# Blind Quality Assessment for in-the-Wild Images via Hierarchical Feature Fusion Strategy

Wei Sun, Huiyu Duan, Xiongkuo Min, Li Chen, and Guangtao Zhai
*Institue of Image Communication and Information Processing, Shanghai Jiao Tong University, China*

2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) | 978-1-6654-6901-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/BMSB55706.2022.9828590

***Abstract*—Image quality assessment (IQA) is very important for both end-users and service-providers since a high-quality image can significantly improve the user's quality of experience (QoE). Most existing blind image quality assessment (BIQA) models were developed for synthetically distorted images, however, they perform poorly on in-the-wild images, which are widely existed in various practical applications. In this paper, inspired by perceptual visual quality being affected by both low-level visual features and high-level semantic information, we propose an effective BIQA model for in-the-wild images by considering rich features extracted from the convolution neural network (CNN). Specifically, we propose a staircase structure to hierarchically integrate the features from intermediate layers of the CNN into the quality-aware feature representation, which enables the model to make full use of visual information from low-level to high-level and are more suitable for the in-the-wild IQA task. Experimental results show that the proposed model outperforms other state-of-the-art BIQA models on six in-the-wild IQA databases by a large margin. Moreover, the proposed model is flexible and can be replaced with popular CNN models to meet the various needs of practical applications.**

***Index Terms*—objective evaluation techniques, quality of experience, blind image quality assessment, in-the-wild images, convolution neural network, feature fusion**

## I. Introduction

With the advent of the mobile era, billions of images are generated in various social media applications every day, most of which are captured by amateur users in various in-the-wild environments. Different from pictures shot by photographers, the quality of ordinary-user-generated images is often degraded by distortions like under/over exposure, low visibility, motion blur, ghosting, etc. A high-quality image can improve the viewer's Quality of Experience (QoE) and also benefit lots of computer vision algorithms. With massive images being generated every moment, there is an urgent need to develop a quality assessment model for in-the-wild images.

Due to the lack of pristine images, only blind image quality assessment (BIQA) models are qualified for evaluating the quality of in-the-wild images. Previous BIQA models [1]–[4] mainly focus on images with synthetic distortions such as JPEG compression, Gaussian blur, etc. However, the difference between images with authentic and synthetic distortions is quite large. For example, synthetic distortions (e.g. JPEG compression, White noise) are usually global uniform since these distortions are introduced to the whole images uniformly, while authentic distortions can be not only global uniform (e.g. low illumination) but also non-uniform (e.g. object moving, ghosting). We illustrate some examples of synthetically and authentically distorted images in Fig. 1. Though existing BIQA models [1]–[5] have achieved remarkable performance on synthetically distorted images, there is still a great challenge to assess the quality of in-the-wild images.

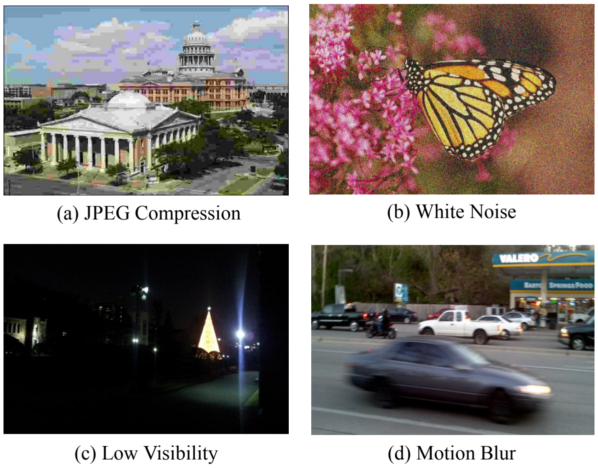
(a) JPEG Compression
(b) White Noise
(c) Low Visibility
(d) Motion Blur

Fig. 1. Images with synthetic distortions and authentic distortions. (a) and (b) are synthetically distorted images, (c) and (d) are authentically distorted images.

