Blind image quality assessment via learnable attention-based pooling

Jie Gu\(^a,b\), Gaofeng Meng\(^a,*\), Shiming Xiang\(^a,b\), Chunhong Pan\(^a\)

\(^a\)National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P.R. China
\(^b\)School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, P.R. China

A R T I C L E   I N F O

Keywords:
Image quality assessment
Perceptual image quality
Visual attention
Convolutional neural network
Learnable pooling

A B S T R A C T

Many recent algorithms based on convolutional neural network (CNN) for blind image quality assessment (BIQA) share a common two-stage structure, i.e., local quality measurement followed by global pooling. In this paper, we mainly focus on the pooling stage and propose an attention-based pooling network (APNet) for BIQA. The core idea is to introduce a learnable pooling that can model human visual attention in a data-driven manner. Specifically, the APNet is built by incorporating an attention module and allows for a joint learning of local quality and local weights. It can automatically learn to assign visual weights while generating quality estimations. Moreover, we further introduce a correlation constraint between the estimated local quality and attention weight in the network to regulate the training. The constraint penalizes the case in which the local quality estimation on a region attracting more attention differs a lot from the overall quality score. Experimental results on benchmark databases demonstrate that our APNet achieves state-of-the-art prediction accuracy. By yielding an attention weight map as by-product, our model gives a better interpretability on the learned pooling.

\(\copyright\) 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Digital images are subject to a broad spectrum of distortions during image acquisition, restoration, compression, transmission, and display. The introduced distortions may probably result in a degradation of visual quality. Since human visual system (HVS) is the ultimate “receiver” [1] in most applications, subjective evaluation is a natural and reliable way to quantify visual image quality. However, subjective test is usually cumbersome, expensive and time-consuming, and thus can hardly be used in the practical applications where a real-time and automated quality assessment is needed.

Recently, there has been a growing interest in developing objective image quality assessment (IQA) algorithms for an automated quality evaluation [1,2]. According to the availability of a “distortion-free” reference image, existing objective image quality metrics can be classified as full-reference (FR), reduced-reference (RR) and no-reference/blind (NR/B) methods. FR methods evaluate visual quality by quantifying the differences between the distorted and corresponding reference images. In the past decades, researchers have made many efforts to develop effective FR-IQA metrics according to the known characteristics of the HVS [3–9].

RR approaches are desirable when only partial information of the reference signal is available [10,11]. However, the requirement of a reference image is often problematic in many real-world applications, so an NR/blind IQA (BIQA) model is needed in these cases for quality evaluation. Particularly, general-purpose BIQA methods are more practical as they do not require prior knowledge about distortion types.

Current general-purpose BIQA methods can be broadly classified as opinion-free (a.k.a., opinion-unaware) and opinion-aware models. The former does not require subjective scores for model training. Mittal et al. [12] first developed a “completely blind” NR-IQA model, which does not require prior knowledge of anticipated distortions and is free of learning (i.e., without training on ground-truth scores). Later this metric is extended by Zhang et al. [13] by further introducing three additional types of quality-aware features and executing the quality estimation locally. A training-free referenceless quality metric for 3D synthesized images is developed in [14] with auto-regression based local image description. Unlike the learning-free ones, some other opinion-unaware models access image quality by learning with objective or synthetic scores. For example, Xue et al. [15] proposed to assign each patch an objective score and perform clustering to learn a set of centroids, which then serve as a codebook to infer quality estimations. In [16], unsupervised rank aggregation is applied to combine several FR metrics to generate synthetic scores, which are used as ground-truth scores for BIQA model training. Moreover, opinion-free models for the quality assessment of screen content and enhanced images are
explored in [17] and [18], respectively. These two methods adopt similar labeling strategy in [16] that uses objective scores derived by FR metrics as ground-truth scores during training.

In contrast, opinion-aware models are commonly learning-based with subjective scores given as ground-truth quality estimations. Typically, quality-aware features are first extracted and then a mapping from the extracted features to the quality score is learned [19]. One can see that a satisfactory feature extraction algorithm is the key to the success of opinion-aware BIQA methods. The most widely used features are natural scene statistics (NSS) based features [20–26]. The assumption behind NSS models is that normalized photographic images follow some statistical laws and therefore the perturbations of these statistics can be used to measure distortion levels. As one of the pioneering models, blind image quality index (BIQI) [20] (a two-step framework) uses distorted image statistics. BIQA methods based on Wavelet domain NSS, DCT domain NSS and spatial domain NSS have been presented in [22], [23] and [24], respectively. In [25], Gu et al. combined spatial NSS features with free energy principle based features to build the NFERM metric. Ghiyaram and Bovik [26] proposed a bag of features approach for an accurate quality prediction on authentically distorted images.

Another kind of opinion-aware BIQA methods are based on feature learning, which aims to learn quality aware representations directly from raw images in an unsupervised or supervised manner. Among the former, Ye and Doermann [27] first proposed to encode images with a visual codebook. The quality estimation of a test image is defined as the weighted average of the quality scores of Gabor-filter-based codewords. In their later work, they [28] proposed to learn a dictionary in an unsupervised way and used soft-assignment coding with max pooling to obtain general image representations. In [29], Zhang et al. proposed an approach similar to [28], but with more elaborate features focusing on semantic obviousness. Xie et al. [30] utilized a bag-of-words model for image representation. The codebook is constructed by clustering local quantized pattern descriptors. Supervised feature learning has also been adopted by many BIQA methods. In [31], Ye et al. developed an efficient and cross-domain model based on filter learning, which can be considered as a supervised extension of their previous work [28]. In recent years, deep learning has achieved wide attention and adoption in a variety of vision tasks. Such a success encourages researchers to explore the potential of applying deep networks to BIQA. A brief review of the representative methods [32–40] is given in Section 2.1.

In the literature, many convolutional neural network (CNN) based BIQA methods share a common two-stage framework [35,39–41], i.e., local quality evaluation followed by weighted pooling (score aggregation). Generally, image quality/distortion is first measured locally by a CNN and then a pooling strategy combines the local quality estimations using a weighting function.

