The maximum or average pooling is formularized as:

$$S_{\text{out}} = \lceil (S_{\text{in}} + 2*\text{padding} - S_{\text{kernel}})/\text{stride} \rceil + 1 \tag{8}$$

where S_{out} denotes the size of output, S_{in} denotes the size of input, and *padding* denotes the fill size. The operation of padding during the pooling process is used to maintain the boundary information of the feature map. S_{kernel} denotes the kernel size and *stride* denotes the step size.

Given the input and output dimensions, the adaptive pooling operation is defined as follows:

$$\text{stride} = \lfloor S_{\text{in}}/S_{\text{out}} \rfloor \tag{9}$$

$$S_{\text{kernel}} = S_{\text{in}} - (S_{\text{out}} - 1) \cdot \text{stride} \tag{10}$$

3.3 Quality prediction process

The calculation of image quality scoring is formularized in Eq. (11).

$$s = P(f_I, U(C(I), \lambda)) \tag{11}$$

f_I denotes the multi-scale semantic feature stream extracted from ResNet50. $C(I)$ denotes the high-level semantic features, which will be fed into the adaptive content understanding module to generates weights and biases for the image quality prediction module. These weights and biases will be calculated with f_I using fully connection to get the quality score.

4 Experiment

4.1 Dataset introduction

The datasets in this experiment includes Live-Challenge [41], KonIQ-10k [42], BID [43], SPAQ [44] and FLIVE [45]. The LIVE Challenge was developed by University of Texas at Austin. It contains 1162 authentic distortion images with more than 350,000 collected subjective scores. The values of these scores are from 3.42 to 92.43. The KonIQ-10k was developed by the University of Konstanz. It contains 10,073 authentic distorted images with 1,459 annotators and 1.2 million subjective data. BID contains 586 images with authentic blur distortion (such as complex motion blur, etc.). SPAQ contains 11,125 images covering a wide range of scene categories such as landscape, human, animal, etc. FLIVE is currently the largest database for IQA, and it contains over 40,000 authentic distorted images.

4.2 Training method

The training was carried out on a Windows 10 computer equipped with a GPU 2080Ti, and we took Pytorch library to develop the experimental programs. A single-size training method proposed in [39] was adopted for model training. To enhance the generalization of the model, each image is randomly sampled and horizontally flipped to get 25 blocks in the size 224×224 pixels. In each round of training, 80% of the dataset are randomly selected as the training set and 20% of the dataset are used as the testing set. In the test, SmoothL1Loss is adopted as the loss function and is formularized as follows:

$$smooth_{L1}(t) = \begin{cases} 0.5t^2 & |t| < 1 \\ |t| - 0.5 & otherwise \end{cases} \tag{12}$$

In Eq. (12), SmoothL1Loss will be assigned as L2Loss when the absolute value of t is less than 1. Then smaller loss is obtained because L2Loss squares the error, which is beneficial for model convergence. Otherwise, SmoothL1Loss is a translation of L1Loss. L1Loss is more insensitive to outliers compared to L2Loss, and the magnitude of the gradient is controllable. SmoothL1Loss is treated as the combination of L2Loss and L1Loss. Therefore, SmoothL1Loss is adopted as loss function. The t in Eq. (12) is defined as below:

$$t = P(f_I, U(C(I), \lambda)) - q \tag{13}$$

The $P(f_I, U(C(I), \lambda))$ in Eq. (13) denotes the predicted score, and q denotes the subjective score. Adam optimizer is adopted for performance optimization. The weight attenuation is set to 5×10^{-4}; meanwhile, the learning rate is set to 2×10^{-5}.

4.3 Assessment metrics

In order to validate the performance of the proposed method, Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC) are taken as assessment metrics to compute the correlation between subjective and objective scores in this experiment. PLCC is adopted for assessing the prediction accuracy of the model, and SROCC is adopted to access the prediction monotonicity.