Existing BIQA models generally follow such routine: 1) extracting quality-aware features, and 2) mapping these features into quality scores via a regression model. Commonly used quality-aware features include natural scene statistics (NSS) features [1], [2], free energy features [3], [4], corners/textures [5], etc., while commonly used regression models include support vector regression, random forest regression, etc. For example, DIIVINE [1] first identifies the distortion type of the image, and then conducts distortion-specific IQA using NSS features extracted in the wavelet domain. BRISQUE [2] uses the scene statistics of local luminance coefficients to quantify possible losses of "naturalness". Gu *et al.* [3] develop a NR free energy based robust metric (NFERM) using three groups of features: features extracted by the free energy model, image structure and gradient features, and NSS features of the mean subtracted contrast normalized coefficients. Min *et al.* [5] integrate the similarities of corners and local binary patterns between distorted images and the corresponding pseudo images as the quality score, where the pseudo images can be in multiple distortion levels.

Recently, deep learning technologies show great ability to solve various visual signal problems. Latest BIQA models

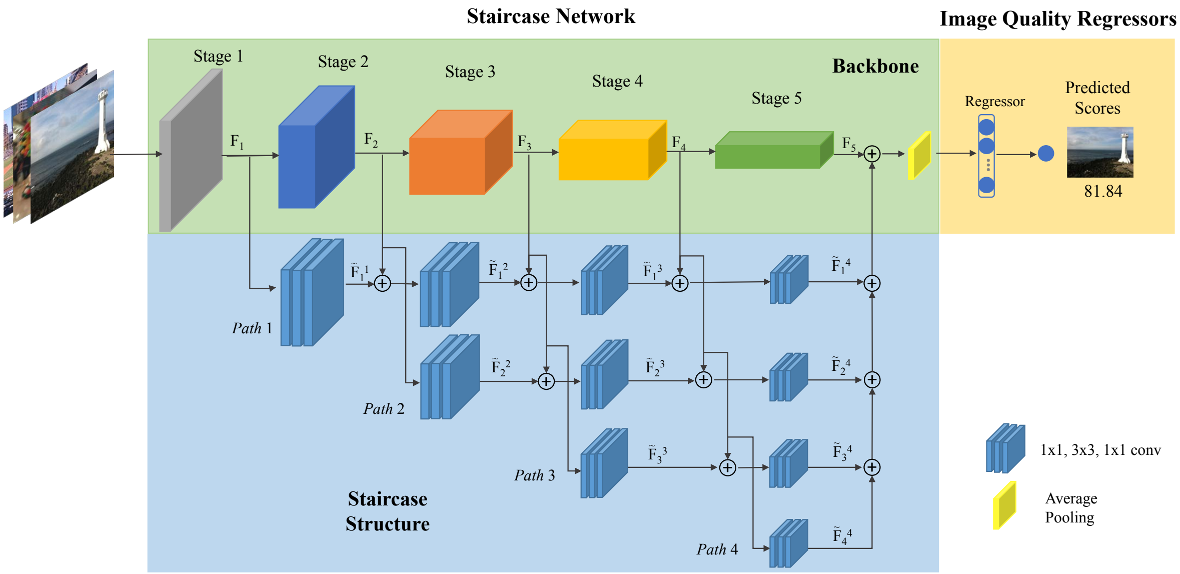

Fig. 2. The network architecture of the proposed model. The proposed model includes the staircase network for quality-aware feature extraction and image quality regressors for mapping the quality-aware features to the quality score.

adopt the deep learning based architecture, which utilizes a convolutional neural network (CNN) to extract quality-aware features of distorted images and then regresses them to quality scores with a fully connected network. This kind of architecture allows it to be trained in an end-to-end manner and has been dominant in the BIQA fields. For example, Kang *et al.* [6] use a shallow CNN model consisting of one convolutional layer and two fully connected layers to estimate the quality of small patches, and then the image level quality score is averaged by the predicted patch scores of the corresponding image. Bosse *et al.* [7] further deepen the CNN model by jointly learning the quality and weight of patches, where the weight is the relative importance of the patch quality to the global quality estimation. Zhang *et al.* [8] propose to merge features extracted from two kinds of CNN models into a better representation by bilinearly pooling, where two CNN are respectively pretrained on the distortion type and level classification task and the image classification task. Su *et al.* [9] develop a self-adaptive hyper network to aggregate local distortion features and global semantic features.