Benefiting from the representative power of CNN, significant progress has been made in evaluating local quality. However, usually the pooling stage is still done in simplistic and heuristic ways in existing methods, based on some prior knowledge about HVS. For example, inspired by the visual attention mechanism, saliency- and object-based pooling have been widely adopted in IQA, where the weights are generated by saliency prediction or generic object detection methods. But such a pooling strategy is often less effective on distorted images, since most saliency prediction and object detection models are learned on natural (distortion-free) images and thus are sensitive to distortions.

This paper presents an attention-based pooling network (APNet), focusing on developing a built-in learnable attention for pooling in BIQA. Specifically, the APNet is built by incorporating an attention module into the common classification network that can automatically learn to assign visual weights. It allows for a joint learning of local quality and local weight. During training, a correlation constraint between the local quality estimations and attention weights is further introduced to our APNet for a better model training. The constraint is designed to penalize the case that the estimated scores of those regions that attract more attention differ a lot from the overall quality score. It is noteworthy that our attention-based pooling is actually learned in a weakly supervised manner. Unlike previous works, the pooling model is learned merely from subjective quality scores, without accessing to attention/saliency supervision.

Our contributions can be summarized as follows.

- We aim to achieve a learnable attention-based pooling that can benefit the IQA performance. In this process, a network is developed to jointly estimate the local quality and execute the global pooling. The proposed model can achieve state-of-the-art correlation with human subjective scores on both authentically and synthetically distorted images.
- We introduce a correlation constraint between the local quality estimations and attention weights to regulate the model training. With this constraint, an effective and interpretable pooling strategy can be learned in a data-driven manner.
- Our learned attention-based pooling model is algorithm-independent. It can be readily applied to other IQA metrics in the pooling stage to help improve their performance.
- More interestingly, even without attention supervision during training, our method still has the ability to figure out “where” the model focuses on during image quality assessment. The visualization of weight maps shows that the learned attention generally puts more emphasis on the object-like or textural regions, which is consistent with human intuition.

To the best of our knowledge, our work is the first attempt to learn an interpretable attention merely from subjective quality scores to help improve the overall IQA performance.

2. Related work

We briefly introduce related works in this section, including a review of network based BIQA methods and the pooling strategies in previous IQA works.

2.1. Neural networks for BIQA

In recent years, deep learning has made tremendous progress in many computer vision tasks, e.g., clustering [42], image processing [43], image classification [44], detection [45] and segmentation [46]. Accordingly, researchers have shown an increasing interest in developing deep learning based BIQA algorithms.

Deep belief networks (DBNs) [47] are hybrid models with several directed layers and a single undirected layer (typically at the top of the nets). Generally, DBNs are used to generate complex feature representations for quality estimation based on pre-extracted features. For example, Ghiyaram and Bovik [32] fed hybrid NSS features to the network and trained a support vector regressor (SVR) to predict image quality. Tang et al. [33] took the features presented in their previous work [21] as network inputs. In [34], BIQA is formulated as a classification problem, and the NSS features are extracted in wavelet domain.

CNN is first applied to the BIQA task in [35], where a shallow network is trained on image patches with a small size, and the average of the predictions is reported. The authors further proposed a multi-task CNN [36] for a better performance, which can be seen as an extension of [35]. Ma et al. [38] focused on retargeted NR-IQA and proposed a pairwise rank learning approach. They explored different kinds of image representations, including VGG features
[48], for the purpose of retargeted IQA. Kim and Lee [40] developed a BIQA method by imitating FR-IQA behavior, which aims to resolve the absence of local ground-truth targets. To realize image-based training with multiple cropped patches, Lu et al. [37] designed a multi-patch aggregation strategy based on two basic network structures, i.e., statistics and fully-connected sorting. However, the presented aggregation strategy can hardly learn the behavior of the HVS as neither the statistics nor the fully-connected sorting structure is designed from a human-oriented perspective.

In [39], Boose et al. explored the use of an attention model in networks for FR- and NR-IQA. However, their method is less effective in the NR-IQA setting. Actually the results reported in [39] indicate that only incorporating an attention model into the network for learning a pooling strategy may even decrease the prediction performance in NR-IQA.

2.2. Weighted pooling strategies in IQA

Many IQA (especially FR-IQA) methods consist of two independent modules: local quality prediction and pooling, i.e., score aggregation. In the pooling stage, combining local quality estimations with subjective visual fixation data is intuitive and technically sound [49,50]. However, since the subjective data is not available in practical applications, a computational pooling strategy is required for an automatic quality assessment.

In [4], Wang and Li divided the existing pooling strategies into four categories: Minkowski pooling, local distortion based pooling, saliency-based pooling and object-based pooling. The motivation behind saliency-based and object-based pooling is that visual attention is driven by certain low-level salient features and high-level cognitive factors like semantic information [51]. The pooling weights, in this case, may be generated by saliency models [752] or object detection algorithms [41]. Local distortion based pooling strategies generally put more emphasis on highly distorted regions. This is supported by the observation that viewing task affects top-down attention [51], hence people tend to focus on the distorted regions (if observable) when judging image quality [53]. In practice, the pooling can be implemented with a nonuniform weighting approach, e.g., based on microstructural similarity [54] or considering distributions of distortion position, distortion intensity, frequency changes and histogram changes [55]. Recently, Kim and Lee [56] realized a local distortion based pooling purely learned from data, but this learning-based method is only designed for FR-IQA. Achieving a learnable pooling in the BIQA setting is more challenging without information from references.

In this study, we attempt to learn an effective and interpretable pooling strategy for BIQA in a data-driven manner by integrating a constrained attention model into a CNN. Experimental results on both authentically and synthetically distorted images show that the learned score pooling strategy tends to be object-based.

