J. Vis. Commun. Image R. 82 (2022) 103399

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Full Length Article

Attention integrated hierarchical networks for no-reference image quality assessment☆

Junyong You [a,*], Jari Korhonen [b]

[a] NORCE Norwegian Research Centre AS, Bergen, Norway
[b] Shenzhen University, Shenzhen, China

ARTICLE INFO

Keywords:
Attention
Hierarchical networks
Image quality assessment (IQA)
Perceptual mechanisms
Quality perception

ABSTRACT

Quality assessment of natural images is influenced by perceptual mechanisms, e.g., attention and contrast sensitivity, and quality perception can be generated in a hierarchical process. This paper proposes an architecture of Attention Integrated Hierarchical Image Quality networks (AIHIQnet) for no-reference quality assessment. AIHIQnet consists of three components: general backbone network, perceptually guided neck network, and head network. Multi-scale features extracted from the backbone network are fused to simulate image quality perception in a hierarchical manner. The attention and contrast sensitivity mechanisms modelled by an attention module capture essential information for quality perception. Considering that image rescaling potentially affects perceived quality, appropriate pooling methods in the non-convolution layers in AIHIQnet are employed to accept images with arbitrary resolutions. Comprehensive experiments on publicly available databases demonstrate outstanding performance of AIHIQnet compared to state-of-the-art models. Ablation experiments were performed to investigate the variants of the proposed architecture and reveal importance of individual components.

1. Introduction

With the rapid development of smart devices and social media platforms, there has been an explosion of user-generated content (UGC). Consequently, evaluation of Quality of Experience (QoE) of UGC images is becoming a critical issue, considering that most images are produced in an unprofessional manner and often suffer from authentic distortions. Full-reference image quality models with access to undistorted image signals as a reference point have already demonstrated high accuracy to assess the fidelity of distorted images compared with the reference [1]. However, UGC images with authentic distortions do not have any references. Thus, no-reference image quality assessment (NR-IQA) is naturally the only option for QoE measurement [2]. Earlier works on NR-IQA mainly concentrated on investigating quality features related to certain distortion types, e.g., blocking artifact [3], image blurring and sharpness [4]. These image quality models have achieved high correlation with subjective judgment, but the range of application scenarios is limited, as prior knowledge about distortion types must be available. With the rapid growth of UGC content, authentic image distortions become more and more relevant. Thus, natural image quality assessment not specialized in distortion types is more demanded.

Based on the premise that the visual perceptual system is designed to interpret statistical regularities in natural scenes, Sheikh [5] approached IQA in his dissertation from a novel direction by assuming that image distortions make natural scenes in images look “unnatural”. Consequently, natural scene statistics (NSS) modeled by information-theoretic approach can be employed to define viewing behavior in a natural task, including quality assessment [6–8]. In [7], the natural image quality evaluator (NIQE) is based on the construction of quality aware collection of statistical features extracted by a NSS model. As NSS can represent the generic distribution of characteristics of image presentations, these approaches can ignore the assumption of image distortion types. Inspired by NIQE, Zhang *et al.* [8] trained an opinion-unaware multivariant Gaussian model (IL-NIQE) based on NSS features for local image quality prediction, which can provide stronger generalization capability than those IQA models following an opinion-aware approach by learning regression model based on subjective quality opinion.

Due to the capability of machine learning in revealing the hidden patterns of target tasks from training data, it has also attracted wide interests in IQA research. Most efforts have been focused on engineering

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: junyong.you@norceresearch.no (J. You).

https://doi.org/10.1016/j.jvcir.2021.103399
Received 27 July 2021; Received in revised form 18 October 2021; Accepted 25 November 2021
Available online 6 December 2021

features related to visual quality perception and employing machine learning models to predict image quality [7–11]. For example, a widely referenced blind image spatial quality evaluator (BRISQUE) derives scene statistics of locally normalized luminance coefficients in the spatial domain to quantify the losses of naturalness, and an image quality score can be estimated by any machine learning based regressor [11].

Even though classical learning-based approaches achieve promising performance in IQA, they often require sophistically designed features, in which both the characteristics of human visual system (HVS) and image attributes should be considered. Lately, with the advancement of deep learning, allowing feature engineering to be an integrated part of a deep learning model, it has been dominating computer vision tasks [12]. Naturally, NR-IQA models based on deep learning have been proposed, following the commonly used architecture in image recognition, e.g., convolution neural networks (CNN) as feature extractors followed by single- or multi-layer perceptron for quality prediction, see e.g. [13–20]. Section 2 will present a detailed review of the related work on deep learning driven NR-IQA models.

The architecture of using CNNs and regression layers cannot fully exploit image information that is relevant for quality perception. For example, spatial details might be lost when image information is abstracted by deeper and deeper layers in a CNN architecture, even though such deep structure is expected to efficiently represent the abstract semantics of the image. In vision research, hierarchical multi-scale perception architectures are often employed to model visual perception process in viewing tasks that can represent relevant features at multiple spatial and semantic scales. For example, multi-scale structural similarity model (MS-SSIM) has shown better performance than single-scale SSIM for IQA [21]. In [22], a simple side pooling net has also been used to pool features extracted from multi-scales in a CNN architecture. Another typical example of hierarchical architecture is object detection, where feature pyramid network (FPN) built on different scales of base network layers can be employed to detect objects at multiple scales [23]. The motivation is that objects with different sizes can both be retained and detected in a multi-scale manner. Considering that the HVS can perceive visual stimuli at different scales, e.g., from coarse to fine [24,25], we assume that multi-scale perception modelling also benefits IQA.

On the other hand, important mechanisms in visual perception should be considered in IQA models. As a key mechanism, attention drives a set of cognitive operations occurring in the HVS to focus on selected visual stimuli, while other perceptible information in the visual field is overtly or covertly ignored [26]. Thus, visual attention can significantly influence how perception is generated in IQA and should be considered in development of visual quality models [27,28]. Psychovisual experiments have also demonstrated that the HVS has different sensitivities to varying visual stimuli at different frequencies modeled by contrast sensitivity function (CSF) [29,30], determining the most essential information for image quality perception.

In this work, a deep architecture of attention integrated hierarchical image quality network (AIHIQnet) for NR-IQA is proposed. Relevant quality features are first extracted at multiple scales from a *backbone* network. Subsequently, a perceptually guided *neck* network is built, consisting of multi-scale feature fusion and a special attention module simulating the spatial attention and contrast sensitivity mechanisms. Finally, following the subjective process of IQA, where multiple participants rate the quality, a *head* network is proposed to predict the distribution of quality scores given to an image. Furthermore, other deep learning driven computer vision problems often employ image rescaling to generate constant input sizes and reduce computing resource requirements. However, such rescaling operation can potentially affect quality perception. For example, viewers might prefer high resolution rather than low resolution images on large displays. In order to handle the issue of varying input resolutions in IQA, the proposed hierarchical networks mainly contain resolution-independent full convolution networks and subnetworks. The main contributions of this work are summarized as follows:

1) A generic NR-IQA model for quality prediction of natural images with authentic distortions and arbitrary resolutions.

2) An architecture based on multi-scale feature fusion to derive image quality perception in a hierarchical manner.

3) An attention module simulating the mechanisms of attention and contrast sensitivity in the HVS to capture the perceived information that is essential for image quality perception.

The remainder of this paper is organized as follows. Related work on deep learning driven NR-IQA models is reviewed in Section 2. Section 3 presents the details of AIHIQnet. Section 4 reports comprehensive experiments, including comparison with state-of-the-art IQA models, analysis, and ablation experiments. Finally, concluding remarks are drawn in Section 5.

2. Related deep learning driven NR-IQA models

A commonly used architecture in deep learning driven IQA models consists of CNNs to extract image features, followed by fully-connected (FC) layers for quality value regression. CNN architectures have natural advantages for IQA, e.g., convolution can capture structural changes of images that also effectively indicate quality change. A widely used approach is based on image patching [13–16]. For example, inspired by the pioneering AlexNet architecture for image recognition [31], Kang *et al.* [13] proposed a simple architecture consisting of a single convolution layer and pooling layers on normalized 32 × 32 pixel image patches. On the other hand, following the successful applications of VGG architecture using small convolution kernels and deeper layers [32], Bosse *et al.* [14,15] proposed deepIQA to predict the quality score of randomly sampled image patches (32 × 32 pixels) and then use their average as full image quality score. In [16], Li *et al.* proposed a deep architecture consisting of four convolution blocks with variant kernel sizes and apply it on image patches of 224 × 224 pixels. However, using small image patches in model training is based on the assumption that each image patch has the same quality level as its source image. Such assumption can be unreliable, as image quality is often spatially inconsistent, and viewers might pay different attention to different areas driven by the attention mechanism.

Several models attempt to exploit existing CNN architectures pretrained on large-scale image sets in the scenario of IQA [17,18]. Gao *et al.* [17] employed the VGG-net pretrained on ImageNet to derive images features at different levels, which are then concatenated and fed into SVR for image quality prediction. Similarly, Bianco *et al.* [18] used AlexNet pretrained on ImageNet and other large-scale databases and then fine-tuned the model for IQA. In order to fit the existing CNN models with fixed input size, rescaling or patching images of different resolutions is often performed. Features from different patches are most commonly pooled by simple averaging, but more advanced pooling strategies have also been proposed. In our earlier work on video quality assessment, a long short-term memory (LSTM) network for spatiotemporal pooling of small cubic video clips has been used [19].

As explained earlier, image rescaling might change the perceived quality, compared to the original image. For image patching based quality assessment, in addition to the unreliable assumption of spatially constant quality as explained above, another disadvantage lies in the difficulty of developing an end-to-end learning approach based on patches. This also applies to the approach of using CNN for feature extraction and other regressors (e.g., SVR) for quality prediction. Furthermore, by directly applying CNN architectures pretrained on image sets for the purposes other than quality assessment might not be the optimal way to fully exploit the benefits of large-scale pretraining. Our ablation experiments will demonstrate that even though CNN architectures pretrained on ImageNet can provide solid foundation for downstream tasks, e.g., IQA, further transfer learning should be performed appropriately to achieve good results.

In order to fully exploit the advancement of deep learning, more studies on deep learning driven end-to-end trainable NR-IQA models have been proposed [20,33–38]. Following the common approach, Hosu *et al.* [20] proposed Koncept512 consisting of a base CNN network (InceptionResNetV2 [39]) followed by several fully connected layers to implement regression to predict image. They also created a large-scale image quality database to train and evaluate the model.