The quality scores achieved by different IQA methods are in different ranges, so a mapping function to regress these quality scores into a common space is defined as follows:

$$Q(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}} \right) + \beta_4 x + \beta_5 \tag{14}$$

where x denotes the input score and $(\beta_1, \cdots, \beta_5)$ is the set of parameters to be fitted.

4.4 Experimental results

4.4.1 SROCC and PLCC Performance

The experiment is carried out on a group of datasets: Live-Challenge, KonIQ-10k dataset, BID, SPAQ and FLIVE dataset. The experiment compares the performance of SROCC and PLCC with the traditional manual feature extraction IQA methods [8, 16–18], synthetic distortion IQA methods [20, 21] and authentic distortion IQA methods [22, 24–27]. As shown in Table 1. The proposed method outperformed in the comparison, indicating that directly inputting the whole image into the network is helpful for reflecting the overall image quality. Because cropping/stretching the original image usually destroys the original structure and content of the image, it weakens the performance of quality scoring.

4.4.2 Comparison of MOS prediction

The MOS prediction was performed using BIECON, FRISQUEE, PQR, DBCNN, HyperIQA and CSPP-IQA on Live-Challenge dataset, respectively. As shown in Fig. 5, the result is illustrated in a group of scatter plots, where the horizontal coordinates are the true MOS values and the

Table 1 The performance comparison of SROCC and PLCC running on different datasets

Database	Live-Challenge		KonIQ-10k		BID		SPAQ		FLIVE
Criterion	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
BRISQUE	0.608	0.629	0.665	0.681	0.574	0.540	0.802	0.806	0.320	0.356
CORNIA	0.629	0.671	0.683	0.713	0.612	0.663	0.709	0.725	0.311	0.349
ILNIQE	0.432	0.508	0.507	0.523	0.516	0.554	0.713	0.721	0.322	0.335
HOSA	0.640	0.678	0.671	0.694	0.721	0.736	0.721	0.733	0.338	0.354
BIECON	0.595	0.613	0.618	0.651	0.439	0.576	0.702	0.722	0.301	0.336
WaDIQaM	0.671	0.680	0.797	0.805	0.725	0.742	0.837	0.845	0.452	0.433
FRIQUEE	0.682	0.705	0.808	0.811	0.728	0.739	0.819	0.830	0.434	0.428
SFA	0.812	0.833	0.685	0.872	0.826	0.840	0.906	0.907	0.542	0.626
PQR	0.857	0.882	0.880	0.884	0.830	0.852	0.902	0.913	0.547	0.635
DBCNN	0.851	0.869	0.875	0.884	0.845	0.859	0.910	0.913	0.554	**0.652**
HyperIQA	0.859	0.882	0.906	0.917	0.869	0.878	**0.916**	0.919	0.535	0.623
CSPP-IQA	**0.882**	**0.898**	**0.912**	**0.921**	**0.875**	**0.891**	**0.916**	**0.922**	**0.556**	0.649