Compared with handcrafted features, features extracted by CNN are more powerful and more suitable for in-the-wild images. However, the commonly used backbone networks such as VGG [10], Resnet [11], etc. are designed for the image classification task, where extracted features are at the semantic level, but perceptual visual quality is also affected by low-level visual features. So, it is not an optimal option to directly use a popular CNN architecture as the backbone of BIQA tasks.

Therefore, in this paper, we propose a novel BIQA model for in-the-wild images by considering the rich features extracted from the CNN model. First, we propose a staircase structure to hierarchically incorporate the features from intermediate layers of the CNN into the final feature representation, which makes the model learn more effective quality-aware features. Previous studies [12], [13] indicate that the features extracted from different stages of a CNN model represent different visual information. For example, the features extracted from bottom convolution layers correspond to low-level information such as edges and corners, while the features extracted from top convolution layers are at the semantic level. Through fusing the features from intermediate layers, the CNN model can fully utilize the visual information from low-level to high-level and learn the better feature representations for quality evaluation. Then, we utilize the fully connected layer to map the quality-aware features into the quality score. Experimental results show that the proposed model achieves the best performance on six in-the-wild IQA databases, and also achieves an excellent performance in the cross-database evaluation, which demonstrate the effectiveness and generalizability of the proposed model. What's more, the proposed staircase structure is flexible and can be replaced with popular CNN models to meet the various needs of practical applications.

The rest of this paper is organized as follows. Section II describes the proposed model in detail. In Section III, we give the results of the proposed model and compare the performance with other popular BIQA models on six benchmark IQA databases. Section IV gives the concluding remarks.

## II. Proposed Method

In this section, we describe the proposed method in detail. A diagram of the network structure is illustrated in Fig. 2, which includes two parts, a staircase network for more effective and powerful feature representation and image quality regressors for mapping the quality-aware features to quality score spaces.

### A. *Staircase Network for Feature Extraction*

Many successful CNN models such as VGG [10], ResNet [11] follow the same design paradigm, which gradually reduce the dimension of feature maps and increase the number of feature maps at the same time. This kind of architecture allows the CNN model to learn features from low-level to high-level as the number of network layers deepen, and achieves promising performance in many computer vision tasks such as image recognition [10], [11], [14], object detection [15], image segmentation [16], etc. However, the perceived quality of images is affected by both the low-level visual features and high-level semantic information [17] [18] [19]. It is not optimal to directly use the popular CNN model as the feature extraction module due to the loss of low-level features. Here, we propose the staircase structure to hierarchically integrate the features extracted from intermediate layers, so the model can make full use of features extracted from low-level to high-level visual information.

Generally, the popular CNN architectures can be divided into several stages according to the dimension of feature maps. In each stage, there are several convolutional layers in series to deepen the network. Assume that there are $N_s$ stages, and $F_i$ is the feature map extracted from the $i$-th stage, where $i \in [1, 2, ..., N_s]$. Since we want to integrate the features extracted from each stage into the final feature representation, a simple method is to fuse the feature maps by element-wise addition operators, i.e.

$$F = \sum_{i=1}^{N_s} F_i. \tag{1}$$

However, there are two problems if we directly use Eq. (1) as the feature fusion method.

First, it is observed that the number of channels and the dimension of feature maps in each stage are not the same. Generally speaking, the dimension of the feature maps at the current stage is half that of the previous stage while the number of channels is twice that of the previous stage. So, it is impossible to add the feature maps from different stages directly. In order to make the number of channels and the dimension of feature maps at different stages the same, we introduce a bottleneck structure consisting of three convolution operations to downscale the dimension and increase the channels. Specifically, we first reduce the channels of feature map $F_i$ to a quarter through the $1\times1$ convolution layer to decrease the computation complexities of the whole procedures. Then we utilize the $3\times3$ convolution layer with a stride of 2 to reduce the resolution of $F_i$ to half. Finally, $F_i$ is passed through the $1\times1$ convolution layer to increase the number of channels for eight times. After that, the feature map $\widetilde{F}_i$ can be represented as:

$$\widetilde{F}_i = W_{1\times1} W_{3\times3} W_{1\times1} F_i = W F_i, \tag{2}$$

where $W_{1\times1}$ and $W_{3\times3}$ are the weight matrices of the $1\times1$ convolution layer and the $3\times3$ convolution layer respectively, and $W$ is the product of $W_{1\times1}$ and $W_{3\times3}$. Then, we can directly add feature maps from different stages:

$$F = \sum_{i=1}^{N_s-1} ( \prod_{j=i}^{N_s-1} W_{ij}) F_i + F_{N_s}, \tag{3}$$

where $W_{ij}$ means the $j$-th weight matrix for the feature map $F_i$.

Second, we notice that adding the features from lower layers to the final stage directly will cause the whole network difficult to train. For example, if we use a short connection (include downscaling and channel maps adding operators) to add the features in Stage 1 to the features in Stage $N_s$, it will make the backward propagated gradients tend to pass through the short connection while ignoring the backbone network during training. As a result, it is hard to train the backbone network to extract deep semantic features. Therefore, we propose to hierarchically merge feature maps from different stages to avoid this problem. More specifically, for the feature map $F_1$ from Stage 1, we first downscale its resolution and increase its channels by two convolution layers to obtain $\widetilde{F}_1$. Then we merge $\widetilde{F}_1$ with the feature map $F_2$ via element-wise addition and derive $\widetilde{F}_1^2$. For $\widetilde{F}_1^2$, we continue to reduce its resolution and increase the channels, and then add it with the feature maps $F_3$ to derive $\widetilde{F}_1^3$. The same operation is repeatedly implemented until fusing $\widetilde{F}_1^{N_s-2}$ with the feature map $F_{N_s-1}$ to derive $\widetilde{F}_1^{N_s-1}$, and $\widetilde{F}_1^{N_s-1}$ is the final feature map extracted from Stage 1 that needs to be merged into final feature maps. We then do similar operations for the feature maps from other stages. These procedures can be formulated as:

$$\widetilde{F}_i^{j+1} = W_{ij} \widetilde{F}_i^{j} + F_{j+1}, \tag{4}$$

where $i \in [1, 2, .., N_s - 2]$, $j \in [i, ..., N_s - 2]$, and $\widetilde{F}_i^i = F_i$.

Finally, the quality-aware features extracted by the staircase network are represented as:

$$F = \sum_{i=1}^{N_s-2} \widetilde{F}_i^{N_s-1} + W_{N_s-1,n} F_{N_s-1} + F_{N_s}. \tag{5}$$

### B. *Image Quality Regressor*

After extracting quality-aware features by the staircase network, we need to map these features to the quality scores with a regression model. We first apply the global average pooling (GAP) on the extracted feature maps to produce a feature vector with a dimension of $P \times 1$, where $P$ is the number of final feature maps. Then two Fully Connected (FC) layers are used as the regression model to predict the image quality. The two FC layers consist of 128 and 1 neurons respectively. Finally, we can train the staircase network and image quality regressor in an end-to-end training manner. The Euclidean distance is used as the loss function:

$$L = \| q_{predict} - q_{label} \|^2, \tag{6}$$

where $q_{predict}$ is the quality score predicted by the proposed model and $q_{label}$ is the ground-truth quality score derived from subjective experiments.

TABLE I
PERFORMANCE OF EIGHT STATE-OF-THE-ART METHODS AND THE PROPOSED MODEL ON SIX IN-THE-WILD IQA DATABASES. THE BEST PERFORMING MODELS IN IQA CATEGORIES ARE HIGHLIGHTED IN EACH COLUMN.