As mentioned above, image quality perception can be significantly influenced by other factors, e.g., distortion types and visual mechanisms. In [33], Ma *et al.* proposed a multi-task end-to-end network (MEON) consisting of two sub-networks sharing early layers for distortion identification and quality prediction. The training of quality prediction network also benefits from the distortion identification network which can be trained with readily generated data. However, MEON limits its fixed input resolution to 256 × 256, and for images with larger resolutions, either majority vote strategy of extracted sub-images or averaging is still required. Zhang *et al.* [34] attempted to tackle two distortion types (synthetic and authentic) in a deep bilinear model consisting of two CNNs (DBCNN). As an important visual feature for image quality perception, structural information has been widely used in IQA [521]. Considering that structural information can be appropriately represented by image gradient, Yan *et al.* [35] have proposed a two-stream NR-IQA model, where the image stream attempts to model pixel features by CNN and another gradient stream attempts to model structural information based on image patches. The two streams are merged to predict the overall image quality.

As an important mechanism in the HVS for quality perception, attention or its derived saliency information has also been widely used in IQA. A typical example of integrating attention or saliency information to IQA models is to multiply attention or saliency map with image or feature map. Yang *et al.* [36] proposed an end-to-end multi-task network (SGDNet) to predict image saliency and quality jointly, and image saliency can also serve as a weight map that is used for element-wise multiplication with the feature map derived from a CNN based model. Chen *et al.* [37] proposed an interesting idea of integrating reinforcement learning (RL) into NR-IQA by forcing the model to learn policy attended to fixation regions. On the other hand, in our previous studies [41–43], we have observed that directly multiplying attention or saliency map with images or low-level feature maps might introduce loss of important information for quality perception. Instead, we hypothesize that it is more appropriate to integrate attention into relevant visual mechanisms. For example, we found that multiplying attention map with the critical frequency in CSF shows promising performance in video quality assessment [43]. This also motivates us to use a deep learning approach to integrate attention mechanism into CSF in this work, e.g., using an attention module in AIHIQnet. Inspired by the fact that the advanced Transformer architecture can efficiently capture the attentional information in the input signals [44], we have developed a hybrid architecture (TRIQ) by applying the encoder of Transformer to the features extracted by a backbone CNN, which achieves high performance compared with other state-of-the-art models in IQA [38].

In addition, multi-scale image representation can provide more perceptual semantcis for visual tasks. In [22], Wu *et al.* proposed a cascaded architecture (CaHDC) to represent the hierarchical perception mechanism in HVS, and then pool the features extracted at different scales by a simple side pooling net (SiPNet). The model was trained with respect to combining several image quality databases. Another interesting approach (hyperIQA) is proposed to use a hyper network to establish perception rule adaptive to image contents [39], in which multi-scale image features representing both local and global distortions by a backbone network (ResNet50) can be aggregated for image quality prediction.

In order to build a general purpose NR-IQA model, it is necessary to appropriately integrate relevant visual mechanisms, e.g., attention, CSF, crossing scale perception, into a deep architecture. This motivates us to design the AIHIQnet model to exploit the advantage of feature representation capability of deep networks and the importance of perceptual mechanisms in IQA. To demonstrate the performance of AIHIQnet, several other models that can well represent different types of deep learning driven NR-IQA algorithms have been included in this study as benchmarks. Table 1 summarizes these models and their settings in our work.

Table 1
Explanations of the Deep Learning Driven NR-IQA Models Included In Our Work.

Models	*Brief explanations*
DeepBIQ [18]	Each image is cropped into 5 × 5 patches with size of 224 × 224, then the AlexNet without top layers pretrained on ImageNet is used to calculate features of each patch. The averaged features over all patches are taken as the features at image level, and then fed into SVR for quality prediction. Model is implemented by PyTorch.
Koncept512 [20]	InceptionResNet-V2 [40] without top fully connected layers to extract image features, and a global average in spatial dimension, and then add three fully connected layers and dropout layers. Finally, a fully connected layer with one filter for MOS prediction. Can be adjusted to accept varying image resolutions by specifying the input shape as [None, None, 3] together with global pooling method.
CaHDC [22]	A VGG-like base network to generate 4 feature maps (i.e., intermediate outputs of convolution layers at different scales). The results are concatenated after max pool, and then a bottleneck convolution layer with dropout, and finally another convolution layer with 1 filter and global average to predict image quality. Can be adjusted to accept varying image resolutions by specifying the input shape as [None, None, 3] together with global pooling method.
MEON [33]	TensorFlow implemented provided by the authors (https://ece.uwaterloo.ca/~zduanmu/tip2018biqa/) was modified in our work. Each image is divided into overlapped patches with size of 256 × 256, the overlap size is 128 × 128. When training the model, we assumed all the patches share same quality score as the shource images, but the maximal prediction quality value over patches is used in the inference.
DBCNN [34]	VGG16 and the S-CNN implemented by the authors are combined by a global bilinear pooling operation. The model can accept arbitrary image resolutions. The official Matlab implementations published on Github (https://github.com/zwx8981/DBCNN) were employed.
SGDNet [36]	Resnet50 [45] to generate a feature map, which is processed by a convolution layer or a squeeze-and-excitation [46] block and then multiplied with the image saliency map. And then followed by two fully connected layers with dropout, and a fully connected layer with one filter for MOS prediction. Can be adjusted to accept varying image resolutions by specifying the input shape as [None, None, 3] together with global pooling method and by resizing the resolution of saliency map accordingly.
TRIQ [38]	Our proposed model and it is publicly available (https://github.com/junyongyou/triq), adaptive positional embedding is used to handle arbitrary image resolutions.
hyperIQA [39]	Official implementation by PyTorch (https://github.com/SSL92/hyperIQA) and default parameters were trained in our database settings. Images are rescaled to the same size during training, and therefore, images with different resolutions than the trained resolution also need to be rescaled in inference.

3. AIHIQnet: Attention integrated hierarchical image quality networks

Since the revolutionary work AlexNet for image recognition using deep neural networks [31], many advanced networks (e.g., VGG [32], InceptionResnetV2 [40], ResNet [45]) have demonstrated that traditional hand-crafted features can be replaced by the networks using images directly as inputs. With a sufficiently deep architecture, the earlier layers can serve as "feature extractor" due to the outstanding representative capacity of deep networks. Consequently, the subsequent layers can perform target tasks based on those extracted features. For example, image classification is often performed by fully connected layers (head) built on top of convolution layers (backbone) to classify an

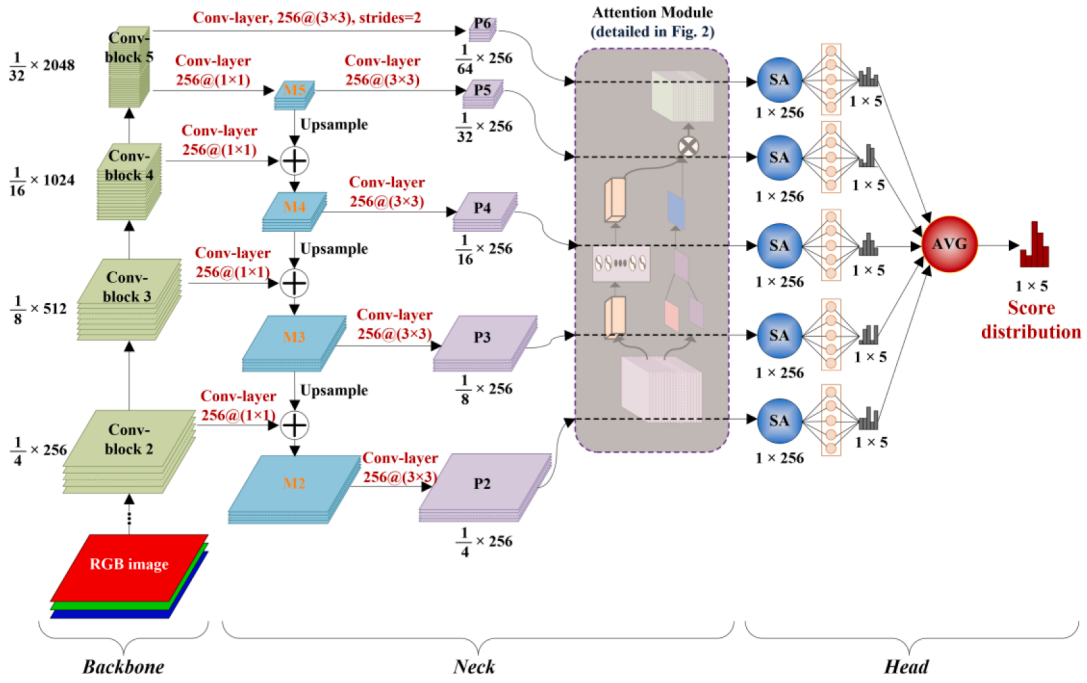

Fig. 1. Architecture of the perceptual hierarchical image quality networks (AIHIQnet). The numbers indicate the output shape of each block/node.

image into predefined categories. In the research of object detection, intermediate layers between backbone network and final head layers are widely employed.

Most convolution networks use increasing number of convolutional filters together with spatial strides to organize the contained layers so that semantic image features are abstracted from space to channels. In other words, later layer outputs have smaller spatial scales whilst higher (channel) depth. In this way, semantic features can be extracted at scales from low to high, when they are propagated from the earlier layers to the later ones. Top layers usually contain more representative semantics for a target task, whereas too small spatial resolutions might also cause limitations for solving the task. Thus, considering that sufficient spatial details carry relevant information for image quality perception, we intend to employ semantic features both from low scales (high resolutions) and high scales (low resolutions). The aim of AIHIQnet is to generate image quality prediction at different scales and combine them to derive the overall quality. Fig. 1 illustrates the architecture of AIHIQnet consisting of three parts: backbone, neck and head networks, as detailed in the following subsections.

A. Backbone network

AIHIQnetThe purpose of backbone network is to extract image features relevant for quality perception through the learning process. In general, the choice of the backbone network is not very crucial for AIHIQnet. A CNN architecture following the common approach to organize the convolution blocks/layers and generate feature maps at different scales can be served as the backbone network. In the AIHIQnet implementation, we have mainly focused on ResNet family, e.g., ResNet50. ResNet152, VGG16 [32] and DenseNet121 [47] have also been included in the ablation experiments. However, the experiments and following descriptions are all based on ResNet50. The implementation of ResNet in the official release of TensorFlow [48] has been employed in this work.