Bold number represents the best performance

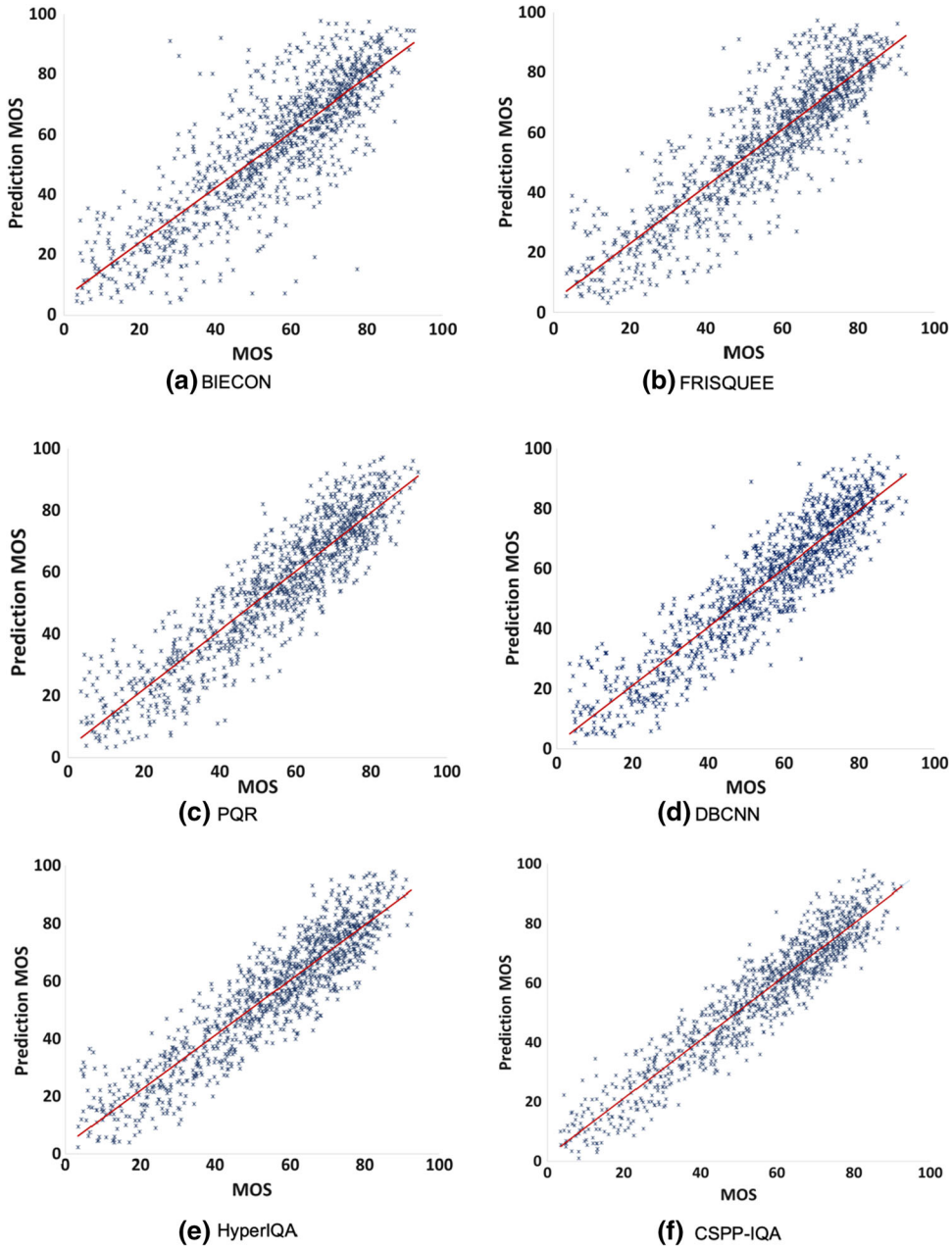

Fig. 5 Scatter plots of predicted MOS values of six different IQA methods on the Live-challenge

Table 2 The comparison of generalization for proposed model and others

Train	Test	PQR	DBCNN	HyperIQA	Proposed
Live-challenge	KonIQ-10k	0.757	0.754	0.772	**0.788**
Live-challenge	BID	0.714	0.762	0.756	**0.763**
KonIQ-10k	Live-challenge	0.770	0.755	0.785	**0.797**
KonIQ-10k	BID	0.755	0.816	**0.819**	0.808

Bold number represents the best performance

vertical coordinates are MOS values predicted by the objective IQA methods. The scatter plot can intuitively show the relationship between the two sets of data. The better the performance of the algorithm, the more the scatter distribution in the plot is clustered around the red fitting line. The scatter distribution of CSPP-IQA is more concentrated and regular than other methods, which also shows that the predicted MOS of CSPP-IQA is more consistent with subjective score.