| Database | CLIVE | | BID | | KonIQ-10k | | SPQA | | FLIVE | | FLIVE Patch | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Criterion | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC | SRCC | PLCC |
| NIQE [20] | 0.4536 | 0.4676 | 0.4772 | 0.4713 | 0.5260 | 0.4745 | 0.6973 | 0.685 | 0.1048 | 0.1409 | 0.3211 | 0.2826 |
| BRISQUE [2] | 0.6005 | 0.6211 | 0.5736 | 0.5401 | 0.715 | 0.7016 | 0.8021 | 0.8056 | 0.3201 | 0.3561 | 0.5372 | 0.5843 |
| BMPRI [5] | 0.4868 | 0.5229 | 0.5154 | 0.4583 | 0.6577 | 0.6546 | 0.7501 | 0.7544 | 0.2737 | 0.3146 | 0.5839 | 0.6142 |
| CNNIQA [6] | 0.6269 | 0.6008 | 0.6163 | 0.6144 | 0.6852 | 0.6837 | 0.7959 | 0.7988 | 0.3059 | 0.2850 | 0.6005 | 0.5379 |
| WaDIQaM-NR [7] | 0.6916 | 0.7304 | 0.6526 | 0.6359 | 0.7294 | 0.7538 | 0.8397 | 0.8449 | 0.4346 | 0.4303 | 0.6995 | 0.7197 |
| SFA [21] | 0.8037 | 0.8213 | 0.8202 | 0.8253 | 0.8882 | 0.8966 | 0.9057 | 0.9069 | 0.5415 | 0.626 | 0.7175 | 0.7501 |
| DB-CNN [8] | 0.8443 | 0.8624 | 0.8450 | 0.8590 | 0.8780 | 0.8867 | 0.9099 | 0.9133 | 0.5537 | 0.6518 | 0.7509 | 0.7869 |
| HyperIQA [9] | 0.8546 | 0.8709 | 0.8544 | 0.8585 | 0.9075 | 0.9205 | 0.9155 | 0.9188 | 0.5354 | 0.6228 | 0.7489 | 0.7850 |
| Proposed | **0.8624** | **0.8821** | **0.8724** | **0.8830** | **0.9186** | **0.9346** | **0.9208** | **0.9245** | **0.5733** | **0.6756** | **0.7669** | **0.8009** |

## III. EXPERIMENTAL VALIDATION

In this section, we first present the experimental protocol in detail and then report the comparison results between the proposed model and other state-of-the-art (SOTA) BIQA models on six in-the-wild IQA databases. Then, the ablation studies are conducted to validate the effectiveness of each module in the proposed model. Finally, we test the generalizability of the proposed model via cross-database evaluation.

### A. Experimental Protocol

*1) Test Database:* The proposed method is mainly validated on six in-the-wild IQA databases, CLIVE [22], BID [23], KonIQ-10K [24], SPAQ [25], FLIVE [26], and FLIVE Patch [26]. CLIVE consists of 1,162 images with diverse authentic distortions captured by mobile devices. BID is a blur image database that contains 586 images with realistic blur distortion such as out-of-focus, motion blur, etc. KonIQ-10K contains 10,073 images selected from the large public multimedia database YFCC100m, which covers a wide and uniform range of distortions in terms of quality indicators such as brightness, colorfulness, contrast, noise, sharpness, etc. SPAQ consists of 11,125 images taken by 66 kinds of mobile devices. These images cover a large range of scene categories like animal, human, plant, indoor scene, cityscape, landscape, night scene, etc. FLIVE is the largest in-the-wild IQA database by far, which contains about 40,000 real-world distorted images and 120,000 randomly cropped patches. We denote the latter as the FLIVE Patch database.

*2) Evaluation Criteria:* Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) are adopted to evaluate the performance of IQA models. These two criteria have different meanings for demonstrating the performance of IQA models, of which PLCC reflects the prediction linearity of the model and SRCC indicates the prediction monotonicity.

*3) Implementation Details:* We use ResNet50 [11] as the backbone of the staircase network. The weights of the backbone are initialized by training on ImageNet, and other weights are randomly initialized. For the FLIVE and FLIVE Patch databases, we use the same pre-processing method in [26] to white fill images to the resolution of 340×340. For images in other databases, we resize the resolution of the minimum dimension of images as 380 while maintaining their aspect ratios. In the training stage, the input images in the FLIVE Patch database and other databases are randomly cropped with resolutions of 224×224 and 320×320 respectively, and in the testing stage, we crop the four corners and center patch with the same resolution of 224×224 for images in the FLIVE Patch database and 320×320 for images in other databases. The quality score of each testing image is averaged by the scores of five patches. The proposed model is implemented in PyTorch. The Adam optimizer [27] with the initial learning rate 0.00001 and batch size 30 is used for training the proposed model on a server with NVIDIA GTX 2080Ti. All databases are split into the training set with 80% distorted images and the test set with 20% distorted images. We randomly split the databases for 10 times, and report the median values of SRCC and PLCC.