ResNet50 consists of 5 stages (residual blocks) each containing a convolution block and identity block. In this work, we choose the outputs from the later 4 stages (named C_2, C_3, C_4, and C_5) to derive feature maps for IQA. The output from the first stage is not included because it contains relatively shallow semantics and requires a lot of memory. The backbone network is in fact a full convolution network, whose configuration is independent of the input shape. In this case, only the parameters of the convolution layers (e.g., filter number, kernel size, strides) need to be specified. Such architecture can accept any image inputs with different resolutions. Assuming an input RGB image of resolution [X, Y, 3], where X and Y denote height and width, the output shapes of the backbone network are:

$$\begin{cases} C_2 : [X/4, Y/4, 256] \\ C_3 : [X/8, Y/8, 512] \\ C_4 : [X/16, Y/16, 1024] \\ C_5 : [X/32, Y/32, 2048] \end{cases} \tag{1}$$

B. Neck network for feature fusion and attention modelling

As explained above, image quality perception can be better modelled in an approach of feature fusion between high-scale (i.e., low resolution or more abstract) and low-scale feature maps. To perform the cross-scale feature fusion without introducing large additional computational cost, a bottleneck convolution layer is first applied to the hierarchical feature maps {C_2, C_3, C_4, C_5}:

$$Conv(K, (1 \times 1), 1)(C_i), i = 2, 3, 4, 5 \tag{2}$$

where K denotes the filter number that is set to the smallest number of channel dimension in {C_2, C_3, C_4, C_5}, thus 256 is used, 1×1 indicates the kernel size, and another 1 is the striding step. The main purpose of the bottleneck convolution layer with kernel size of 1×1 is to reduce the channel dimensions of {C_2, C_3, C_4, C_5} without changing their spatial resolutions.

Subsequently, up-sampling is performed for cross-scale fusion, i.e., up-sampling the spatial resolution of high-scale feature map to match its next low-scale map. The up-sampled feature maps are fused with convolution results of the next high-scale map to generate intermediate maps {M_5, M_4, M_3, M_2}.

$$\begin{aligned} M_5 &= Conv(K, (1 \times 1), 1)(C_5) \\ M_i &= Conv(K, (1 \times 1), 1)(C_i) + Up - sample(M_{i+1}), i = 4, 3, 2 \end{aligned} \tag{3}$$

where M_5 is directly generated from C_5 by the convolution layer, as C_5 is the highest scale map from the backbone network.

Consequently, another convolution layer using K filters with kernel size (3×3) is applied to {M_2, M_3, M_4, M_5} to generate pyramidic feature maps {P_2, P_3, P_4, P_5}, as in Eq. (4). Other kernel sizes have also been

tested, e.g., 5 × 5, while it has been found to perform worse for AIHIQnet in IQA.

$$P_i = Conv(K, (3 \times 3), 1)(M_i), i = 2, 3, 4, 5 \tag{4}$$

In addition, a common understanding about CNN architecture is that more semantic information is conveyed by more abstract representation. Thus, we intend to more fully exploit the highest scale feature map, i.e., C_5, for IQA by another convolutional operation with a striding step 2. Same as generating the other pyramid feature maps $\{P_2, P_3, P_4, P_5\}$, the kernel size of this convolutional operation is also set to (3 × 3). Thus, another feature map P6 can be generated from C_5 directly, as formulated in Eq. (5).

$$P_6 = Conv(K, (3 \times 3), 2)(C_5) \tag{5}$$

The pyramidic feature maps $\{P_2, P_3, P_4, P_5, P_6\}$ carry perceptible information extracted from the input image for quality assessment. However, due to the selective attention mechanism, not all the perceptible information will contribute to quality perception evenly. As explained by the spatial selective attention mechanism, the HVS often selectively perceives the most important or informative stimuli, while ignoring other perceptible information [26,49]. In other words, viewers pay more attention to certain spatial areas than others in the field of vision.

In addition, contrast sensitivity measured as a function of retinal eccentricity for visual stimuli in spatial frequency is maximized at the fovea and declined with eccentricity to the gaze [30], as described in the following equation.

$$CT(f, e) = CT_0 \cdot exp\left(\alpha \cdot f \cdot \frac{e + e_2}{e_2}\right) \tag{6}$$

where f denotes the spatial frequency, e is the retinal eccentricity. A critical frequency f_c, beyond which the contrast will be imperceptible, can be obtained by setting CT to 1.0 (the maximum possible contrast) and solving for f:

$$f_c = \frac{e_2 \cdot log(1/CT_0)}{\alpha \cdot (e + e_2)} \tag{7}$$

In our earlier work [43], a psychovisual experiment has been conducted demonstrating that the critical frequency can be further adjusted according to spatial attention map, approximating a dot-product way as following:

$$f_c^{'} = f_c \cdot [\rho + (1 - \rho)] \cdot AM \tag{8}$$

where AM denotes distribution of spatial attention, and ρ is a control parameter.

Attention guided contrast sensitivity mechanism indicates that the HVS has stronger reflection to the components at certain frequency in visual stimuli than other components. Thus, such mechanism can also be represented by a means of attention network, i.e., more attention is paid to components with certain contrast frequencies than the others. For example, no attention is paid to the frequency components beyond the critical frequency. Therefore, even though perceptible information for IQA has already been extracted and underlies in the feature maps $\{P_2, P_3, P_4, P_5, P_6\}$, an attention module implementing contrast sensitivity and selective attention mechanism is still conducive to detecting the most crucial information for IQA.

As contrast sensitivity and selective attention are two separate but related mechanisms in the HVS, we also implement the attention module in such an approach. The attention module contains two attention blocks, namely channel attention representing contrast sensitivity and spatial attention for selective attention.

Even though the channel features in each of the pyramidic feature maps $\{P_2, P_3, P_4, P_5, P_6\}$ are not same as the frequency components used in CSF, we hypothesize that the channel features still represent similar mechanism of contrast sensitivity. The channel features actually represent the responses of image signals to different convolution kernels, corresponding to different frequency components. A simplified example is Sobel operator, which is a convolutional operator that can distinguish high frequency information (e.g., edges) from low frequency information (e.g., plain areas). Thus, our hypothesis has a solid foundation. In the channel attention block for contrast sensitivity mechanism, inspired by the squeeze-and-excitation networks [46] and convolutional block attention module [50], we first squeeze a pyramidic feature map using two methods, namely spatial maximum and average pooling schemes. Subsequently, a fully connected layer shared by the maximum and average pooled feature vectors is employed to derive attention weights from the two pooled feature vectors. Individual components in the maximum or average pooled feature vector represent features at different frequencies, and we aim to use the attention module to derive the relative important levels for individual components of quality perception. Therefore, Sigmoid function has been chosen as the activation function in the fully connected layer. The number of filters in this layer is set to the number of channels of the pyramidic feature map, i.e., K, so that the dimension of channel weights matches the feature map. Consequently, the average between the two channel weights derived respectively from the maximum and average pooled feature vectors is obtained, which serves as the channel weight to represent the attention distribution over frequencies for quality perception. The channel attention block can be represented roughly as follows:

$$CA = avg\{s\{FC[AvgPool(P)], FC[MaxPool(P)]\}\} \tag{9}$$

where FC and s denote the fully connected layer with Sigmoid as activation function, and avg is the average of the two channel weights. As CSF indicates that visual perception can be influenced by frequencies beyond certain threshold, the channel attention block is appropriate for simulating CSF, e.g., by using the unbalanced attention distribution over different channels.

As explained earlier, multiplying a spatial attention or saliency map directly with an image or low-level feature maps might cause loss of information essential to quality perception. Thus, we intend to apply spatial attention to the high-level feature representations. Inspired by Eq. (8), the spatial attention block is applied to the output of the channel block in AIHIQnet.

In the spatial attention block, channel-wise maximum and average pooling schemes are first applied to a pyramidic feature map to squeeze information in the channel domain. Consequently, two squeezed feature maps are concatenated. As the spatial attention block is expected to derive attention allocation over spatial locations in the feature map for quality perception, a convolution layer using one filter is employed to generate a single weight map with the same spatial size as each of the pyramidic feature maps $\{P_2, P_3, P_4, P_5, P_6\}$. A relatively large kernel size 7 × 7 is used in the convolution layer to increase its receptive field. Other kernel sizes, including 5 × 5 and 9 × 9, have also been tested in this work, and we found that kernel size of 7 × 7 provides the best results. Similar to the channel attention block, Sigmoid function is employed as the activation function in the convolution layer. A spatial weight map can be derived as follows:

$$SA = s\{Conv(1, (7 \times 7), 1)[AvgPool(P), MaxPool(P)]\} \tag{10}$$

Following the idea of the Transformer model [44], and also inspired by the relationship between contrast sensitivity and attention that can be represented by dot-product as in Eq. (8), dot-product attention without scaling factor is employed to implement the attention module in this work. The perceived quality feature maps (PF) at five spatial scales are derived as follows, with the same shape as the pyramidic feature maps $\{P_2, P_3, P_4, P_5, P_6\}$ in Eqs. (4) and (5):

$$PF_i = SA \otimes (CA \otimes P_i), i = 2, 3, 4, 5, 6 \tag{11}$$

Finally, it should be noted that the same attention module is used to

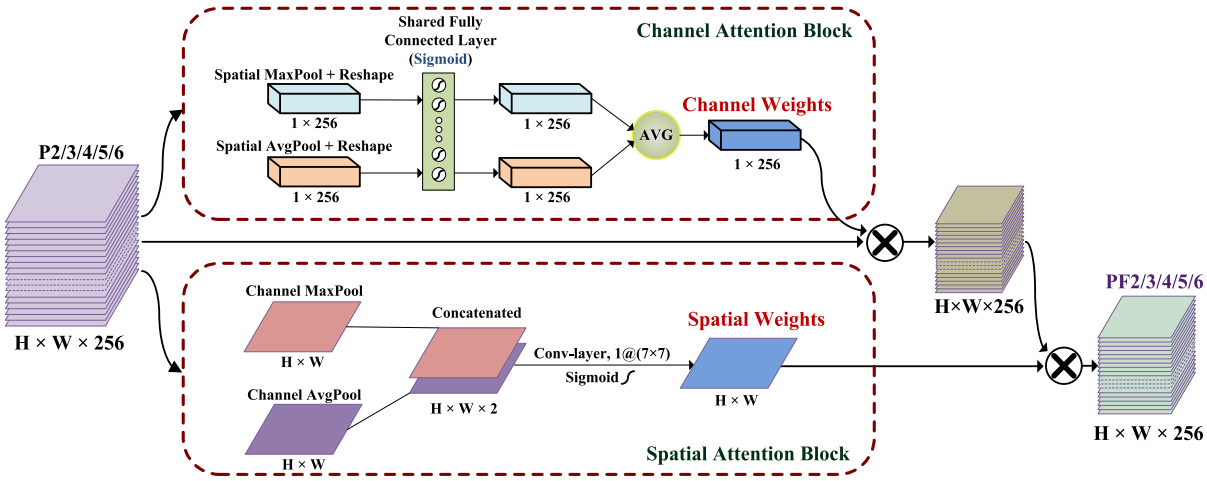

Fig. 2. Attention module in IQA including channel attention block for contrast sensitivity mechanism and spatial attention block for selective attention mechanism, and the number below each node indicates the output shape.

process all the five pyramidic feature maps. Fig. 2 shows the flowchart of the attention module. Several alternative operations in the two attention blocks have also tested in this work, e.g., summation or dot product to replace *avg* in Eq. (9), as well as using averaging instead of concatenation in the spatial attention block. We have found that the approach described above achieves the best performance in IQA in general.