Table 3 The performance comparison of training process (100 rounds) running on the dataset KonIQ-10k

Algorithm	Hours/h	SROCC	PLCC
BIECON	73	0.595	0.613
SFA	84	0.685	0.872
DBCNN	112	0.851	0.869
HyperIQA	127	**0.906**	**0.917**
CSPP-IQA	**26**	0.905	**0.917**

Bold number represents the best performance

4.4.3 Analysis of model generalization

To evaluate the generalization of the proposed model, we took different image datasets for training and testing. The model is trained on a dataset, but tested on a different dataset. The proposed method outperforms the competition with PQR, DBCNN and HyperIQA (as shown in Table 2). Moreover, by integrating the CBAM into the proposed model, the achieved quality score of the model is close to human subjective perception.

4.4.4 Comparison of training time

We performed a 100-round experiment to evaluate the training time of the proposed model and others. The original image is inputted directly in the training phase on the KonIQ-10k dataset without cropping. As shown in the Table 3, compared with HyperIQA, under the premise of ensuring accuracy, the training speed of CSPP-IQA is improved by 4.88 times. Besides, the proposed model requires less computational resources.

Table 4 The performance comparison of different pooling modules integrated to SPP running on the dataset Live-Challenge

Pooling	SROCC	PLCC
Max	0.875	0.887
Avg	**0.882**	**0.898**

Bold number represents the best performance

4.4.5 Experimental analysis on SPP structure

By introducing the SPP, there is no need for stretching and cropping the original image, which could avoid causing extra distortions to the original image. Figure 6 shows the result of CSPP-IQA adopted different levels of SPP (i.e., SPP-3, SPP-4 and SPP-5). SPP-3 denotes three-level SPP structure and so on. The result indicates that the three-level SPP outperformed the performance comparison. That is because more features extracted by higher-level SPP could result in certain redundant features, which could affect the final quality assessment. Meanwhile, the introduction of