*4) Compared Algorithms:* We compare the proposed models with eight state-of-the-art BIQA models including handcrafted feature based BIQA models: NIQE [20], BRISUE [2], and BMPRI [5], and deep learning based BIQA models: CNNIQA [6], WaDIQaM-NR [7], SFA [21], DB-CNN [8], and HyperIQA [9]. We retrained the compared models on the six IQA databases for the fair comparison.

### B. Performance Comparison with the SOTA Methods

The performance results on the in-the-wild databases are summarized in Table I. From Table I, we first observe that the proposed model achieves the best performance on all six in-the-wild IQA databases and it leads by a significant margin, which indicates that the proposed model has more powerful representation abilities for the quality of in-the-wild images than other deep learning based methods as well as handcraft features based methods. Then all handcraft features based models perform poorly on in-the-wild IQA databases, and their performance is obviously lower than deep learning based models, which reflects that handcraft features are difficult to model the quality of images captured under various in-the-wild environments. Third, other deep learning based models such

TABLE II
THE PERFORMANCE OF MODELS WITH DIFFERENT CONVOLUTIONAL PATHS ON THE KONIQ10K DATABASE. S IN THE FIRST ROW MEANS STAGE.

| Convolutional Path | S1 | S2 | S3 | S4 | SRCC | PLCC |
|---|---|---|---|---|---|---|
| $Path$ 1 | √ | √ | √ | √ | 0.9169 | 0.9324 |
| $Path$ 2 | × | √ | √ | √ | 0.9157 | 0.9328 |
| $Path$ 3 | × | × | √ | √ | 0.9146 | 0.9316 |
| $Path$ 4 | × | × | × | √ | 0.9128 | 0.9289 |
| None | × | × | × | × | 0.9100 | 0.9259 |

as HyperIQA and SFA also use ResNet50 as the backbone for extracting features, but their performance is all inferior to the proposed model, which indicates the superiority of the staircase structure for improving the representation ability of the model.

### C. Ablation Experiment

*1) Analyzing Features Extracted from Different Stages:* In this section, we train the backbone network (i.e. ResNet50) with these four convolutional paths individually on the KonIQ10k database to verify the contributions of features extracted from different stages. We list the results in Table III. First, when comparing the performances of $Path4$ to $Path1$, we find the performance increases monotonously as the features extracted from Stage4 to Stage1 are added in sequence to the model, which indicates that the features extracted from all stages make contributions to the overall performance. Then, we observe that the performance gains of $Path3$ to $Path4$ and $Path4$ to fusing no features are larger than the performance gains of $Path1$ to $Path2$ and $Path2$ to $Path3$, which means the features extracted from Stage3 and Stage4 are more important to the image quality evaluation. Finally, the performances of $Path4$ to $Path1$ are all inferior to the proposed staircase structure, which indicates that combing the fused features from these convolutional paths can further improve the model's performance.

*2) The Effects of Different Backbones:* In this section, we test different backbones to show the effect of backbone networks on the performance of the model. Specifically, we train three CNN models, MobileNetV2 [28], ResNext50 [29], and ResNest50 [30] with the staircase structure on the KonIQ10k database. MobileNetV2 is a lightweight CNN model for mobile applications while ResNest and ResNext are two more powerful CNN structures than ResNet. The results are listed in Table III. From Table III, we observe that the performances of ResNest50 and ResNext50 are both superior to the ResNet50 though they have a similar number of parameters. MobileNetV2 has ten times fewer parameters than ResNet, but the SRCC value of Mobilenetv2 is only 0.0123 less than ResNet50. Therefore, the staircase structure is a flexible and effective module for BIQA, which can be integrated with popular CNN models, and we can choose the corresponding model to meet requirements such as a greater emphasis on high accuracy or faster running time.