C. Head network for prediction of image quality scores

In practical quality assessment, multiple subjects are often recruited to vote for image quality over a predefined rating scale, e.g., a five-scale rating (1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent) is often used. Subsequently, mean opinion score (MOS) of all participants is calculated to indicate the perceived quality of the image. An IQA model can target at predicting either distribution of quality scores over the rating scale or single MOS value. However, the distribution of ratings might be useful in certain applications, e.g., user experience survey. Previous studies also demonstrated that predicting score distribution appears to provide more robust prediction than predicting MOS directly [51]. This will also be confirmed in our experiments. We assume that such phenomenon is related to the fact that label smoothing often improves the robustness of multi-class classification [52]. Using the distribution of quality scores rather than one-hot coding (i.e., MOS) is assumed to work like label smoothing. Thus, AIHIQnet is set to predict the distribution of ratings, rather than the quality score directly, and MOS values can be calculated from the distribution straightforwardly as:

$$IQ = \sum_{x\in\{1,2,3,4,5\}} x\cdot p(x) \tag{12}$$

where $p(x)$, $x\in\{1,2,3,4,5\}$ denotes the normalized distribution of the quality scores.

As the perceived quality features at different scales are generated from the neck network, the goal of the head network is to combine them to derive the normalized distribution of quality scores. At each spatial scale, global spatial average pooling is first performed to squeeze the perceived quality feature map to generate a feature vector with K dimensions. Such global pooling can also handle image inputs with arbitrary resolutions. Other advanced pooling approaches can also be used to combine the features across space and channels. We have tested the spatial pyramid pooling proposed by K. He *et al.* [53] that can handle arbitrary input image resolutions, but no performance gain was observed for AIHIQnet in our experiments. We assume that the features derived through the backbone and neck networks with the integrated attention blocks can efficiently model quality perception, and then simple global spatial pooling can easily combine the features in the head network for quality prediction. Subsequently, a fully connected layer with five outputs is employed to predict the normalized distribution of quality scores at each scale, as the five-point rating is used in this work. It should be noted that such rating scale is different from the spatial scale of quality features, even though they are both set to five, by coincidence, in this model. As the last fully connected layer should predict the normalized probability distribution representing the quality score distribution at each spatial scale, Softmax is employed as the activation function of this layer, as follows:

$$p_i(x) = FC_i\{avg(PF_i); softmax\}, \quad \begin{matrix} x\in\{1,2,3,4,5\}, \\ i = 2,3,4,5,6 \end{matrix} \tag{13}$$

where $FC_i\{\text{avg}(PF_i)\}$ denotes the fully connected layer on the global average of PF_i.

Finally, the distribution probabilities derived at the five spatial scales are averaged to obtain the overall distribution of the image quality scores, i.e.,

$$p(x) = \overline{p_i(x)}, x\in\{1,2,3,4,5\}, i = 2,3,4,5,6 \tag{14}$$

As the distribution of quality scores represents the normalized distribution of image quality votes over the rating scales, cross entropy is used as the loss function for the predicted distribution against the ground truth score distribution. We have also tested other loss functions, including loss for ordinal classification [54] and the earth mover's distance (EMD), as used in neural image assessment [55], and observed that in general, cross entropy produces the most robust results for AIHIQnet.

On the other hand, it is also easy to predict MOS directly, instead of the distribution of quality scores, in the model. This is done by changing the number of outputs from five to one, and activation function from Softmax to linear in Eq. (13) for single value prediction. Consequently, a single value will be produced at each scale, and then the average over the five scales is calculated for MOS prediction, as given in Eq. (15).

$$MOS_i(x) = \overline{FC_i}\{avg(PF_i); linear\}, i = 2,3,4,5,6 \tag{15}$$

Accordingly, cross entropy is replaced by mean squared error (MSE) for direct MOS prediction. This model is named AIHIQnet-MOS in the experiments.

4. Experiments and discussions

In order to evaluate the performance of AIHIQnet, several state-of-the-art models representing typical machine learning based NR-IQA are included in our experiments, namely NIQE[7], IL-NIQE[8], BRISQUE [11], and the deep learning driven models explained in Table 1.

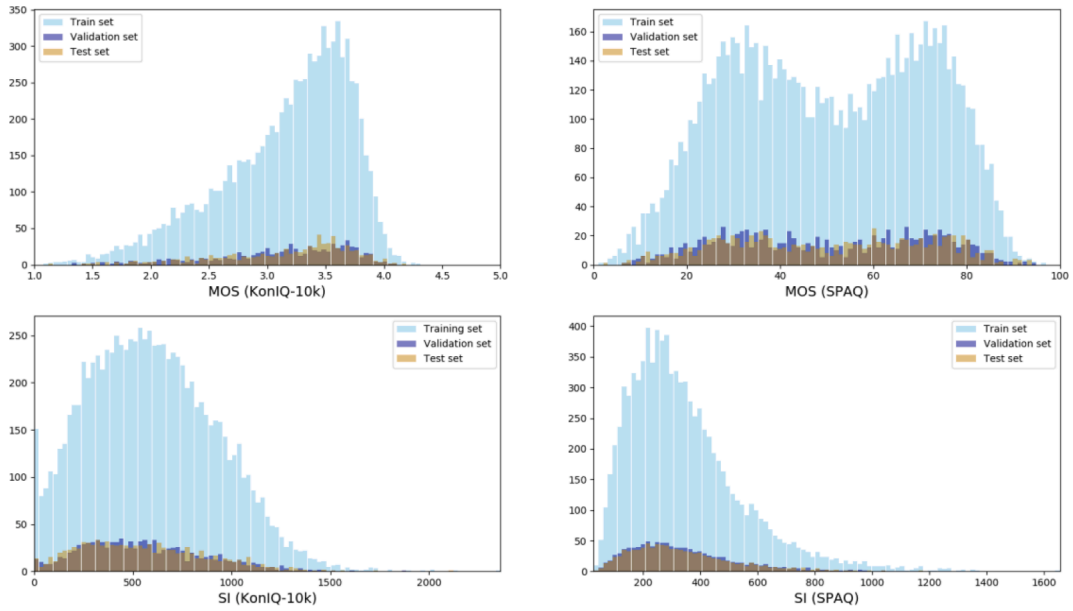

Fig. 3. Histograms of MOS and SI of the train, validation and test images in KonIQ-10 k (left) and SPAQ (right) databases.

Most studies report correlation coefficients, e.g., Pearson linear correlation (PLCC) and Spearman rank-order correlation (SROCC), between the predicted quality scores and ground truth scores to evaluate the performance of quality models, e.g., see [20,22,34,35,39]. A machine learning driven IQA model often aims to minimize the distance measure between single predictions and ground truth, while a correlation coefficient only provides a measure of statistical characteristics between multiple predictions and ground truth. A correlation coefficient is not always positively correlated with a distance measure. Therefore, in addition to PLCC and SROCC, the root mean squared error (RMSE) between the predicted quality scores and ground truth has also been included as an evaluation criterion. Finally, ablation experiments have also been conducted to reveal impact of different mechanisms of AIHIQnet on the performance, providing us with useful insight to advance development of deep learning based IQA models.

A. Image quality databases

Training a deep learning driven IQA model often requires a large amount of annotated data. There are many publicly available IQA databases containing images with authentic or synthetic distortions in either with or without reference scenarios. Even though some attempts have been made to combine different databases in training a single IQA model, e.g., an uncertainty-aware unified model (UNIQUE) for handling cross-distortion-scenario issue [56], it is still difficult or impossible to directly combine multiple quality databases produced in different subjective assessment experiments. Therefore, we intend to employ individual large-scale IQA databases in our experiments.

To the best of our knowledge, there are currently three large-scale IQA databases with authentic distortions publicly available, namely KonIQ-10 k [20], SPAQ [57], and LIVE-FB [58]. In [20], Hosu *et al.* collected over 10,000 natural images with resolution of 1024×768 and conducted a crowdsourcing experiment to produce the KonIQ-10 k image quality database. The MOS values and original quality score distributions were both published. The authors also evaluated their quality model using half-sized images and the original quality scores, and reported better performance than using the original resolution images [20]. It is difficult to judge the legitimacy of the approach of using quality models with smaller resolution images than those used in the subjective study to obtain the ground-truth subjective scores, as the performance is influenced by two independent factors: model design and the impact of downscaling. The SPAQ database was produced in a controlled laboratory environment, in which several human opinions (image quality, image attributes and scene categories) were collected on 11,125 images with diverse resolutions taken by smart phones. In addition, Ying and Niu *et al.* [58] introduced the largest (by far) IQA database (LIVE-FB) containing 39,806 images. However, even though the LIVE-FB database contains a large number of images, about 92% images have the MOS values located in a narrow range of [60,80] out of the full range [0,100]. This is mainly due to the fact that the LIVE-FB database is actually assembled from several other existing databases, which are not designed for IQA purpose and the used images often have fair quality. Thus, LIVE-FB cannot appropriately represent the distribution of varying image qualities. We have actually tested the NR-IQA models on the LIVE-FB database, as will be briefly discussed in Section IV.C, while none of them achieves promising performance. Thus, in our experiments, only the images with original resolution in KonIQ-10 k database and the SPAQ database were mainly used. Considering that these two databases were produced from different assessment environments, i.e., crowdsourcing and controlled laboratory, it is also interesting to compare the two assessment methodologies based on IQA models.

Subsequently, following the standard protocol, the two databases were split into train, validation and test sets. A crucial rule for splitting is that the resulting sets should be independent and identically distributed. In our work, we split the individual database *randomly* according to both the MOS values and complexity of images, namely spatial perceptual information (SI) as defined in the ITU Recommendation [59]. All images in a database were first roughly classified into two complexity categories in terms of SI: high SI and low SI, and then the images in each category were further divided into five quality categories based on their MOS values. Consequently, we randomly chose 80% of the images from each quality category within each complexity category as train images, 10% images as the validation images, and the rest 10% served as test images. Such random split were performed in ten separate sessions. Fig. 3 shows the average histograms of MOS and SI values of the train, validation and test images over the ten sessions, illustrating that the images in the split sets are independent and similarly distributed.

It should be noted that different ranges of MOS values are used in the two databases, i.e., [1,5] in KonIQ-10 k and [0,100] in SPAQ. For a cross-database performance evaluation, we have also linearly normalized the MOS values in SPAQ into the range of [1,5]. In addition, as the raw distribution of scores given by individual participants in the KonIQ-10 k database has been published, we have calculated the normalized

Table 2
Average and Standard Deviation of Evaluation Criteria on KonIQ-10 k and SPAQ Test Sets.