3. BIQA via learnable attention-based pooling

3.1. Basic principle of learnable pooling

Image quality prediction can always be conducted by fine-tuning a pre-trained classification network. In this process, a global average pooling has to be appended to the local quality estimation layer to obtain the overall quality score. Specifically, let \( X \) be the extracted feature map with size of \( H \times W \times C \), where \( H \), \( W \) and \( C \) are the height, width and channel dimension, respectively. Denote \( x_{ij} \) the feature vector at the location \((i, j)\) in \( X \) with size of \( 1 \times 1 \times C \). Let \( \omega \) be the weight vector with the same size as \( x_{ij} \) for the quality estimation mapping. The estimated overall quality score of an input image, denoted as \( \hat{y} \), can be formulated as

\[
\hat{y} = \frac{1}{HW} \sum_{i,j} \left( x_{ij} \cdot \omega \right),
\]

where the operator “ \( \cdot \) “ represents dot product between vectors.

The predicted score \( \hat{y} \) in (1) is actually an average pooling of local quality estimations. That is, convolving the feature map \( X \) with the filter \( \omega \) gives a local quality map, where the pixel value \( x_{ij} \cdot \omega \) can be viewed as the quality estimation of the receptive field of the feature vector \( x_{ij} \). The overall score \( \hat{y} \) is the average of the quality estimations over the local quality map. However, the average pooling is inconsistent with our current understanding about the HVS because not every region in an image draws the same amount of attention of viewers [51,53]. Many previous works on visual attention have shown that humans always focus on some particular regions, e.g., salient objects, when looking at an image [51,57].

A common way to improve the pooling in many quality metrics is to assign weights with saliency prediction or generic object detection methods. But since the saliency prediction models are usually learned on natural distortion-free images, it is difficult to assure the best performance when generalizing them to the quality pooling on distorted ones. This motivates us to find a way to enable a learnable pooling on the IQA data, with only distorted images and subjective quality scores. To achieve this, we explore the use of an attention module and propose an attention-based pooling network (APNet).

Generally speaking, we attempt to learn an attention-based pooling strategy that assigns a positive weight \( \alpha_{ij} \) to each location \((i, j)\) in \( X \). The weight \( \alpha_{ij} \) can be explained as the contribution or importance of the location \((i, j)\) in combining the local quality estimations together. The overview of the proposed method is illustrated in Fig. 1. Our APNet consists of two branches, generating a local quality map and an attention weight map, respectively. In practice, we use the “soft” attention [58] model to compute the weight \( \alpha_{ij} \), i.e.,

\[
\begin{align*}
\alpha_{ij} & = g(X) \\
\alpha_{ij} & = \frac{e^{\hat{y}_{ij}}}{\sum_{i,j} e^{\hat{y}_{ij}}} ,
\end{align*}
\]

where \( g(\cdot) \) is a mapping from the features to the attention weights. Applying the weights (which sum to 1) to (1) gives

\[
\hat{y} = \sum_{i,j} \alpha_{ij} (x_{ij} \cdot \omega) .
\]

Different from many previous studies (e.g., [4,7,41]), the pooling in our method is integrated into the network by employing an attention module. It is learned in a data-driven manner, without any prior knowledge and assumptions of the HVS. Moreover, unlike many previous IQA works that assign weights to the pixels or small patches of input images, the attention weights in our APNet are assigned to the receptive fields of feature vectors. Specifically, the weight \( \alpha_{ij} \) corresponds to the feature vector \( x_{ij} \) in pooling, and can be viewed as the visual importance of the receptive field of this feature vector in overall quality assessment. A large \( \alpha_{ij} \) means that the local area, i.e., the receptive field of \( x_{ij} \), is likely to attract visual attention.

In practice, however, we found that a pooling that corresponds well to visual attention can hardly be learned by only using quality score regression. To address this, we further introduce a constraint on the correlation between the quality estimations and attention weights in the network to assist the training. We notice that the introduced correlation constraint can not only benefit the performance, but also improve the interpretability of the learned pooling. An example is given in Fig. 1(b) and (c), which are the attention
weight maps generated by the APNet trained with or without the constraint, respectively. One can see that with the constraint, the learned pooling would pay more attention to object-like or textural regions. The detailed descriptions about the correlation constraint are given in the next section.

3.2. Correlation constraint to regulate training

The objective function of training our APNet can be defined as a quality score regression loss, i.e.,

\[
L = \left( \sum_{i,j} \alpha_{ij}(x_{ij} \cdot \omega) - y \right)^2 .
\]  

(4)

where \( y \) is the ground-truth quality score. For simplicity, the above objective function contains only one training image.

However, we found that with only the score regression loss (4), the learned pooling is inconsistent with human intuition. It fails to give more weight to those image regions that are likely to be attended by humans. We then explore to further introduce a correlation constraint between the quality estimations and attention weights in our APNet to regulate the model training.

The correlation constraint comes from an intuitive observation. Specifically, due to the strong interaction between visual attention and quality perception [51], the quality of those regions that attract more visual attention would have a relatively large impact on the overall perceived quality. That is to say, it is unlikely that the quality evaluations of these regions differ a lot from the image level quality score. In practice, the constraint can be implemented as a regularization term applied to the score regression loss, which is formulated as

\[
\min \sum_{i,j} \alpha_{ij}(x_{ij} \cdot \omega - y)^2 .
\]  

(5)

If the receptive field of \( x_{ij} \) attracts more visual attention (i.e., \( \alpha_{ij} \) is large), the regularization would penalize the case that the estimated local score \( x_{ij} \cdot \omega \) is very different from the global score \( y \). Moreover, note that a small score difference \( (x_{ij} \cdot \omega - y)^2 \) is not required when the region is less likely to be attended (i.e., \( \alpha_{ij} \) is small), since the impact on the overall quality is limited in this case. Accordingly, the overall objective function of our model is

\[
L = \lambda_1 \left( \sum_{i,j} \alpha_{ij}(x_{ij} \cdot \omega) - y \right)^2 + \lambda_2 \sum_{i,j} \alpha_{ij}(x_{ij} \cdot \omega - y)^2 ,
\]  

(6)

where the ratio of the loss weights \( \lambda_1/\lambda_2 \) is a hyper-parameter. We minimize the objective function (6) over the training set.

