Multiple Image Distortion DNN Modeling Individual Subject Quality Assessment

Multiple Image Distortion DNN Modeling Individual Subject Quality Assessment

LOHIC FOTIO TIOTSOP, Control and Computer Engineering, Politecnico di Torino, Torino, Italy
ANTONIO SERVETTI, Control and Computer Engineering, Politecnico di Torino, Torino, Italy
PETER POCTA, Multimedia and Information-Communication Technology, Zilinska univerzita v Ziline, Zilina, Slovakia
GLENN VAN WALLENDAEL, Ghent University - imec, Ghent, Belgium
MARCUS BARKOWSKY, Deggendorf Institute of Technology, Deggendorf, Germany
ENRICO MASALA, Control and Computer Engineering, Politecnico di Torino, Torino, Italy

A recent research direction is focused on training Deep Neural Networks (DNNs) to replicate individual subject assessments of media quality. These DNNs are referred to as Artificial Intelligence-based Observers (AIOs). An AIO is designed to simulate, in real-time, the quality ratings of a specific individual, enabling an automatic quality assessment that accounts for subjects characteristics and preferences. Training AIOs is a promising but challenging research area due to the greater noise in individual raw opinion scores compared to the Mean Opinion Score. Effective learning from noisy labels necessitates the training of complex models on large-scale datasets. Unfortunately, this is challenging for AIOs as the media quality assessment community lacks extensive datasets that include individual opinion scores. To address the complexity of the task, we first created a dataset comprising two million samples, with synthetic labels derived from human annotation. We then trained a customized network for image quality assessment, named Multi-Distortion ResNet50 (MDResNet50), on this dataset. The weights of the MDResNet50 were subsequently utilized to initialize the learning process of each AIO, thereby avoiding the need to train a complex model from scratch on a small-scale dataset with raw individual opinion scores. Computational experiments show that our approach significantly advances the state-of-the-art in the AIO research. In particular: (i) we demonstrate through a simulation the ability of AIOs to mimic two well-known behavioral characteristics of a subject, i.e., bias and inconsistency, when scoring the media quality; (ii) we train and release DNN-based AIOs that, compared to the state-of-the-art, exhibit a higher performance with a statistical significance in assessing multiple image distortions; (iii) we train AIOs that more accurately mimic the sensitivity of real subjects to noise and color saturation and also better predict the opinion score distribution compared to the state-of-the-art AIOs.

CCS Concepts: • **Information systems → Multimedia information systems**; • **Computing methodologies → Machine learning;**

This work has been supported in part by PIC4SeR (http://pic4ser.polito.it).
Authors' Contact Information: Lohic Fotio Tiotsop, Control and Computer Engineering, Politecnico di Torino, Torino, Torino, Italy; e-mail: lohic.fotiotiotsop@polito.it; Antonio Servetti, Control and Computer Engineering, Politecnico di Torino, Torino, Torino, Italy; e-mail: antonio.servetti@polito.it; Peter Pocta, Multimedia and Information-Communication Technology, Zilinska univerzita v Ziline, Zilina, Slovakia; e-mail: peter.pocta@feit.uniza.sk; Glenn Van Wallendael, Ghent University - imec, Ghent, Belgium; e-mail: glenn.vanwallendael@ugent.be; Marcus Barkowsky, Deggendorf Institute of Technology, Deggendorf, Bayern, Germany; e-mail: Marcus.Barkowsky@th-deg.de; Enrico Masala, Control and Computer Engineering, Politecnico di Torino, Torino, Italy; e-mail: enrico.masala@polito.it.

ACM 1551-6865/2024/06-ART255
https://doi.org/10.1145/3664198

Additional Key Words and Phrases: Individual image quality assessment, artificial intelligence-based observers, synthetic labels, individual quality perception, transfer learning

ACM Reference Format:
Lohic Fotio Tiotsop, Antonio Servetti, Peter Pocta, Glenn Van Wallendael, Marcus Barkowsky, and Enrico Masala. 2024. Multiple Image Distortion DNN Modeling Individual Subject Quality Assessment. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 8, Article 255 (June 2024), 27 pages. https://doi.org/10.1145/3664198

1 INTRODUCTION

Subjective experiments provide individual opinion scores on the quality of media content from a group of subjects with different characteristics, experiences and expectations. In general, the individual opinion scores are averaged to obtain the so-called **Mean Opinion Scores (MOS)** of that content. Traditionally, the research on media quality assessment has been mainly focused on designing algorithms that can predict this MOS [2, 4, 22, 32, 32, 54]. Unfortunately, the MOS does not preserve all the information contained in the individual opinion scores obtained from a subjective experiment. This precludes the possibility to thoroughly assess the end users' **Quality of Experience (QoE)** if one uses algorithms that predict only the MOS. Predicting individual opinion scores rather than the MOS has therefore naturally become an interesting research topic [21, 48, 51]. In recent articles [48, 53], **Neural Networks (NNs)** have been trained (one for each subject) to mimic individual quality perception. These NNs are called **Artificial Intelligence-based Observers (AIOs)**.