Models	KonIQ-10 k test set			SPAQ test set		
	PLCC ↑	SROCC ↑	RMSE ↓	PLCC ↑	SROCC ↑	RMSE ↓
NIQE [7]	0.597 (±0.011)	0.601 (±0.036)	0.527 (±0.136)	0.765 (±0.021)	0.728 (±0.045)	0.582 (±0.094)
IL-NIQE [8]	0.573 (±0.040)	0.552 (±0.029)	0.496 (±0.037)	0.770 (±0.032)	0.708 (±0.107)	0.594 (±0.088)
BRISQUE [11]	0.637 (±0.009)	0.634 (±0.016)	0.419 (±0.058)	0.803 (±0.012)	0.810 (±0.019)	0.501 (±0.062)
DeepBIQ [18]	0.873 (±0.021)	0.864 (±0.037)	0.284 (±0.029)	0.858 (±0.027)	0.861 (±0.028)	0.389 (±0.035)
Koncept512 [20]	0.916 (±0.116)	0.909 (±0.085)	0.267 (±0.094)	0.831 (±0.097)	0.830 (±0.080)	0.384 (±0.060)
CaHDC [21]	0.856 (±0.027)	0.817 (±0.025)	0.370 (±0.041)	0.824 (±0.030)	0.815 (±0.019)	0.486 (±0.068)
MEON [33]	0.704 (±0.136)	0.794 (±0.073)	0.405 (±0.059)	0.683 (±0.088)	0.733 (±0.049)	0.483 (±0.100)
DBCNN [34]	0.856 (±0.027)	0.843 (±0.021)	0.375 (±0.064)	0.894 (±0.024)	0.865 (±0.020)	0.459 (±0.057)
SGDNet [36]	0.868 (±0.027)	0.811 (±0.016)	0.312 (±0.047)	—	—	—
TRIQ [38]	0.922 (±0.018)	0.910 (±0.011)	0.223 (±0.030)	0.916 (±0.027)	**0.925** (±0.015)	**0.324** (±0.021)
hyperIQA [39]	0.916 (±0.030)	0.907 (±0.027)	0.242 (±0.031)	0.910 (±0.026)	0.915 (±0.020)	0.329 (±0.028)
AIHIQnet	**0.932** (±0.012)	**0.919** (±0.019)	**0.207** (±0.012)	—	—	—
AIHIQnet-MOS	0.929 (±0.020)	0.915 (±0.014)	0.209 (±0.022)	**0.929** (±0.022)	**0.925** (±0.019)	0.326 (±0.027)

distribution of quality scores over the quality categories as the predictive object in AIHIQnet. Whereas, only MOS values were published for the SPAQ database. Therefore, the AIHIQnet-MOS model has been used for SPAQ.

B. Model training strategies

NR-IQA models based on classical machine learning, e.g., BRISQUE and DeepBIQ using SVR, NIQE and IL-NIQE based on Gaussian distribution, can achieve consistent performance by following the common approach of grid search to find the best hyperparameters based on the image features derived from the training images. On the other hand, the performance of deep learning models can be heavily affected by the training strategy. We have adopted the original implementations of other compared deep learning models and the training strategies provided by their authors. The original implementations have been slightly adjusted as explained in Table 1, so that the models can take arbitrarily sized images as inputs.

AIHIQnet and AIHIQnet-MOS were first built and compiled by individual loss functions and Adam as the optimizer, and the training process was performed in two phases: pretrain and finetune. In the pretrain phase, we have found that a base learning rate 5e-5 provided the best performance in general. The maximum number of epochs was set to 100. A learning rate scheduler of cosine decay with warmup was employed. The first 10 epochs with linearly increasing learning rate were for warmup, then the base learning rate was held for another 30 epochs, and subsequently, the cosine decay of learning rate was applied until the end of training. Early stop has also been used by monitoring the PLCC value on the validation set. In the finetune phase, the base learning rate was set to 1e-6 with the same scheduler applied. Finally, the best model was determined according to the maximal PLCC value on the validation images produced from the training process.

As the images contained in the SPAQ database have diverse resolutions, we implemented an image generator that can serve a batch of images with same resolution in each training step. The images were shuffled after every epoch. In addition, image augmentation is often employed when training deep learning models to increase the amount of training data and accordingly improve the generalization capability of the trained models. For example, transformation, mixture of multiple images, etc., are used in model training for object detection, image recognition, etc. However, due to the particularity of IQA, most of the image augmentation techniques can potentially affect the perceived quality of augmented images. In this work, we have investigated popular image augmentation techniques implemented in a Python library [60] and gauged if they affected the perceived image quality. Finally, we concluded that only horizontal flip did not influence image quality significantly, and therefore, it was included as an augmentation strategy in our experiments.

The models were trained on two GeForce RTX 3090 GPUs. As some models employ base networks, e.g., ResNet50 in AIHIQnet/AIHIQnet-MOS, SGDNet, TRIQ, InceptionResNet-V2 in Koncept512, publicly available weights pretrained on ImageNet for the base networks were employed for applying transfer learning. Accordingly, all the images were normalized by the preprocessing methods in TensorFlow or PyTorch based on how the base networks were implemented, when pretrained weights have been applied. Otherwise, if a base network is not used, e.g., in CaHDC, the images were still normalized by subtracting the mean and dividing the standard deviation of all images in the training sets.

In addition, the compared models (Koncept512, SGDNet, MEON, DBCNN, CaHDC, TRIQ) were also trained by the proposed strategy above, and we have found that better results were often obtained for some models, compared to the original training strategies. Naturally, the better results are reported in the paper.

C. Comparison experiments

In the training process from each of the ten random train/validation/

Table 3
Average and Standard Deviation of Evaluation Criteria of NR-IQA Models Trained On One Database and Tested On Another Database

Models	Trained on KonIQ-10 k tested on SPAQ			Trained on SPAQ tested on KonIQ-10 k		
	PLCC ↑	SROCC ↑	RMSE ↓	PLCC ↑	SROCC ↑	RMSE ↓
NIQE [7]	0.572 (±0.026)	0.584 (±0.030)	0.746 (±0.105)	0.524 (±0.026)	0.610 (±0.040)	0.691 (±0.086)
IL-NIQE [8]	0.585 (±0.039)	0.602 (±0.044)	0.723 (±0.048)	0.597 (±0.038)	0.626 (±0.035)	0.725 (±0.098)
BRISQUE [11]	0.659 (±0.010)	0.654 (±0.019)	0.650 (±0.049)	0.518 (±0.018)	0.526 (±0.023)	0.684 (±0.120)
DeepBIQ [18]	0.793 (±0.027)	0.807 (±0.030)	0.562 (±0.050)	0.714 (±0.019)	0.712 (±0.030)	0.520 (±0.020)
Koncept512 [20]	0.825 (±0.105)	0.828 (±0.094)	0.726 (±0.110)	0.728 (±0.103)	0.753 (±0.079)	0.528 (±0.130)
CaHDC [21]	0.712 (±0.038)	0.727 (±0.030)	0.595 (±0.051)	0.536 (±0.038)	0.589 (±0.033)	0.870 (±0.082)
MEON [33]	0.693 (±0.060)	0.704 (±0.067)	0.671 (±0.047)	0.648 (±0.065)	0.692 (±0.034)	0.583 (±0.091)
DBCNN [34]	0.686 (±0.039)	0.733 (±0.050)	0.646 (±0.105)	0.677 (±0.040)	0.644 (±0.035)	0.601 (±0.087)
TRIQ [38]	0.853 (±0.017)	**0.865** (±0.020)	0.479 (±0.037)	0.786 (±0.025)	**0.798** (±0.025)	0.463 (±0.035)
hyperIQA [39]	0.832 (±0.046)	0.830 (±0.029)	0.472 (±0.044)	0.743 (±0.036)	0.720 (±0.030)	0.574 (±0.050)
AIHIQnet	**0.864** (±0.019)	**0.865** (±0.013)	**0.458** (±0.036)	—	—	—
AIHIQnet-MOS	0.852 (±0.023)	0.854 (±0.019)	**0.458** (±0.030)	**0.799** (±0.031)	0.791 (±0.024)	**0.455** (±0.0242)

test split sessions, the NR-IQA models with the best weights determined by the maximal PLCC values on the validation set have been applied to the test set to evaluate the performance of individual models. Such approach has been performed on the KonIQ-10 k and SPAQ databases, respectively. Table 2 reports the average values and standard deviations of the evaluation criteria PLCC, SROCC, and RMSE over the ten split sessions on the two databases, respectively. It should be noted that SGDNet requires saliency maps as inputs. Even though the authors of SDGNet also proposed another approach without saliency to predict image quality, we have found that including saliency as input indeed improves the performance. As the authors of SDGNet only published saliency maps for the KonIQ-10 k database, the evaluation of SDGNet was not performed on the SPAQ databases.

Subsequently, in order to evaluate the generalization capability of the NR-IQA models, we have also applied the models trained on one database to another database, e.g., AIHIQnet was trained on the train and validation sets in KonIQ-10 k database and then tested on the entire SPAQ database, and vice versa. The results are reported in Table 3. As no saliency maps are available for the SPAQ database, SGDNet was excluded from the cross-database experiment. It is also noted that AIHIQnet can be trained with KonIQ-10 k database while not on SPAQ, as the distribution of MOS values is required. Thus, only AIHIQnet-MOS trained on SPAQ has been tested with respect to the KonIQ-10 k database.

According to the evaluation results, deep learning driven models generally outperform the representative NR-IQA models (BRISQUE, NIQE, IL-NIQE) based on classical machine learning approach using hand-crafted features. Furthermore, the results demonstrate that the proposed AIHIQnet, as well as AIHIQnet-MOS, shows significantly better performance and generalization capability than the other deep learning driven models with respect to both the two databases. The possible reasons are discussed as follows.

DeepBIQ shows promising performance in NR-IQA, which on one hand confirms that large-scale pretraining indeed provides solid foundation for downstream tasks, e.g., IQA. On the other hand, the employed image patching and averaging approach can be influenced significantly by image resolution, spatially inconsistent quality and unbalanced attention distribution. This is also suspected to be a potential reason why MEON obtains poor performance in our experiments, in which the image resolutions are relatively large.