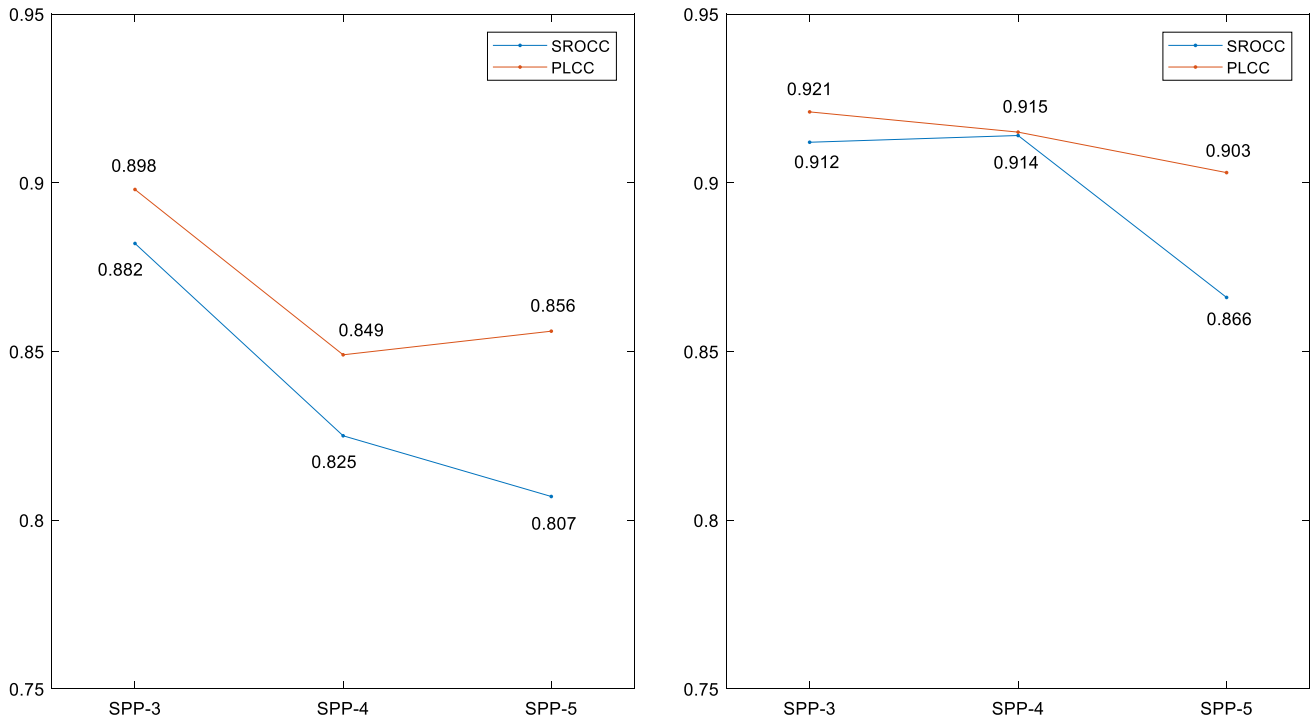

Fig. 6 Performance comparison of SPP structure in different levels: **a** Result based on the dataset Live-Challenge; **b** Result based on the dataset KonIQ-10k

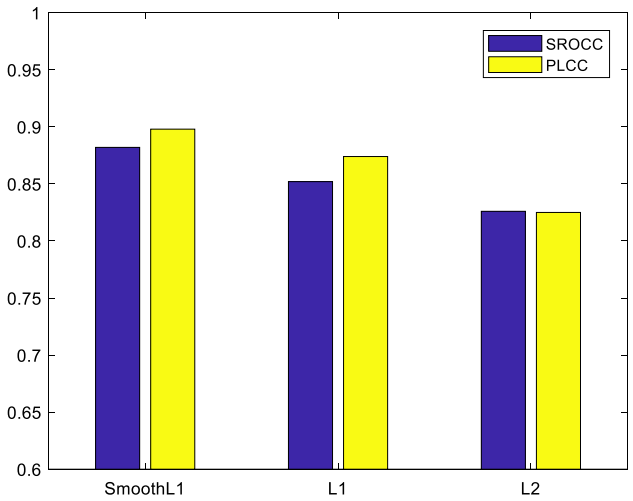

Fig. 7 The comparison of loss functions adopted in this experiment on the dataset Live-challenge

higher-level SPP structure could hinder the converge process of the model.

In the experiment based on the dataset Live-Challenge, the maximum pooling and average pooling are introduced in the three-layer SPP structure respectively. As shown in Table 4, the comparative result indicates that average pooling achieved better performance as it retains the overall characteristics of the image. In this sense, the main function of average pooling is more consistent than maximum pooling, which aims to preserving texture features and reducing the impact of irrelative information. The idea of the proposed method is to learn the global and local semantic information by fusing multi-scale features, and thus average pooling is adopted in SPP structure.

4.4.6 Comparison of SROCC performance for loss functions

Figure 7 shows the comparative results of different loss functions running on Live-Challenge dataset. Compared with L1 / L2 loss function, the SmoothL1 loss function achieved the best performance thanks to the integration of L1 and L2 loss functions.

Table 5 Comparison of prediction MOS between the proposed method and others

Image number	(a)	(b)	(c)	(d)	(e)
MOS	49.73	82.38	85.33	91.95	75.5
BIECON	28.79	89.36	75.99	83.78	59.2
DBCNN	32.43	70.33	91.49	81.74	56.32
HyperIQA	43.5	86.45	**86.77**	88.46	71.86
CSPP-IQA	**45.62**	**79.37**	89.28	**88.55**	**73.49**