3.3. APNet with vector regression

Previous BIQA methods are commonly developed within a single regression framework. A group of features are first extracted and then a regression model (e.g., SVR or neural network) maps the features to a quality score. Recently, Gu et al. [41,52] explored the uncertainty in quality assessment and proposed a vector regression framework to improve the quality prediction accuracy. In practice, we notice that the performance of our APNet can also benefit from the vector regression.

The basic idea of the vector regression is to learn a mapping from the features to a vector of pre-defined belief scores, rather than a single quality score (typically mean opinion score (MOS)) as in previous ones. Specifically, Gu et al. divided the continuous scoring scale into several quality intervals, and designed the belief scores to implicitly measure the probabilities of an input image being assigned to these intervals by the population. The belief score of the \( k \)th quality interval, denoted as \( s_k \), is defined as

\[
s_k = y - \mu_k \quad (k = 1, 2, \ldots, K) ,
\]  

(7)

where \( y \) is the MOS of an image, \( K \) is the number of the quality intervals, and \( \mu_k \) is the center of the \( k \)th pre-defined interval.

Within the vector regression framework, our APNet is implemented by replacing \( \omega \) with more filters, denoted as \( \omega_k \) \( (k = 1, 2, \ldots, K) \), for mapping the feature vectors to the corresponding belief scores. Accordingly, convolving \( X \) with \( \omega_k \) \( k = 1, 2, \ldots, K \), generates multiple local belief score maps (instead of a single local quality map), and the network finally outputs a \( K \)-dimensional vector of the overall belief score estimations. Applying the belief
scores to (6) gives the final objective function of our APNet, i.e.,
\[
L = \lambda_1 \cdot \sum_k \left[ \sum_{i,j} \alpha_{ij} (x_{ij} \cdot \omega_k) - s_k \right]^2 \\
+ \lambda_2 \cdot \sum_{i,j} (x_{ij} \cdot \omega_k - s_k)^2,
\]
(8)
where \(s_k\) is the \(k\)th ground-truth belief score of the input image. During testing, the predicted \(k\)th overall belief score is \(\sum_{i,j} \alpha_{ij} (x_{ij} \cdot \omega_k)\), and thus the overall quality score can be expressed as \(\frac{1}{N} \sum_k \left[ \sum_{i,j} \alpha_{ij} (x_{ij} \cdot \omega_k) + \mu_k \right]\).

3.4. Implementation details

Training the proposed APNet consists of two steps. ResNet [59] (50 layers) is chosen as the basic network due to its successful adaptation on a number of visual tasks. We first fine-tune the ResNet with vector regression on the BIQA task as in [41]. Then we build the APNet by replacing the global average pooling in ResNet with the attention module and continue the network training. The weights of the APNet are initialized as the weights of the fine-tuned ResNet. More training configurations and details are given in the following sections.

Network architecture: The mapping \(g(\cdot)\) in the soft attention model is implemented by using a small network, which contains two convolutional layers with 8 kernels of size \(3 \times 3 \times 2048\) and 1 kernel of size \(1 \times 1 \times 8\), respectively. The convolutional stride and padding are both set as 1 to preserve the spatial resolution.

Details of fine-tuning with vector regression: We first adapt the ResNet-50 to the BIQA task by fine-tuning with vector regression. The size of the input image is fixed to \(224 \times 224 \times 3\), and the length of the belief vector \(K = 7\) as set in [41]. We substitute the softmax layer in ResNet with a 5-D regression layer and then fine-tune the whole network by using stochastic gradient descent (SGD) with a momentum of 0.9, weight decay of 0.0001, and mini-batch size of 8 examples. The fine-tuning process is performed for roughly 10 epoches. The learning rate is initially set as 0.001 and decreased by a factor of 0.1 for two times. The settings of the quality interval centers for computing belief scores, namely \(\mu_k\) (\(k = 1, 2, \ldots, K\)) in (7), are the same as in [41].

Details of training attention-based pooling network: After fine-tuning with vector regression, we continue SGD training of the parameters of the attention model, the score mapping \(\omega_k (k = 1, 2, \ldots, K)\), and the last three “bottleneck” blocks in the network. The input size is fixed to \(384 \times 384 \times 3\) to enlarge the size of the feature map \(X\). The training configurations are basically the same as those in the fine-tuning process. Specifically, we start SGD at a learning rate of 0.001 and decrease it by a factor of 0.1. The training is performed for about 10 epoches and the learning rate is decreased about two times. The momentum, weight decay and batch size are set as 0.9, 0.0001 and 8 respectively. The hyper-parameter \(\lambda_1/\lambda_2\) in (8) is set as 10 in the experiments.

Generation of training data: We collect training examples (with size of \(224 \times 224 \times 3\) or \(384 \times 384 \times 3\)) by cropping image patches from distorted images. For training we assign each patch a quality score equal to the MOS of its source image, and then the ground-truth belief scores can be computed by (7). Note that such a data augmentation and quality assignment strategy could be acceptable in our work since we crop patches with a large size, which helps reduce the quality differences between the cropped patches and the source image.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source images</th>
<th>Distorted images</th>
<th>Distortion types</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIVE [60]</td>
<td>29</td>
<td>779</td>
<td>5</td>
</tr>
<tr>
<td>TID2008 [63]</td>
<td>25</td>
<td>1700</td>
<td>17</td>
</tr>
<tr>
<td>TID2013 [64]</td>
<td>25</td>
<td>3000</td>
<td>24</td>
</tr>
<tr>
<td>CSIQ [61]</td>
<td>30</td>
<td>886</td>
<td>6</td>
</tr>
<tr>
<td>MICT/LCD [62]</td>
<td>14</td>
<td>168</td>
<td>2</td>
</tr>
<tr>
<td>LIVEMD [65]</td>
<td>15</td>
<td>405</td>
<td>3</td>
</tr>
<tr>
<td>MDDID2013 [66]</td>
<td>12</td>
<td>324</td>
<td>3</td>
</tr>
<tr>
<td>LIVEC [67]</td>
<td>–</td>
<td>1162</td>
<td>–</td>
</tr>
</tbody>
</table>

4. Experiments and discussions

In this part, we conduct six experiments on eight commonly used databases to evaluate the performance of the proposed method and discuss several performance issues. Specifically, the experiments on both authentically and synthetically distorted databases aim to validate how the objective quality estimations correspond to the subjective assessments. The cross database evaluation verifies the generalization ability. The weights for pooling generated by the attention module are visualized to intuitively study the relevance of the learned pooling strategy to the visual attention.