Koncept512 is a representative of deep learning models widely used in image classification, i.e., using a sufficiently deep CNN architecture (InceptResnet-V2) to extract relevant features and then fully connected layers for quality prediction. This approach achieves promising performance in image classification problems, and such architecture of CNNs followed by FC layers achieves fair performance in IQA. On the other hand, we have also noticed that three dropout layers with relatively large dropping rate (0.5) used in Koncept512 brings strong instability in model training. Similarly, DBCNN also follows the common approach of constructing CNNs for image classification, even though two separate CNNs for synthetic and authentic distortions are combined. However, such approach does not fully exploit the features and hidden characteristics extracted by CNN layers for quality perception. As we will illustrate by activation maps in the next Section, the semantical features can vary significantly from image recognition to IQA.

SGDNet integrates image saliency into IQA by multiplying the saliency map derived externally with an intermediate quality feature map generated by the CNN architecture. It shows fair performance on the KonIQ-10 k database, which also confirms that image saliency can contribute to quality assessment significantly. However, a direct multiplication between a saliency map and the quality feature map as in SGDNet might not be the best approach to fully exploit saliency or attention mechanism in quality assessment, as we have investigated in earlier studies [41]-[43]. Furthermore, as the saliency map is obtained externally, SGDNet cannot learn the importance of saliency on quality perception during the training process, which also potentially impedes the advantage of using saliency in IQA. In this regard, the Transformer architecture can better exploit the attention mechanism for target tasks, and thus, TRIQ shows very promising performance in our experiments. A potential approach to further improve TRIQ is to combine Transformer and hierarchical structure based on CNNs in IQA. On the other hand, the hierarchical structure has been employed in CaHDC, which can take advantage of deep CNN architectures. However, the side pooling nets using simple maximum pooling to combine feature maps at different CNN scales might not sufficiently capture the crucial perceptible information for IQA. As a comparison, hyperIQA using an elaborate architecture based on multi-scale image features has obtained high performance, which demonstrates that image quality perception can be appropriately modelled in a hierarchical fashion. Whereas, the hyperIQA model requires that images should be rescaled to the resolution that has been used in model training, which can potentially introduce bias in image quality prediction.

Table 4
Evaluation results of models trained on combined KonIQ-10 k & SPAQ and tested on combination of CLIVE & CID2013

	PLCC ↑	SROCC ↑	RMSE ↓
NIQE [7]	0.553	0.538	0.921
IL-NIQE [8]	0.525	0.576	0.884
BRISQUE [11]	0.617	0.628	0.803
DeepBIQ [18]	0.786	0.758	0.668
Koncept512 [20]	0.718	0.733	0.694
CaHDC [21]	0.529	0.551	0.830
MEON [33]	0.683	0.706	0.701
DBCNN [34]	0.583	0.682	0.740
TRIQ [38]	0.814	0.820	0.658
hyperIQA [39]	0.795	0.801	0.694
AIHIQnet	0.839	**0.852**	0.637
AIHIQnet-MOS	**0.843**	0.843	**0.628**

AIHIQnet follows the multi-scale perceptual process and benefits from a deep network architecture deriving relevant features. Such hierarchical structure of generating quality feature maps can accurately capture the intrinsic mechanism of IQA. Furthermore, by integrating the perceptually guided attention module into pyramidic feature maps, AIHIQnet can also appropriately simulate the generation of image quality perception in the HVS. In most cases, AIHIQnet obtains better performance than AIHIQnet-MOS, which also supports the prior observation that predicting score distribution provides more robust results than predicting MOS values directly.

On the other hand, the cross-database experiment shows that the models still have limited generalization capability, as indicated by the results in Table 3. This is expectable because the two IQA databases were established in two subjective experiments with different settings and environmental conditions. In addition, the alignment of quality scores in different databases is also a challenging issue. By comparing the cross-database evaluation results, it can be observed that most models trained on the KonIQ-10 k database demonstrate better generalization capability than being trained on SPAQ. We suspect that this is due to that different assessment environments used for the subjective studies. As the SPAQ experiment was performed in a controlled laboratory environment, the results are probably more consistent than the results in a crowdsourcing study, limiting the generalizability of the results.

Furthermore, it is also interesting to evaluate how a deep learning based NR-IQA model trained on large-scale datasets performs on new small-scale datasets. Many subjective IQA experiments have been conducted and the datasets are published, e.g., CLIVE [61], CSIQ [62], TID2013 [63], CID2013 [64]. However, most datasets, e.g., CSIQ and TID2013, were developed for full-reference (FR) IQA purpose, i.e., the reference images were available for viewers . A significant difference between FR IQA and NR IQA is that the former has reference images, and both subjective participants and objective QA models assess the relative quality of a distorted image against its reference, not the absolute perceived quality. Therefore, it might be inappropriate to evaluate NR-

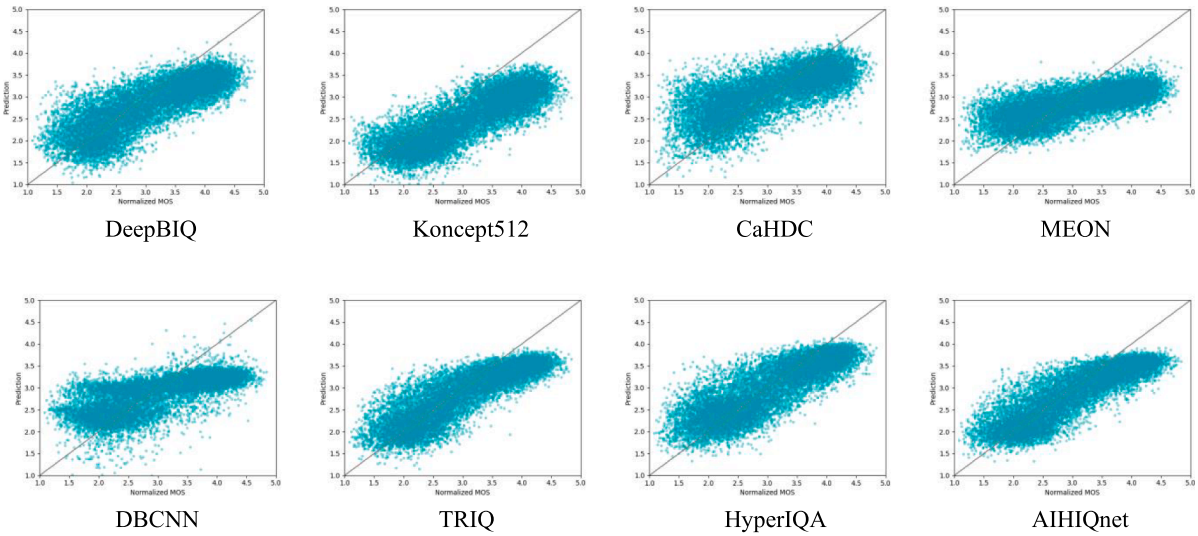

Fig. 4. Scatter plots of different NR-IQA models tested on the SPAQ database, models were trained on KonIQ-10 k database, horizontal axis: ground-truth MOS normalized into [1,5], vertical axis: average of predicted MOS by models trained in the ten split sessions.

IQA models using IQA datasets developed for FR scenario. Consequently, we have trained the models on a combined dataset consisting of KonIQ-10 k and SPAQ, and then evaluated them on the combination of NR datasets CLIVE and CID2013. Please note that this approach might confuse the differences across IQA experiments and datasets, e.g., KonIQ-10 k vs. SPAQ, and CLIVE vs. CID2013, whereas it can roughly demonstrate the strategy of training models on large datasets and testing on small datasets. Table 4 reports the evaluation results of this experiment, demonstrating that the proposed AIHIQnet/AIHIQnet-MOS models have promising generalization capability.

In addition, we have also drawn the scatter plots of MOS values predicted by the trained NR-IQA models against the ground-truth quality assessment. Fig. 4 shows the case of models trained on KonIQ-10 k and tested with respect to the ground-truth MOS on the SPAQ database. The scatter plots illustrate that AIHIQnet, TRIQ and hyperIQA provide more consistent predictions of image quality in general. However, the scatter plots also show that the imbalanced distribution of quality assessment of the KonIQ-10 k database, i.e., most subjective MOS values located in the range of [1.5, 4.0] as shown in Fig. 3, was introduced to the model prediction. Such phenomenon is even severer with the LIVE-FB database. We have trained and evaluated the NR-IQA models with respect to the LIVE-FB database, and almost all the models fail to obtain promising performance with only AIHIQnet-MOS and TRIQ achieving correlation slightly over 0.5 against the ground-truth. This partially demonstrates that an IQA database with balanced distribution of quality scores is crucial for benchmarking image quality models.

D. Activation map visualization

It is important to investigate what AIHIQnet has learned from the training process. A common approach in CNN architectures is to visualize the intermediate activation maps from individual channels [65]. In this experiment, we visualize the activation maps from the trained AIHIQnet model and also compare them with the map derived from the original pretrained backbone network. AIHIQnet trained on the KonIQ-10 k database has been selected for this experiment, and we have chosen the split session that produced the highest PLCC value on the validation set.

The perceived quality feature maps (*PF*) derived in Eq. (11) carry information of image quality perception. Considering that an individual activation map at different scales contains many channels, e.g., 2,048 channels at the highest scale, we performed maximum pooling over channel dimension of *PF* that roughly represents the distribution of perceived information within an image at different scales. A similar approach has also been employed to visualize the attention maps in

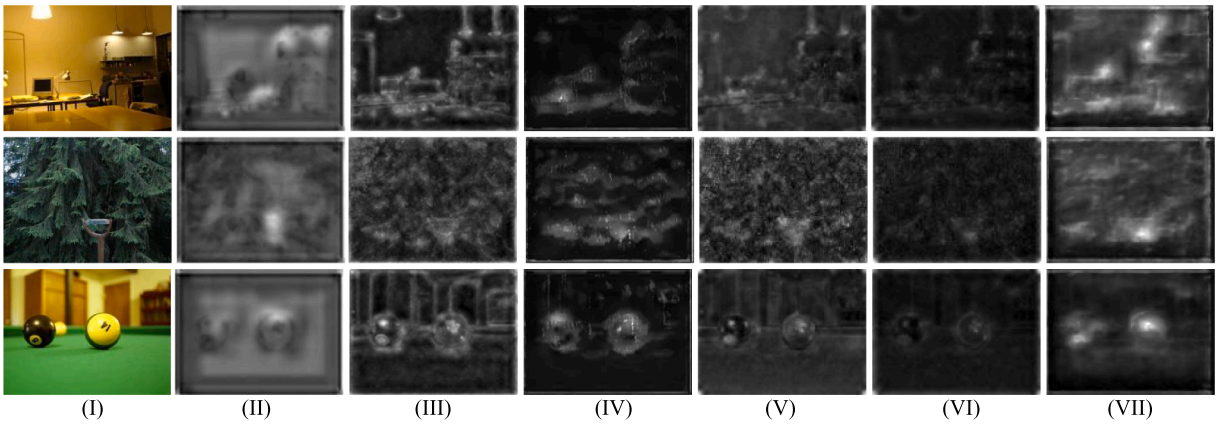

Fig. 5. Example images, activation maps of *PF*, and activation maps from the backbone network from AIHIQnet trained weights and ImageNet pretrained weights. Columns (I): image, (II) backbone map from ResNet50 ImageNet pretrained, (III) backbone map from IQA training with ResNet50 as backbone, (IV) *PF* map with ResNet50 as backbone, (V) backbone map from DenseNet121 ImageNet pretrained, (VI) backbone map from IQA training with DenseNet121 as backbone, (VII) *PF* map with DenseNet121 as backbone.