Bold number represents the best performance

Table 6 The result of ablation experiment

	Live-challenge		KonIQ-10k	
Components	SROCC	PLCC	SROCC	PLCC
HyperIQA	0.859	0.882	0.906	0.917
HyperIQA + SPP	0.877	0.889	0.911	0.918
HyperIQA + CBAM	0.864	0.881	0.897	0.918
CSPP-IQA	**0.882**	**0.898**	**0.912**	**0.921**

Bold number represents the best performance

4.4.7 Comparison of quality scoring accuracy

As shown in Fig. 8, we took 5 randomly selected images from the SPAQ dataset for the experiment, and the predicted MOS of four IQA methods (BIECON, DBCNN, HyperIQA and CSPP-IQA) for these images are listed in Table 5. CSPP-IQA achieved better performance in terms of consistency with MOS while assessing the quality of inputting images of different contents (e. g. animals, landscapes, portraits, buildings, etc.). As shown in Table 5, DBCNN achieved poor scores on Fig. 8e because the traditional IQA model that predicts the score directly without

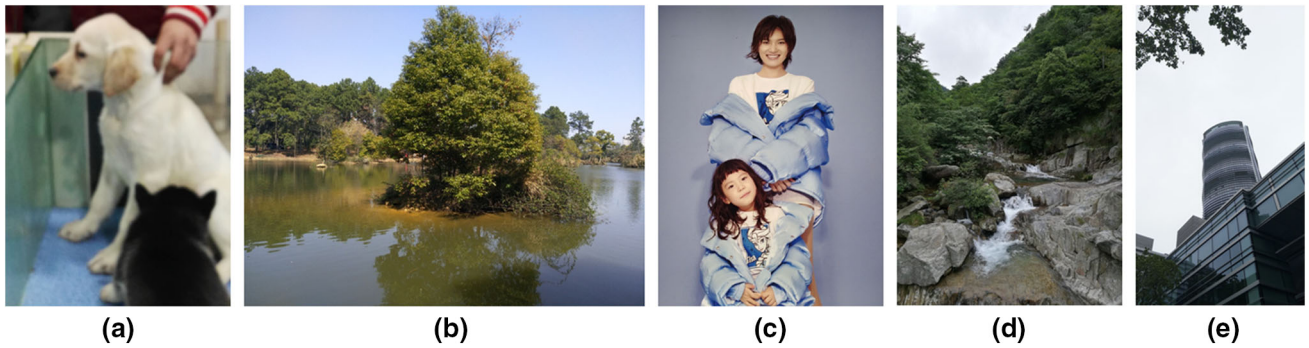

Fig. 8 A group of randomly selected images from the SPAQ dataset

content understanding is prone to mistakenly treat flat regions (e. g. the sky) as distorted images [24].

4.4.8 Ablation Analysis

As shown in Table 6, ablation experiment was carried out based on the datasets Live-Challenge and KonIQ-10k to verify the effectiveness of each component. With Live-Challenge dataset, the introduction of either the SPP structure or CBAM could contribute to the improvement of SROCC by 1.5% and by 0.7% and PLCC by 2.5% and by 0.9%, respectively. With the dataset KonIQ-10k, the introduction of SPP structure could contribute an improvement of SROCC by 0.4% and PLCC by 0.5%, and the introduction of CBAM mechanism could contribute an improvement of PLCC by 0.6%.

5 Conclusion

In this study, we proposed a blind image quality assessment method (named CSPP-IQA) for distorted images. It introduces the spatial pyramid pooling, as well as the attention mechanism to successfully address the issue caused by the constraint of the image size in the fully connected layer. In the proposed method, by introducing the CBAM and adaptive content understanding module, the assessment accuracy is further improved, and it archives stronger generalization and less time cost compared with these existing IQA methods. Besides, CSPP-IQA requires less training time and computational resources and could be widely used in many real-time applications.

Because the human eye is easy to be attracted by the salient areas of the image when observing an image, the quality of these areas has great impact on the overall quality of the image. In the future, the visual attention could be taken into the basis of current study to further improve the prediction performance for the IQA tasks.