4.1. Experimental protocol

Datasets: A total of eight IQA databases are used to evaluate our method. Five of them, namely LIVE [60], CSIQ [61], MICT/LCD [62], TID2008 [63] and TID2013 [64], are traditional synthetically distorted databases that contain images impaired by a single distortion. The LIVEMD [65] and MDDID2013 [66] databases are composed of synthetically distorted images with multiple distortions. The LIVE in the Wild Image Quality Challenge Database (LIVEC) [67] is recently presented, which contains authentically distorted images captured by a wide variety of mobile devices. The characteristics of these eight databases are summarized in Table 1. Note that there are no “distortion-free” reference images in the LIVEC database, and thus both FR- and RR-IQA methods are not applicable in this case.

Evaluation criteria: In our experiments, the performance of IQA algorithms is evaluated by two widely used criteria, namely linear correlation coefficient (LCC) and Spearman rank order correlation coefficient (SROCC), which measure the linear dependence and the monotonicity between objective scores and subjective scores, respectively. Following the suggestions by Sheikh et al. [60], a nonlinear mapping from objective scores to subjective ones is performed before computing the LCC metric, which is a logistic function

\[
f(x) = \beta_1 \left( \frac{1}{2} - \frac{1}{1 + \exp \left( \beta_2 (x - \beta_3) \right)} \right) + \beta_4 x + \beta_5.
\]
(9)
where \(x\) is the predicted objective score and \(\beta_i, i = 1, 2, \ldots, 5\), are the parameters to be fitted.

4.2. Evaluation on authentic distortions

IQA on realistic distortions has attracted many attention in recent years. Authentically distorted images usually contain diverse distortion types and mixtures that are introduced during image acquisition and processing. It is a challenging task for the difficulty of modeling the complex mixtures of distortions and highly varied lighting conditions, e.g., low-light blur and noise.

We first validate the performance of our approach on the LIVEC [67] data-base, which contains a wide variety of image contents
including pictures of people, animals, nature scenes and distinct foreground objects. The experimental configurations are the same as in [26,41]. We randomly select 80% distorted images for training and the rest 20% for testing. To reduce the influence of random selection, the training-testing split is repeated 20 times and the median is reported. The prediction performance measured by LCC and SROCC is listed in Table 2, where the best model is highlighted in boldface. The result of WaDiQAqM [39] is taken from its original paper, and the other results of the competing methods are reported by ourselves. The source codes of these methods, namely BRISQUE [24], CORNIA [28], NFERM [25] and FRIQUEE [26], are obtained from original authors except for CNN [35], which is implemented by ourselves.

To demonstrate the effectiveness of the attention module for score aggregation, we use the ResNet after fine-tuning on the LIVEC database (i.e., the network for initializing APNet) as the baseline. During testing, we randomly select 50 and 10 patches for the baseline and our APNet, respectively. The global quality estimation is the average score of the selected patches. From Table 2, one can observe that our approach outperforms the other competitors, and the attention module indeed helps improve the performance.

We then visualize the attention weights ($\alpha_{ij}$) to study the relevance of the learned pooling strategy to the visual attention. Some examples are shown in Fig. 2, where the weight maps have been resized to the same size as the input images (up-sampling 32 times using bilinear interpolation in this case). In the weight maps, brighter pixels indicate larger values. All these images are selected from the testing set in one training-testing iteration. It can be seen from Fig. 2 that the learned pooling strategy is broadly object-based. More weights are assigned to the regions that contain more semantic contents, which agrees with the observation that objects are likely to be noticed due to the top-down visual attention mechanism.

4.3 Evaluation on synthetically distorted databases

4.3.1 Evaluation on singly distorted images

In this part, we compare the performance of our approach against other top-performing BIQA methods on singly distorted images. Three standard benchmark databases are used, i.e., LIVE [60], CSIQ [61] and TID2008 [63]. To ensure a fair comparison, the same experimental configurations as in previous works are adopted. Specifically, for LIVE and CSIQ, we use the entire databases and do not exclude any types of distortions. For TID2008, we conduct experiments on the first thirteen distortions as in [28,35,40]. The other four distortions (non eccentricity pattern noise, local blockwise distortions of different intensity, intensity shift and contrast change) that are very inhomogeneous are excluded [28,40].

The main difficulty of applying our method to the legacy databases is that the image contents are limited in these datasets. As listed in Table 1, there are only 29, 30 and 25 reference images in the LIVE, CSIQ and TID2008 databases, respectively. As a result, it is hard to learn a good attention model based only on these databases. We address this issue by adapting the APNet trained on authentically distorted images into the synthetic ones. Specifically, after pre-training the attention branch of the APNet on the LIVEC database, we continue SGD training of the parameters of the whole network on LIVE, CSIQ and TID2008. The training configurations are basically the same as those listed in the Section 3.4, including the number of training iterations, the settings of the learning rate, momentum, weight decay and batch size. During testing, the image level quality score is reported as the average of the estimated scores of 10 randomly selected image patches. The input size of the network is fixed to $384 \times 384 \times 3$ for the LIVE and CSIQ databases, while $224 \times 224 \times 3$ for TID2008 because the image resolution is relatively small ($512 \times 384$) in this case.

Table 3 lists the comparisons, where the best BIQA method is highlighted in boldface. The results are reported as the medians across 20 training-testing iterations. For each iteration, we randomly divide the distorted images according to their reference images into two groups, 80% for training and 20% for testing. This is to ensure that no image contents used in testing have been seen by the models during training. We compare our method with three FR-IQA methods, i.e., SSIM [3], FSIQ [5] and VSI [7], and seven BIQA methods, i.e., BRISQUE [24], CORNIA [28], CNN [35], NFERM [25], FRIQUEE [26], WaDiQAqM [39] and BIECON [40]. All the results of these methods are reported by the original authors or implemented by ourselves with the source codes.