AIOs enable a far more complete objective assessment of the quality of media content perceived by the end user. In fact, if one can accurately predict numerous individual opinion scores, then they can be aggregated (for bias removal) as desired to obtain a scalar estimate of the perceived quality to be optimized under the constraints imposed by the available network and storage resources. For instance, the average can be taken, yielding the MOS, or the content might be processed so that the minimum opinion score is above a certain threshold. However, the MOS or the minimum opinion score do not tell the whole story, since, for instance, it is not possible to determine the percentage of subjects dissatisfied with the quality being delivered with the available resources. In such cases, individual opinion scores might be aggregated to estimate the distribution of opinion scores and determine such a percentage. This distribution, in turn, might not answer all questions of interest. Other typical questions: Which content providers might be interested in are: Who is not satisfied with the quality of my content? How can I process my content in a way that a cluster of customers with specific characteristics will appreciate its quality? Only individual opinion scores can provide answers to these questions. Obviously, we cannot train an AIO for all possible subjects but similar to how we conduct a subjective test, we may recruit subjects with different characteristics and train their AIOs. Then, in practice, each AIO will allow us to predict the behavior of customers with similar characteristics to the subject that the AIO is mimicking.

Although the prediction of individual opinion scores rather than the MOS allows a more complete assessment of the end users' QoE, obtaining effective models of individual subjects is very challenging. In fact, the raw opinion scores directly collected from a subject are known to be noisier than the MOS [18, 25, 52]. Therefore, the development of a model that can predict these opinion scores has to deal with noisy ground-truth data. Unfortunately, effectively learning from noisy data is still an open research question in the machine learning community [41]. The first attempt to predict individual opinion scores on the quality of media content was presented in Reference [21]. The authors predicted the opinion scores of a subject through a random forest-based model. The authors of References [48, 51], instead, used a shallow NN to predict the individual

opinion scores. Random forests and shallow NNs are regression approaches that require hand-crafted features. The authors of Reference [53] highlighted a number of issues related to the use of hand-crafted features when modeling individual quality perception and suggested the use of a deep **Convolutional NNs (CNNs)** that can opportunely compute, from the input signal, the right features for each subject. Moreover, the mechanism guiding the choice of an individual in terms of the quality perception is rather complex and influenced by many factors [35]. It is therefore more reasonable to mimic such a process with complex models such as deep CNNs with several hidden layers rather than simple regression models.

This work aims at advancing the state-of-the-art toward mimicking the quality perception of an individual with a deep CNN through a three-fold contribution.

(1) Simulation shows that a deep CNN-based AIO can model and mimic two fundamental behavioral characteristics, i.e., the subject bias and inconsistency of the mimicked human subject when scoring media quality.
(2) 19 Deep CNN-based AIOs that, compared to the state-of-the-art, exhibit higher performance with statistical significance in assessing multiple image distortions, are trained and released, together with the **Multi-Distortion ResNet50 (MDResNet50)**, that is a deep CNN that we trained in a way that makes it a suitable starting point for transfer learning in media quality assessment.
(3) With respect to state-of-the-art DNN-based AIOs, the proposed AIOs can more effectively mimic the sensitivity of real subjects to the noise and reduction of the color saturation, and can predict the opinion score distribution with greater accuracy.

Our approach to train the AIOs involved two main learning steps. Specifically, we began by utilizing a small-scale, subjectively annotated, dataset to create a large-scale dataset with synthetic labels. As per the ITU-T guidelines (see Reference [15] for more details), the created dataset encompasses a wide diversity of content. This was achieved by sampling 100,000 images from the ImageNet dataset, which is well-known for the variety of its content. Furthermore, the range of each distortion (noise, blur, JPEG, and JPEG2000 compression) applied to the selected 100,000 images was carefully chosen to cover the entire quality spectrum. This was accomplished by mapping the distortion parameters to the MOS values gathered in a previous subjective test. Subsequently, we employed this synthetically annotated dataset to train the MDResNet50 in the initial learning step. Finally, we conducted a second and final learning step, which utilized a small-scale, subjectively annotated, dataset. This step commenced with the weights of the MDResNet50 as a starting point, ultimately leading to the development of 19 AIOs.

Computational experiments conducted on several datasets highlighted the effectiveness of our approach. In particular, it was shown that the trained MDResNet50 can represent a suitable starting point for the training of new DNN-based models for **Image Quality Assessment (IQA)**, as it is able to extract features that generalize better to different contexts than those of several state-of-the-art DNN-based models developed for blind IQA. Each trained AIO can assess multiple image distortions with an accuracy that outperforms that of the state-of-the-art AIOs with a statistical significance. Finally, when comparing the previously published deep CNN-based AIOs to the ones proposed in this article, it was observed that the latter can better mimic the sensitivity of real subjects to the noise and reduction of the color saturation in an image. For research reproducibility and benchmarking, the 20 trained deep CNNs, i.e., the 19 AIOs and the MDResNet50, are made freely available and can be downloaded at this link: https://media.polito.it/MD-AIOs.