Acknowledgements This work was supported by Ministry of Education humanities social sciences research project (20YJC760101); China Postdoctoral Science Foundation (2020T130102ZX); Natural Science Foundation of Zhejiang Province (LQ20E080021, LQ21H190004); and the Educational Commission of Zhejiang Province of China (Y202147553).

Data availability The datasets generated during and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interests The authors declare that they have no interest conflict.

References

1. Zhai G, Min X (2020) Perceptual image quality assessment: a survey. Sci China Inf Sci 63(11):1–52
2. Rajevenceltha J (2022) Gaidhane V (2022) An efficient approach for no-reference image quality assessment based on statistical texture and structural features. Eng Sci Technol Int J 30:101039
3. Zhao L, Li K, Pu B, Chen J, Li S, Liao X (2022) An ultrasound standard plane detection model of fetal head based on multi-task learning and hybrid knowledge graph. Futur Gener Comput Syst 135:234–243
4. Wu X, Tan G, Zhu N, Chen Z, Li K (2021) CacheTrack-YOLO: Real-time detection and tracking for thyroid nodules and surrounding tissues in ultrasound videos. IEEE J Biomed Health Inform 25(10):3812–3823
5. Fang Y, Zhu H, Zeng Y, Ma K, Wang Z (2020) Perceptual quality assessment of smartphone photography. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3677–3686.
6. Liu X, Yang L, Chen J, Yu S, Li K (2022) Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation. Biomed Signal Process Control. https://doi.org/10.1016/j.bspc.2021.103165
7. Zhou X, Liang W, Wang K, Yang L (2021) Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations. IEEE Trans Comput Soc Syst 8(1):171–178
8. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
9. Mittal A, Soundararajan R, Bovik AC (2012) Making a "completely blind" image quality analyzer. IEEE Signal Process Lett 20(3):209–212
10. Moorthy AK, Bovik AC (2010) A two-step framework for constructing blind image quality indices. IEEE Signal Process Lett 17(5):513–516
11. Moorthy AK, Bovik AC (2011) Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Trans Image Process 20(12):3350–3364
12. Saad MA, Bovik AC, Charrier C (2010) A DCT statistics-based blind image quality index. IEEE Signal Process Lett 17(6):583–586
13. Mahmoudpour S, Kim M (2016) No-reference image quality assessment in complex-shearlet domain. SIViP 10(8):1465–1472
14. Lu F, Zhao Q, Yang G (2015) A no-reference image quality assessment approach based on steerable pyramid decomposition using natural scene statistics. Neural Comput Appl 26(1):77–90
15. Ye P, Kumar J, Kang L, Doermann D (2012) Unsupervised feature learning framework for no-reference image quality assessment. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 1098–1105
16. Zhang L, Zhang L, Bovik AC (2015) A feature-enriched completely blind image quality evaluator. IEEE Trans Image Process 24(8):2579–2591
17. Xu J, Ye P, Li Q, Du H, Liu Y, Doermann D (2016) Blind image quality assessment based on high order statistics aggregation. IEEE Trans Image Process 25(9):4444–4457
18. Kang L, Ye P, Li Y, Doermann D (2014) Convolutional neural networks for no-reference image quality assessment. In: IEEE conference on computer vision and pattern recognition, pp 1733–1740
19. Kim J, Lee S (2016) Fully deep blind image quality predictor. IEEE J Sel Topics Signal Process 11(1):206–220