It can be seen from Table 3 that our APNet achieves state-of-the-art results on these three synthetically distorted databases. It delivers a higher performance than other competitors on LIVE and CSIQ, and achieves a comparable performance to the state of the art on TID2008. In general, one can observe from Tables 2 and 3 that APNet works consistently well on all the benchmark databases. The competing BIQA models, by contrast, may work well on some databases but fail to achieve good results on other cases. For example, BIECON shows a superior performance on TID2008, but performs relatively poorly on CSIQ. Moreover, it cannot be used for the quality assessment on realistic distortions, since there are no “distortion-free” reference images in this case for model training. WaDiQAqM is also less effective on the challenging LIVEC database due to the difficulty of modeling inhomogeneous realistic distortions.

To intuitively validate the attention model learned on the synthetically distorted databases, we visualize the weight maps ($\alpha_{ij}$) in Figs. 3–5, where the images are selected from LIVE, CSIQ and TID2008, respectively. For each database, two reference images and their distorted versions with four different distortion types are selected from the testing set. Specifically, the distorted images

---

**Table 2** Performance comparison on the LIVEC database. The baseline is the ResNet after fine-tuning on the LIVEC database (namely the network without the attention module for initializing our APNet).

<table>
<thead>
<tr>
<th>Metric</th>
<th>BRISQUE</th>
<th>CORNIA</th>
<th>CNN</th>
<th>NFERM</th>
<th>WaDiQAqM</th>
<th>FRIQUEE</th>
<th>baseline</th>
<th>APNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCC</td>
<td>0.603</td>
<td>0.635</td>
<td>0.566</td>
<td>0.609</td>
<td>0.860</td>
<td>0.712</td>
<td>0.867</td>
<td>0.879</td>
</tr>
<tr>
<td>SROCC</td>
<td>0.581</td>
<td>0.622</td>
<td>0.576</td>
<td>0.586</td>
<td>0.671</td>
<td>0.694</td>
<td>0.849</td>
<td>0.859</td>
</tr>
</tbody>
</table>

---

**Table 3** Performance comparison on three synthetically distorted databases (the images are all singly distorted).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LIVE</th>
<th>CSIQ</th>
<th>TID2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>SROCC LCC</td>
<td>SROCC LCC</td>
<td>SROCC LCC</td>
</tr>
<tr>
<td>SSIM [3]</td>
<td>0.948 0.945 0.876</td>
<td>0.861 0.878 0.857</td>
<td>0.945 0.941 0.876</td>
</tr>
<tr>
<td>FSIQ [5]</td>
<td>0.965 0.961 0.931</td>
<td>0.919 0.926 0.913</td>
<td>0.965 0.962 0.931</td>
</tr>
<tr>
<td>VSI [7]</td>
<td>0.952 0.948 0.942</td>
<td>0.928 0.917 0.891</td>
<td>0.952 0.948 0.942</td>
</tr>
<tr>
<td>BRISQUE [24]</td>
<td>0.940 0.942 0.725</td>
<td>0.791 0.768 0.795</td>
<td>0.940 0.942 0.725</td>
</tr>
<tr>
<td>CORNIA [28]</td>
<td>0.942 0.935 0.713</td>
<td>0.775 0.813 0.837</td>
<td>0.942 0.935 0.713</td>
</tr>
<tr>
<td>CNN [35]</td>
<td>0.956 0.953 0. -</td>
<td>0.862 0.873 0.</td>
<td>0.956 0.953 0.</td>
</tr>
<tr>
<td>NFERM [25]</td>
<td>0.941 0.946 0.767</td>
<td>0.817 0.842 0.849</td>
<td>0.941 0.946 0.767</td>
</tr>
<tr>
<td>FRIQUEE [26]</td>
<td>0.947 0.962 0.859</td>
<td>0.871 0.851 0.856</td>
<td>0.947 0.962 0.859</td>
</tr>
<tr>
<td>WaDiQAqM [39]</td>
<td>0.954 0.963 0. -</td>
<td>0.923 0.901 0.</td>
<td>0.954 0.963 0.</td>
</tr>
<tr>
<td>BIECON [40]</td>
<td>0.961 0.962 0.825</td>
<td>0.838 0.923 0.</td>
<td>0.961 0.962 0.825</td>
</tr>
<tr>
<td>APNet</td>
<td>0.967 0.970 0.905</td>
<td>0.922 0.887 0.901</td>
<td>0.967 0.970 0.905</td>
</tr>
</tbody>
</table>
Fig. 2. Illustration of the attention weight map generated by the learned attention model. The weight maps have been resized to the same size as the input images for visualization. The brighter pixel indicates larger weight.

Fig. 3. Illustration of the attention weight map generated by the attention model learned on LIVE. The first to the fourth rows exhibit the images impaired by JP2K, JPEG, WN and GB, as well as their corresponding weight maps, respectively. Two reference images are included, and for each reference image and each distortion type, we select two distorted versions at the first and the last degradation levels (with the best and the worst quality) for visualization.
subjecting to JP2K, JPEG, WN and GB, as well as their corresponding weight maps, are shown in the first to the fourth rows in Figs. 3–5 respectively. For each reference image and each type of distortion, we select two images that have the best and the worst visual quality for illustration. One can observe that our method tends to focus on object-like or textural regions, e.g., the plane and the fisherman in Fig. 3, the geckos in Fig. 4 and the rowboat or parts of the rowboat in Fig. 5. Moreover, we can notice that the weight maps are observably affected by the distortions. Generally, as the degree of distortion increases, our model would take more informative regions into consideration (e.g., more distorted regions as in the case of WN), and the weight map tends to be smoother.