TABLE III
THE PERFORMANCE OF DIFFERENT BACKBONES WITH THE STAIRCASE STRUCTURE ON THE KONIQ10K DATABASE.

| Backbones | MobileNetV2 | ResNet50 | ResNext50 | ResNest50 |
|---|---|---|---|---|
| SRCC | 0.9063 | 0.9186 | 0.9202 | 0.9228 |
| PLCC | 0.9235 | 0.9346 | 0.9363 | 0.9402 |

TABLE IV
SROCC EVALUATIONS ON CROSS DATABASE TESTS.

| Training | Testing | DBCNN | HyperIQA | Proposed |
|---|---|---|---|---|
| KonIQ-10k | LIVEC | 0.755 | 0.785 | **0.795** |
| | BID | 0.816 | **0.819** | 0.813 |
| BID | LIVEC | 0.725 | 0.770 | **0.793** |
| | KonIQ-10k | 0.724 | 0.688 | **0.734** |

### D. Cross-Database Evaluation

In this section, we test the generalization ability of the proposed model via cross-database evaluation. Specifically, we choose KonIQ10k and BID as the training database because KonIQ10k is a large and distortion-rich database, while BID is a small and distortion-single database, so we can observe the generalization ability of the proposed model in different ways. Then we respectively test LIVEC, BID and LIVEC, KonIQ10k on the two trained models. The two most competitive models DBCNN and HyperNet are selected for comparison, and the results are listed in Table IV. It is observed that among four cross-database tests, the proposed model achieves three times of top performance, and the other is very close to the compared models, which demonstrates the strong generalization ability of the proposed model.

## IV. CONCLUSION

In this paper, we propose a new BIQA model for in-the-wild images. The proposed model consists of two parts: the staircase network for better quality-aware feature extraction and the image quality regressor for mapping the quality-aware features to the quality score. The staircase structure makes the model integrate the features from intermediate layers into the final feature representation, so the model can make full use of visual information from low level to high level. Experimental results show that the proposed model outperforms other state-of-the-art BIQA models on six in-the-wild IQA databases, and also achieves an excellent performance in the cross-database evaluation, which demonstrate the effectiveness and generalizability of the proposed model. What's more, the proposed model is very flexible and can be replaced with popular CNN models to meet the various needs of practical applications.

## V. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation of China (61901260, 61831015) and National Key R&D Program of China 2021YFE0206700.

# References

[1] Anush Krishna Moorthy and Alan Conrad Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[2] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[3] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2014.

[4] Guangtao Zhai, Xiongkuo Min, and Ning Liu, "Free-energy principle inspired visual quality assessment: An overview," *Digital Signal Processing*, vol. 91, pp. 11–20, 2019.

[5] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.

[6] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[7] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[8] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[9] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning 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.

[10] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR 2015 : International Conference on Learning Representations 2015*, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[12] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[13] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[17] Fei Gao, Yi Wang, Panpeng Li, Min Tan, Jun Yu, and Yani Zhu, "Deepsim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, 2017.

[18] Seyed Ali Amirshahi, Marius Pedersen, and Stella X Yu, "Image quality assessment by comparing cnn features between images," *Electronic Imaging*, vol. 2017, no. 12, pp. 42–51, 2017.

[19] Wei Sun, Xiongkuo Min, Guangtao Zhai, Ke Gu, Huiyu Duan, and Siwei Ma, "Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 64–77, 2019.

[20] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[21] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang, "Which has better visual quality: The clear blue sky or a blurry animal?," *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1221–1234, 2018.

[22] Deepti Ghadiyaram and Alan C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

[23] Alexandre Ciancio, Eduardo AB da Silva, Amir Said, Ramin Samadani, Pere Obrador, et al., "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Transactions on image processing*, vol. 20, no. 1, pp. 64–75, 2010.

[24] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[25] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang, "Perceptual quality assessment of smartphone photography," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3677–3686.

[26] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3575–3585.

[27] Diederik P. Kingma and Jimmy Lei Ba, "Adam: A method for stochastic optimization," in *ICLR 2015 : International Conference on Learning Representations 2015*, 2015.

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[29] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[30] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al., "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.