20. Bosse S, Maniry D, Müller KR, Wiegand T, Samek W (2017) Deep neural networks for no-reference and full-reference image quality assessment. IEEE Trans Image Process 27(1):206–219
21. Ghadiyaram D, Bovik AC (2015) Scene statistic of authentically distorted images in perceptually relevant color spaces for blind image quality assessment. In 2015 IEEE International conference on image processing (ICIP), pp 3851–3855
22. Bianco S, Celona L, Napoletano P, Schettini R (2017) On the use of deep learning for blind image quality assessment. SIViP 12(2):355–362
23. Li D, Jiang T, Lin W, Jiang M (2019) Which has better visual quality: The clear blue sky or a blurry animal? IEEE Trans Multimedia 21(5):1221–1234
24. Zeng H, Zhang L, Bovik AC (2017) A probabilistic quality representation approach to deep blind image quality prediction. arXiv preprint arXiv:1708.08190
25. Zhang W, Ma K, Yan J, Deng D, Wang Z (2020) Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans Circuits Syst Video Technol 30(1):36–47
26. Su S, Yan Q, Zhu Y, Zhu Y, Zhang C, Ge X, Sun J, Zhang Y (2020) Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: IEEE/CVF Conference on computer vision and pattern recognition, pp 3667–3676
27. Zhou X, Li Y, Liang W (2021) CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. IEEE/ACM Trans Comput Biol Bioinf 18(3):912–921
28. Zhou X, Liang W, Wang K, Huang R, Jin Q (2021) Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data. IEEE Trans Emerg Top Comput 9(1):246–257
29. Sun W, Min X, Zhai G, Ma S (2021) Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. arXiv preprint arXiv:2105.14550
30. Zhai G, Sun W, Min X, Zhou J (2021) Perceptual quality assessment of low-light image enhancement. ACM Trans Multimedia Comput Commun Appl (TOMM) 17(4):1–24
31. Zhang W, Ma K, Zhai G, Yang X (2020) Learning to blindly assess image quality in the laboratory and wild. In: International conference on image processing (ICIP), IEEE, pp 111–115
32. Zhou X, Liang W, Wang K, Shimizu S (2019) Multi-modality behavioral influence analysis for personalized recommendations in health social media environment. IEEE Trans Comput Soc Syst 6(5):888–897
33. Jiang Q, Xu J, Zhou W, Min X, Zhai G (2022) Deep decomposition and bilinear pooling network for blind night-time image quality evaluation. arXiv preprint arXiv:2205.05880
34. Pu B, Li K, Li S, Zhu N (2021) Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT. IEEE Trans Indus Inf 17(11):7771–7780
35. Zhou X, Xu X, Liang W, Zeng Z, Yan Z (2021) Deep-learning-enhanced multitarget detection for end-edge-cloud surveillance in smart IoT. IEEE Internet Things J 8(16):12588–12596
36 Azam M et al (2022) A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2022.105253
37. Chen J, Yang N, Zhou M, Zhang Z, Yang X (2022) A configurable deep learning framework for medical image analysis. Neural Comput Appl 34(10):7375–7392
38. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778
39. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916
40. Woo S, Park J, Lee, JY, Kweon S (2018) CBAM: Convolutional block attention module. In: Proc. ECCV, pp 3–19
41. Ghadiyaram D, Bovik AC (2015) Massive online crowdsourced study of subjective and objective picture quality. IEEE Trans Image Process 25(1):372–387
42. Hosu V, Lin H, Sziranyi T, Saupe D (2020) KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Trans Image Process 29:4041–4056
43. Ciancio A, Targino da Costa ALN, da Silva EAB, Said A, Samadani R, Obrador P (2010) No-reference blur assessment of digital pictures based on multifeature classifiers. IEEE Trans Image Process 20(1):64–75
44 Chen J, Li K, Zhang Z, Li K, Yu PS (2022) A survey on applications of artificial intelligence in fighting against COVID-19. ACM Comput Surv. https://doi.org/10.1145/3465398
45. Ying Z, Niu H, Gupta P, Mahajan D, Ghadiyaram D, and Bovik AC (2020) From patches to pictures (PaQ-2-PiQ): mapping the perceptual space of picture quality. In: IEEE Conference on computer vision and pattern recognition, pp 3575–3585.