Table 5
Average of Evaluation Criteria of AIHIQnet and AIHIQnet-MOS in Ablation Experiments

Ablation experiments	AIHIQnet on KonIQ-10 k			AIHIQnet-MOS on SPAQ		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Full model for comparison	*0.932*	*0.919*	*0.207*	*0.929*	*0.925*	*0.326*
1.1) Only channel attention	0.907	0.904	0.227	0.902	0.903	0.358
1.2) Only spatial attention	0.898	0.899	0.234	0.909	0.910	0.352
1.3) No attention module	0.899	0.901	0.233	0.905	0.906	0.361
2) No cross-scale fusion	0.907	0.905	0.226	0.910	0.904	0.349
3) No image augmentation	0.894	0.902	0.221	0.901	0.899	0.360
4) No ImageNet pretrained weights	0.805	0.811	0.263	0.826	0.814	0.405
5) Freeze backbone network	0.789	0.750	0.336	0.801	0.793	0.449
6.1) Vgg16 as backbone	0.918	0.920	0.213	0.915	0.917	0.336
6.2) DenseNet121 as backbone	0.933	0.924	0.209	0.926	0.921	0.331
6.3) ResNet152 as backbone	0.911	0.908	0.228	0.913	0.908	0.339

other attention integrated vision models, e.g., vision Transformer (ViT) [66]. Consequently, the average of *PF* maps over different scales is used to indicate the feature activation map, where the higher-scale maps $\{PF_3, PF_4, PF_5, PF_6\}$ are all up-sampled to match the resolution of the lowest scale map PF_2.

In addition, AIHIQnet has been trained by starting with ImageNet pretrained weights for the backbone network, which has been commonly considered as a promising approach, demonstrated also in our ablation experiment described in the following subsection. However, image content recognition task that the ImageNet pretraining is aimed for, is apparently different from IQA task. Thus, it is interesting to see how the target tasks influence the derived image features. A feature map computed by averaging the activation maps of the backbone networks, i.e., $\{C_2, C_3, C_4, C_5\}$ in Eq. (1), is used to represent the features extracted by the backbone network after the training processing. The first feature map is generated using the trained AIHIQnet, and the second map using the original backbone with ImageNet pretrained weights.

Fig. 5 illustrates three examples randomly chosen from the KonIQ-10 k training set, including the images and activation maps generated by two different backbone networks, namely ResNet50 and DenseNet121. Because the two backbone networks employ different architectures, it is expected that the activation maps (columns II and V) of ImageNet pretrained weights are different. In addition, the two backbone networks show a similar behavior in the IQA task as discussed in the following. By comparing the activation maps before and after training for IQA tasks, i.e., columns II vs. III, and columns V vs. VI, it clearly demonstrates that the features have been dramatically updated for the purpose of IQA, and quality assessment is often driven by smaller regions within an image than the regions in content recognition. Finally, columns IV and VII showing the activation maps of *PF* illustrate that the proposed attention module can accurately capture the most important or attended areas for IQA tasks. According to the activation maps in columns III & IV, and VI & VII, roughly representing image features learned from IQA training process, structural information related to objects and especially those objects of interest (e.g., the lamp and monitor in the first row image, the road mark in the second row image, and the snooker balls in the third row image) are accurately captured by the model. This observation also supports the assumption that image quality perception is heavily influenced by structural information and visual attention mechanism [7,21,27,28]. More in-depth analysis on explaining the AIHIQnet model, e.g., from visual-psychological perspective, will be studied in our next work.

E. Ablation experiments

As AIHIQnet provides a general framework for NR-IQA that can be built on varying components in different ways, it is important to investigate the performance of the components of AIHIQnet(-MOS). We have conducted a series of ablation experiments on the two databases. Naturally, the same training strategy as that employed in the comparison experiments has been used. Table 5 reports the average evaluation results of the ablation experiments over the ten split sessions, and the details about the ablation experiments are discussed below. The original results from the full models, as reported in Table 2, are also included for easy comparison.

1) AIHIQnet(-MOS) with variants of attention module.

Three ablation experiments have been performed: AIHIQnet(-MOS) using only channel attention without the spatial attention block, i.e., Eq. (11) is changed to $PF = CA \otimes P$; AIHIQnet using only spatial attention block, i.e., $PF = SA \otimes P$; and AIHIQnet without the attention module. When not using the attention module, the pyramidic feature maps $\{P_2, P_3, P_4, P_5, P_6\}$ in Eqs. (4) and (5) derived from cross-scale fusion were fed into the head network directly. Sufficiently high accuracy was achieved from the three ablation experiments, which confirms that the proposed architecture consisting of backbone, neck and head networks is able to model image quality perception appropriately. However, integrating both the spatial and channel blocks definitely boosts the performance further, which demonstrates that the mechanisms of selective attention and contrast sensitivity indeed play an important role in IQA.

2) AIHIQnet(-MOS) without using cross-scale fusion.

The feature maps $\{C_2, C_3, C_4, C_5\}$ were fed into the attention module. As C_2, C_3, C_4, and C_5 have different channel dimensions, we have applied a bottleneck convolution layer (K filters, kernel size 1×1 and stride 1) to reduce the channel dimensions to K for generating the perceived feature maps $\{PF'_2, PF'_3, PF'_4, PF'_5\}$. In addition, another bottleneck layer with K filters, kernel size 1×1 and stride 2 was applied to C_5 again to derive another perceived feature map PF'_6. The perceived feature maps $\{PF'_2, PF'_3, PF'_4, PF'_5, PF'_6\}$ have same shapes as $\{PF_2, PF_3, PF_4, PF_5, PF_6\}$ in Eq. (11), and they are fed into the head network, subsequently. The result shows that abandoning the fusion slightly drops the performance of AIHIQnet(-MOS), but not as much as removing the attention module. In other words, integrating the attention and contrast sensitivity mechanisms is more helpful than cross-scale feature fusion in AIHIQnet(-MOS).

3) AIHIQnet(-MOS) without using horizontal flip for augmentation.

In this scenario, training was purely based on the original images without any augmentation. The result confirms that augmentation by horizontal flip indeed improves the performance. We have also evaluated other augmenters, including color changes, contrast enhancement, geometric transforms, and vertical flip. However, these image augmentations do not boost the performance of AIHIQnet(-MOS), and they can even reduce it. We believe this is because inappropriate image augmentations can in fact affect the perceived image quality.

In addition, the backbone network forms a solid base in the proposed architecture, and the number of parameters of ResNet50 backbone is in fact over 70% of the total number of parameters in AIHIQnet(-MOS). Therefore, it is interesting to test the influence of backbone network on AIHIQnet(-MOS), and the following ablation experiments all focus on variants of backbone networks.

4) AIHIQnet(-MOS) without using ImageNet pretrained weights for the backbone network.

The experiments described above were all using the ImageNet pretrained weights to initialize the backbone network. However, ImageNet is mainly for visual recognition, and IQA has not been considered when the backbone networks were trained on ImageNet. As explained in Section IV.D, we also observed that the weights of backbone network can be changed dramatically from the initial weights when training AIHIQnet(-MOS) for IQA. According to the result, AIHIQnet(-MOS)

Table 6
Model Complexity and Inference Time Ratios

	Koncept512	SDGNet	TRIQ	CaHDC	hyperIQA	AIHIQnet
FLOPS	$2.58\cdot10^{11}$	$1.23\cdot10^{11}$	$1.86\cdot10^{11}$	$0.373\cdot 10^{11}$	$1.98\cdot10^{11}$	$2.13\cdot10^{11}$
Inference time	1.307	0.734	0.986	0.612	1.078	1

shows worse performance without using the ImageNet pretrained weights for the backbone network. This confirms that the ImageNet pretrained weights provide an appropriate initial point for a wide range of downstream computer vision tasks, including IQA, even though IQA was not the target task during pretraining. This also suggests that even larger databases dedicated to visual quality assessment will advance development of quality models.

5) AIHIQnet(-MOS) with frozen backbone network.

We also trained AIHIQnet(-MOS) by freezing the backbone network, meaning that the pretrained ImageNet weights for the backbone network were kept for IQA. Even though using the pretrained ImageNet weights as a starting point for transfer learning significantly boosts the performance of IQA models, the result demonstrates that the original weights do not work well for IQA without updating them during the transfer learning process.

6) AIHIQnet(-MOS) with other backbone networks.

Three other backbone networks have been employed to replace ResNet50, including VGG16 [32], DenseNet121 [47] and ResNet152 [45]. VGG16 has a relatively simple architecture, DenseNet121 employs direct connections between any two layers with compact set of parameters and reduced complexity, whereas ResNet152 represents a typical very deep architecture. Similarly, the ImageNet pretrained weights were loaded before transfer learning. The evaluation results demonstrate that the choice of the backbone network is not crucial for the model performance. We believe that the reason is twofold: relatively small datasets (i.e., KonIQ-10 k & SPAQ compared with ImageNet) cannot fully exploit the discriminative capability of deep architectures; and employing the perceptually driven neck and head networks significantly empowers AIHIQnet(-MOS) for modeling image quality perception based on image features that can be extracted by various backbone networks. This experiment suggests that a more dedicated backbone network pretrained on other large databases might improve the performance of AIHIQnet(-MOS) further. This will be studied in the future work.

F. Model complexity

We have also analyzed the model complexity in terms of floating point operations per second (FLOPS). Table 5 reports the FLOPS of those deep learning driven NR-IQA models implemented in Python when uniformly specifying the input image size as $1024 \times 768 \times 3$. AIHIQnet-MOS has almost the same complexity as AIHIQnet. In addition, we have run the trained models on a NVIDIA GeForce RTX 3090 GPU to predict image quality on the entire SPAQ database (batch size = 1) as it consists of diverse image resolutions. Subsequently, the average inference time per image was calculated as the computational time of the models. AIHIQnet takes about 47 ms on average to predict a single image quality, and the ratios of inference time of other models are also reported in Table 6. It should be noted that the inference time of SDGNet was calculated by the ratio with AIHIQnet on the KonIQ-10 k database, as no saliency maps are available for the SPAQ database. As deriving the saliency maps also requires complex computations, the computation time for SDGNet will be significantly increased when including saliency map derivation. The complexity analysis indicates that AIHIQnet is not the fastest IQA model, while the inference time is still acceptable given that accurate image quality prediction is achieved. We assume that the attention module performed on multiple scales contributes heavily to the model complexity, and an efficient optimization will be further studied in the next work.