4.3.2. Evaluation on multiply distorted images

We also validate the performance of our approach on synthetically distorted databases that contain images impaired by multiple distortions. LIVEMD [65] and MDID2013 [66] are two representative databases, and thus are used in our experiments. Due to the limited data, we do not partition the databases into two sub-sets for training and testing. Instead, the experiment is conducted by first training on LIVEMD and then testing on MDID2013. The training/testing configurations, as well as the parameter settings, are the same as those in the experiment on LIVE. Specifically, with the attention branch of APNet pre-trained on LIVE, we continue SGD training of the parameters of whole network on the LIVEMD database. During testing on MDID2013, ten patches are randomly selected for an image, and the global quality estimation is the average score of the selected patches.

The results are listed in Table 4. Three FR-IQA methods, i.e., SSIM [3], VSI [7] and PSIM [54], and two BIQA methods, i.e., BRISQUE [24] and SISBLIM [66], are used for comparison. The results of the five competitors are taken from Gu et al. [54,66]. We have two main observations from Table 4. First, our model outperforms the two BIQA competitors, e.g., about 5% SROCC improvements over SISBLIM, which is a state-of-the-art method in the IQA of multiple distortions. Second, although our APNet is a blind metric, it is still superior to the powerful FR-IQA methods SSIM and VSI, and is comparable to the recently developed FR PSIM metric.
Table 5
Cross evaluation with singly distorted images. The models are trained on LIVE and then tested on the following three databases. Only those types of distortions that also appear in the LIVE database are included.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CSIQ sub</th>
<th>CSIQ full</th>
<th>TID2013 sub</th>
<th>TID2013 full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>SROCC</td>
<td>LCC</td>
<td>SROCC</td>
<td>LCC</td>
</tr>
<tr>
<td>FR-IQA [24]</td>
<td>0.882</td>
<td>0.898</td>
<td>0.842</td>
<td>0.863</td>
</tr>
<tr>
<td>CORMA [28]</td>
<td>0.889</td>
<td>0.903</td>
<td>0.879</td>
<td>0.888</td>
</tr>
<tr>
<td>CNN [35]</td>
<td>0.927</td>
<td>0.934</td>
<td>0.892</td>
<td>0.890</td>
</tr>
<tr>
<td>NFRM [25]</td>
<td>0.897</td>
<td>0.914</td>
<td>0.869</td>
<td>0.875</td>
</tr>
<tr>
<td>Wi-FiQA [39]</td>
<td>0.866</td>
<td>-</td>
<td>0.872</td>
<td>-</td>
</tr>
<tr>
<td>APNet</td>
<td>0.866</td>
<td>0.896</td>
<td>0.926</td>
<td>0.936</td>
</tr>
</tbody>
</table>

Table 6
Performance evaluation when the models are trained on authentically distorted images and tested on synthetically distorted ones. Four types of distortions (namely JP2K, JPEG, WN and GB) are included in CSIQ sub and TID2013 sub for a comparison with the results in Table 5 (trained on LIVE).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CSIQ sub</th>
<th>CSIQ full</th>
<th>TID2013 sub</th>
<th>TID2013 full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>SROCC</td>
<td>LCC</td>
<td>SROCC</td>
<td>LCC</td>
</tr>
<tr>
<td>FR-IQA [26]</td>
<td>0.701</td>
<td>0.732</td>
<td>0.584</td>
<td>0.632</td>
</tr>
<tr>
<td>APNet</td>
<td>0.672</td>
<td>0.701</td>
<td>0.609</td>
<td>0.649</td>
</tr>
</tbody>
</table>

4.4. Cross database evaluation

We conduct cross database experiments in this Section to evaluate the generalization ability of the learning-based IQA models. To this end, we first train the IQA models on the full LIVE database and test the performance on other three singly distorted databases, i.e., CSIQ, TID2013 and MICT/LCD. Note that only those types of distortions appearing in the LIVE database are included for evaluation (JP2K, JPEG, WN and GB in CSIQ and TID2013, JP2K and JPEG in MICT/LCD). Five representative IQA models are tested for comparison. The results of these methods are reported by original authors or by ourselves with released models or source codes. Table 5 shows the results, where the best one is highlighted in boldface. One can observe that our APNet performs quite well in this setting. While not as good as other state-of-the-art methods on CSIQ, it achieves the best performance on the TID2013 and MICT/LCD databases.

We next investigate how well state-of-the-art IQA metrics generalize when they are trained on authentically distorted images and tested on synthetically distorted ones. This is a challenging issue and rarely discussed in previous IQA works. In this experiment, the model is trained on LIVEC and tested on subsets of CSIQ and TID2013 (containing JP2K, JPEG, WN and GB), as well as the full databases. Table 6 shows the results, where a recent measure FRIQUEE developed for IQA on real-world images is used for comparison. We have two main observations from Table 6. One is that the performance of our APNet is generally comparable to that of the state-of-the-art FRIQUEE metric. APNet performs better on CSIQ full and TID2013 sub, but is less effective on CSIQ sub and TID2013 full. The other one is that actually the performance of both our APNet and FRIQUEE suffer from this issue. This may be probably due to the differences between modeling authentic and synthetic distortions (a similar observation was given in [26]). Developing an IQA model suitable for both real-world and synthetic images is challenging but worth further exploring.

4.5. Applying learned pooling to other IQA metrics

After training our APNet, an effective and interpretable pooling model can be obtained. An interesting question is whether the learned attention-based pooling would still be effective when applying it to other IQA algorithms. Two experiments are conducted in the following to address this issue. In specific, we validate the performance gains when using our learned pooling model to combine local quality estimations generated by two representative IQA metrics, i.e., the FR-IQA metric SSIM [3] and the vector regression based BIQA model [41].