5. Conclusion

This paper proposed a perception based hierarchical architecture AIHIQnet for NR-IQA, consisting of backbone, neck and head networks. Relevant features that are captured by a general backbone network are filtered by cross-scale fusion and an attention module to produce perceptual quality information, based on which the perceived image quality can be finally predicted following the essential process of quality perception. Two crucial mechanisms in IQA, contrast sensitivity and selective attention, are appropriately modeled in the attention module. Comprehensive evaluation and ablation experiments on publicly available IQA databases demonstrated outstanding performance of AIHIQnet and revealed important characteristics that can be used in future work for further improvement of NR-IQA models.

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.

Acknowledgement

This work is in part supported by the basic grant (Grunnbevilgning) of NORCE funded by the Research Council of Norway, and in part by National Natural Science Foundation of China under Grant 61772348, Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603, and Shenzhen Fundamental Research Program under Grant JCYJ20200109110410133.

References

[1] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (Oct. 2006) 3440–3451.
[2] S.S. Hemami, A.R. Reibman, No-reference image and video quality estimation: Applications and human-motivated design, Signal Process. Image Commun. 25 (7) (2010) 469–481.
[3] L. Meesters, J.-B. Martens, A single-ended blockiness measure for JPEG-coded images, Signal Process. 82 (3) (2002) 369–387.
[4] R. Ferzli, L.J. Karam, A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB), IEEE Trans. Image Process. 18 (4) (2009) 717–728.
[5] H.R. Sheikh, Image quality assessment using natural scene statistics, Ph.D. dissertation, The University of Texas at Austin, USA, 2004.
[6] A.K. Moorthy, A.C. Bovik, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Trans. Image Process. 20 (12) (Apr. 2011) 3350–3364.
[7] A. Mittal, R. Soundararajan, A.C. Bovik, Making a Completely Blind Image Quality Analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.
[8] L. Zhang, L. Zhang, A.C. Bovik, A feature-enriched completely blind image quality evaluator, IEEE Trans. Image Process. 24 (8) (Apr. 2015) 2579–2591.
[9] M.A. Saad, A.C. Bovik, C. Charrier, Blind image quality assessment: A natural scene statistics approach in the DCT domain, IEEE Trans. Image Process. 21 (8) (Aug. 2012) 3339–3352.
[10] D. Ghadiyaram, A.C. Bovik, Perceptual quality prediction on authentically distorted images using a bag of features approach, J. Vis. 17 (1) (Jan. 2017) 1–25.
[11] A.K.M. Mittal, A.C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Trans. Image Process., Dec. 21 (12) (2012) 4695–4708.
[12] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[13] L. Kang, P. Ye, Y. Li, and D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, Jun. 2014.
[14] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, A deep neural network for image quality assessment, in: Proc IEEE. Int. Conf. Image Process. (ICIP), Sep. 2016, Phoenix, AZ, USA.

[15] S. Bosse, D. Maniry, K.-R. Muller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Trans. Image Process. 27 (1) (2018) 206–219.
[16] Y. Li, L-M. Po, L. Feng, and F. Yuan, No-reference image quality assessment with deep convolutional neural networks, in: Proc. IEEE Int. Conf. Digit. Signal Process. (DSP), Oct. 2016, Beijing, China.
[17] F. Gao, J. Yu, S. Zhu, Q. Huang, Q.i. Tian, Blind image quality prediction by exploiting multi-level deep representations, Pattern Recognit. 81 (2018) 432–442.
[18] S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, Signal, Image and Video Process. 12 (2) (2018) 355–362.
[19] J. You, J. Korhonen, Deep neural networks for no-reference video quality assessment, in: Proc. IEEE. Int. Conf. Image Process. (ICIP), Sep. 2019, Taipei, Taiwan.
[20] V. Hosu, H. Lin, T. Sziranyi, D. Saupe, KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment, IEEE Trans. Image Process. 29 (Jan. 2020) 4041–4056.
[21] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multi-scale structural similarity for image quality assessment, in: Proc. IEEE Conf. on Signals, Syst. and Comput., Nov. 2003, Pacific Grove, CA, USA.
[22] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, W. Lin, End-to-end blind image quality prediction with cascaded deep neural network, IEEE Trans. Image Process. 29 (Jun. 2020) 7414–7426.
[23] T-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, Honolulu, HI, USA.
[24] S. Mallet, Wavelets for a vision, Proc. IEEE 84 (4) (Apr. 1996) 604–614.
[25] K. Petras, S. ten Oever, C. Jacobs, V. Goffaux, Coarse-to-fine information integration in human vision, NeuroImage 186 (2019) 103–112.
[26] L. Itti, C. Koch, Computational modelling of visual attention, Nat. Rev. Neurosci. 2 (3) (2001) 194–203.
[27] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, P. Ndjiki-Nya, Visual attention in quality assessment, IEEE Signal Process. Mag. 28 (6) (2011) 50–59.
[28] H. Liu, I. Heynderickx, Visual attention in objective image quality assessment: based on eye-tracking data, IEEE Trans. Circuits Syst. Video Technol. 21 (7) (Jul. 2011) 971–982.
[29] J.G. Robson, Spatial and temporal contrast-sensitivity functions of the visual system, J. Opt. Soc. Am. 56 (8) (1966) 1141–1142.
[30] W.S. Geisler, J.S. Perry, A real-time foveated multi-resolution system for low-bandwidth video communication, in: Proc. SPIE Human Vision Electron. Imaging, vol. 3299, San Jose, CA, USA, Jan. 1998, pp. 294–305.
[31] Krizhevsky, I. Sutskverer, G. Hilton, ImageNet classification with deep convolutional neural networks, in: Proc. Adv. Neural Inf. Process. Syst. (NIPS), Dec. 2012, Lake Tahoe, NV, USA.
[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. Int. Conf. on Learning Representations (ICLR), May 2015, San Diego, CA, USA.
[33] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-end blind image quality assessment using deep neural networks, IEEE Trans. Image Process. 27 (3) (Mar. 2018) 1202–1213.
[34] 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.
[35] Q. Yan, D. Gong, Y. Zhang, Two-stream convolutional networks for blind image quality assessment, IEEE Trans. Image Process. 28 (5) (Nov. 2018) 2200–2211.
[36] S. Yang, Q. Jiang, W. Lin, Y. Wang, SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment, in: Proc. ACM Int. Conf. Multimed. (MM), Oct. 2019, Nice, France.
[37] D. Chen, Y. Wang, W. Gao, No-reference image quality assessment: An attention driven approach, IEEE Trans. Image Process. 29 (May 2020) 6496–6506.
[38] J. You, J. Korhonen, Transformer for image quality assessment, in: Proc IEEE. Int. Conf. Image Process. (ICIP), Sep. 2021, Anchorage, Alaska, USA.
[39] 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: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, Virtual.
[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun.-Jul. 2016, Las Vegas, NV, USA.
[41] J. You, J. Korhonen, A. Perkis, T. Ebrahimi, Balancing attended and global stimuli in perceived video quality assessment, IEEE Trans. Multimed. 13 (6) (Oct. 2011) 1269–1285.
[42] J. You, Video gaze prediction: Minimizing perceptual information loss, in: Proc. IEEE Int. Conf. Multimed. & Expo (ICME), Melbourne, Australia, Jul. 2012.
[43] J. You, T. Ebrahimi, A. Perkis, Attention driven foveated video quality assessment, IEEE Trans. Image Process. 23 (1) (Jan. 2014) 200–213.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in: Proc. Adv. Neural Inf. Process. Syst. (NIPS), Dec. 2017, Long Beach, CA, USA.
[45] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun.-Jul. 2016, Las Vegas, NV, USA.
[46] J. Hu, L. Shen, and G. Sun, Squeeze-and-excitation networks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, Salt Lake City, UT, USA.
[47] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, Honolulu, HI, USA.
[48] TensorFlow applications, available online: https://www.tensorflow.org/api_docs/python/tf/keras/applications.
[49] T. Moore, M. Zirnsak, Neural mechanisms of selective visual attention, Annu. Rev. Psychol. 68 (1) (2017) 47–72.
[50] S. Woo, J. Park, J-Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: Proc. European Conf. Comput. Vis. (ECCV), Sep. 2018, Munich, Germany.
[51] H. Zeng, L. Zhang, A.C. Bovik, Blind image quality assessment with a probabilistic quality representation, in: Proc. IEEE. Int. Conf. Image Process. (ICIP), Oct. 2018, Athens, Greece.
[52] R. Müller, S. Kornblith, G. Hinton, When does label smoothing help? in: Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Dec. 2019, Vancouver, Canada.
[53] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (Sep. 2015) 1904–1916.
[54] P.A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, C. Hervas-Martinez, Ordinal regression methods: Survey and experimental study, IEEE Trans. Knowl. Data Eng. 28 (1) (2016) 127–146.
[55] H. Talebi, P. Milanfar, NIMA: Neural image assessment, IEEE Trans. Image Process. 27 (8) (Apr. 2018) 3998–4011.
[56] W. Zhang, K. Ma, G. Zhai, X. Yang, Uncertainty-aware blind image quality assessment in the laboratory and wild, IEEE Trans. Image Process 30 (Mar. 2021) 3474–3486.
[57] Y. Fang, H. Zhu, Y. Zeng, K. Ma, Z. Wang, Perceptual quality assessment of smartphone photography, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Virtual, Jun. 2020.
[58] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A.C. Bovik, From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), Virtual, Jun. 2020.
[59] ITU-T Recommendation P.910, Subjective video quality assessment methods for multimedia applications, ITU, Apr. 2008.
[60] A.B. Jung, K. Wada, J. Crall, et al., Imgaug, available online: https://github.com/aleju/imgaug.
[61] D. Ghadiyaram, A.C. Bovik, Massive online crowdsourced study of subjective and objective picture quality, IEEE Trans. Image Process 25 (1) (2016) 372–387.
[62] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Digit. Imaging 19 (1) (2010) Mar.
[63] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. Jay Kuo, Image database TID2013: Peculiarities, results and perspectives, Signal Process. Image Commun., Jan. 30 (2015) 57–77.
[64] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, J. Hakkinen, CID2013: A database for evaluating no-reference image quality assessment algorithms, IEEE Trans. Image Process 24 (1) (2015) 390–402.
[65] M.D. Zeiler, R. Fergus, Visualizing and Understanding Convolutional Networks, in: Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, Sep., 2014..
[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: Proc. Int. Conf. on Learning Representations (ICLR), Virtual, May 2021.