The performance of SSIM incorporated with different pooling methods on three IQA benchmarks are compared in Table 7. In this case, the local quality maps are generated by SSIM. Our APNet is trained on the LIVE dataset, and only the attention branch is used here for global pooling, i.e., generating attention weight maps for local score aggregation. Four other pooling methods used in previous IQA works [5,7,41,52,55] are included for comparison. They are the saliency method SDSP [68] based pooling, object oriented pooling (OOP), phase congruency based pooling (PC), and analysis of distortion distribution based pooling (ADD). For simplicity, we adopt a fast implementation of object oriented pooling. The pooling weights are the counts of pixels being covered by the object proposals generated by a generic object detection method MCG [69]. One can see from Table 7 that all the five pooling methods can help to improve the performance of SSIM. We also see that SSIM-ADD achieves the best results. However, the ADD-based pooling requires information from references for calculating the pooling weights, and thus it can hardly be used in the BIQA setting. In contrast, the other four pooling methods are more practical as they can generate weights with only accessing to distorted images. Among them, our model achieves the highest SROCC and LCC accuracies on CSIQ and TID2008, as well as the most consistent performance gains across all the three databases.

We further compare the performance of applying different pooling methods to the vector regression based model presented in [41]. The model is trained on LIVE as in [41] and then used to generate local quality maps for test images in CSIQ, TID2013 and MICT/LCD. Our APNet is also trained on LIVE, and the learned attention-based pooling is employed to convert the local quality maps into image-level quality estimations. The results are listed in Table 8, where the average pooling (i.e., directly averaging the local quality estimations) is used as the baseline (abbreviated to VR). SDSP-, OOP- and PC-based pooling are also included for comparison as in Table 7 (Note that ADD-based pooling cannot be used here because there are no available references in this case). We can see that VR-APNet works consistently better than the baseline and delivers higher performance than other competitors on CSIQ and MICT/LCD. This observation once again
verifies the effectiveness of our learned attention-based pooling model.

4.6. Replacing learned pooling with other pooling methods

In this part, we investigate the performance effects of replacing the learned pooling in APNet with other pooling methods. The experiments are conducted on LIVEC, LIVE and CSIQ. For each database, we first select 80% distorted images for training APNet and the remaining 20% for testing. The detailed rules of the training and testing data partition, as well as the training procedures and configurations, are the same as those in Sections 4.2 and 4.3. After training, the learned APNet is used to generate local quality maps for test images. Then an attention/saliency based pooling method is employed to convert the generated quality maps into their corresponding overall (image-level) quality estimations. Such a training and testing iteration is performed 20 times and the median results are reported.

Table 8 shows the results, where three pooling strategies are compared, i.e., the SDSP-based pooling (APNet-SDSP), object oriented pooling (APNet-OOP) and phase congruency based pooling (APNet-PC). The performance of our APNet is used as the baseline. One can observe that the baseline APNet achieves the best results on LIVEC. Moreover, it is comparable to APNet-PC and slightly better than the other two competitors on LIVE and CSIQ. These results indicate that even without accessing to prior knowledge of HVS and attention/saliency supervision, an effective attention model for pooling in IQA can still be learned. The attention-based pooling in our APNet, though learned merely from subjective quality scores, achieves convincing results in comparison with the competitors that are elaborately designed.

4.7. Effects of correlation constraint

We introduce a correlation constraint in our method to emphasize the close relationship between quality perception and visual attention. In practice, it is implemented as a regularization term in the objective function. We investigate the effect of the introduced constraint in this part, as well as that of the hyper-parameter $\lambda_1/\lambda_2$ in (8) for weighting the MSE loss and the regularization.

To this end, two networks with different training configurations, denoted as APNet$_{1000}$ and APNet$_{noc}$, are used for comparison. For simplicity, we denote the APNet used in the above Sections as APNet$_{10}$, i.e., the hyper-parameter is set as 10 in this case. APNet$_{1000}$ and APNet$_{noc}$ are the same as APNet$_{10}$ except the setting of the constraint during training. APNet$_{1000}$ trains the network with $\lambda_1/\lambda_2$ set to 1000, and APNet$_{noc}$ is trained without the constraint.

We conduct the experiments on the LIVEC database and show the results in Fig. 6. All the experimental settings are the same as those in the Section 4.2. In general, one can observe that the performance decreases if we weaken the impact of the constraint...
5. Conclusion

In this paper, we attempt to learn an attention-based pooling model merely from subjective quality scores to help improve the overall IQA performance. To this end, a network is developed by incorporating an attention module that can automatically learn to assign visual weights while generating quality estimations in a data-driven manner. We further introduce a constraint on the correlation between the estimated local quality and attention weight to regulate the network training. It is noteworthy that our attention model is actually learned in a weakly supervised manner, without accessing to attention/saliency supervision. Extensive experiments on benchmark databases have shown that our approach achieves a great performance and generalization ability in comparison with current state-of-the-art BIQA models. Moreover, another advantage of our method is the good interpretability of the learned pooling strategy. By visualizing the attention weight maps, one can see that more weights are assigned to object-like or textural regions. In the future, we plan to advance our method to FR-IQA or to other quality assessment tasks like photo aesthetic assessment.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 91646207, 61773377, and 61573352, and the Beijing Natural Science Foundation under Grant L172053.

References

and (2016)


K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio, Show, attend and tell: visual image caption generation with attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.


Jie Gu received the B.S. degree in applied mathematics from North China Electric Power University, Baoding, China, in 2014. He is currently working toward the Ph.D. degree at the National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China. His research interests include image processing, computer vision, and machine learning.

Gaofeng Meng received the B.S. degree in applied mathematics from Northwestern Polytechnical University, Xi’an, China, in 2002, and the M.S. degree in applied mathematics from Tianjin University, Tianjin, China, in 2005, and the Ph.D. degree in control science and engineering from Xi’an Jiaotong University, Xi’an, China, in 2009. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include document image processing and computer vision. He is an Associate Editor for Neurocomputing.

Shiming Xiang received the B.S. degree in mathematics from Chongqing Normal University, Chongqing, China, in 1996, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2004. From 1996 to 2001, he was a Lecturer with the Huazhong University of Science and Technology, Wuhan, China. He was a Postdoctorate Candidate with the Department of Automation, Tsinghua University, Beijing, China, until 2006. He is currently a Professor with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include pattern recognition, machine learning, and computer vision.

Chunhong Pan received the B.S. degree in automatic control from Tsinghua University, Beijing, China, in 1987, the M.S. degree from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, China, in 1990, and the Ph.D. degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2000. He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, image processing, computer graphics, and remote sensing.