This article is organized as follows. Section 2 reviews the state-of-the-art, followed by Section 3, where the ability of an AIO to mimic two well-known characteristics of the subject behavior is studied. The training process of the MDResNet50 is discussed in Section 4, followed by Section 5

that highlights the suitability of the MDResNet50 as a starting point to train AIOs on a small-scale subjectively annotated dataset. In Section 6, the derivation of the AIOs from the MDResNet50 is explained. The performance of the proposed AIOs is benchmarked in Section 7. Conclusions are drawn in Section 8.

2 RELATED WORK

The MOS obtained from the opinion scores of a group of subjects has long been considered as the gold standard or ground-truth quality of a media content. Several approaches to predict the MOS have been proposed during the last decades, e.g., References [28, 34, 57, 58]. Although the effectiveness of these approaches has been proven in many applications, several authors [11, 17, 31, 37, 42, 47, 65] have pointed out the need to go beyond the MOS to achieve an exhaustive assessment of the end users' QoE.

The authors of Reference [12] argued that the MOS alone is not enough as a measure of the end users' QoE and that the **Standard deviation of the Opinion Scores (SOS)** should also be considered; they proposed a second-order polynomial function to estimate the SOS from the MOS. The authors of Reference [50] proposed a neural network-based approach to predict the SOS from the scores of several objective metrics.

Other authors have proposed approaches to predict the whole distribution of the opinion scores for a given stimulus [6, 17, 44, 55]. These approaches constitute a first step toward answering the following question that is of interest for any media content provider: What is the percentage of end users satisfied with the quality of the stimulus under evaluation?

Although the distribution of opinion scores addresses the aforementioned question, it does not allow us to answer other important questions in this context. For instance, what are the characteristics of the users that might not be satisfied with the quality of a given stimulus? How to encode a given content in a way that an audience with specific characteristics is satisfied?

To overcome the limits of the approaches that predict the distribution of opinion scores and thus address the aforementioned two questions, researchers have recently focused on how to train a model that can predict the opinion scores of one subject [21, 48, 51, 53]. Predicting individual opinion scores is however a challenging task, since the raw individual ratings gathered in subjective tests are noisy [18, 52]. In fact, several approaches have been proposed to "clean" raw individual opinion scores and thus to recover the true subjective quality of a media content [14, 25–27, 52, 59]. Training a model to predict raw individual opinion scores is therefore a learning task performed with noisy labels. This type of learning tasks is data demanding [5] and unfortunately subjective experiments to gather individual opinion scores are time consuming and resource demanding [49].

To overcome the lack of training samples for the training of AIOs, the authors in References [48, 51] proposed an approach to combine different subjectively annotated datasets. In Reference [53], the authors proposed to first train a deep CNN to classify JPEG compressed images into five different levels of compression on a large-scale dataset, then to fine-tune its weights on a small-scale subjectively annotated dataset. The idea of pretraining a network with synthetic labels before fine-tuning it on a small scale dataset with human annotation has also been explored by several authors for the MOS prediction, e.g., References [29, 63]. Unlike the synthetic labels used in these papers, our labeling procedure is derived from human annotations of the considered distortions. This makes our pretrained network with synthetic labels, as shown later in Section 5, a model that is not only able to extract features that generalize well to different contexts, but can also be readily used to predict the perceived quality, outperforming several state-of-the-art DNN-based models fine-tuned on small-scale subjectively annotated datasets.

This work is different from the previous research on modeling individual opinion scores, since it is shown for the first time that an AIO can mimic two well-known characteristics of

the individual subject scoring behavior, i.e., the bias and inconsistency. Furthermore, this article proposes an approach to train more robust AIOs. In particular, the performance of the proposed DNN-based AIOs outperform with a statistical significance that of the state-of-the-art AIOs in assessing multiple image distortions.

3 ON THE SUITABILITY OF DNNS AS A TOOL TO MIMIC INDIVIDUAL SCORING BEHAVIOR

The neural process underlying the choices of an individual subject when assessing the quality of media content is quite complex. Researchers have tried to explain it by proposing *simplified* models of human behavior when assessing media quality. We will explain our approach to train a network which tries to mimic the entire complexity of the process that guides the choices of an individual subject in the next sections. First, in this section, through simulation, we show that the response of the softmax layer of a deep CNN-based AIO can allow reproducing two behavioral characteristics considered by a well-known and widely used model that explains subject behavior when assessing media quality.

The results presented in this section show that the uncertainty of predictions of a deep CNN-based AIO is correlated to the inconsistency of the mimicked subject. Thus, AIOs mimic the confidence of the mimicked human subjects when scoring the quality although this is not explicitly set during their training process. This result is therefore a preliminary step toward showing that a deep CNN-based AIO is not just a mathematical function that simply maps an input media content to the opinion scores of the related subject but it actually mimics some of the subject's behavioral characteristics when assessing the quality.

It is worth noting here that we rely on simulated subjects in this section, so that we know *a priori* the exact behavioral characteristics of the simulated subject. We will first introduce the model explaining the scoring behavior of a subject, which forms the basis of our simulation, then we will train a deep CNN-based AIO to mimic a simulated subject with predefined characteristics. Finally, we will present results showing the ability of the AIO to preserve the behavioral characteristics of the subject it is mimicking.

3.1 Subject's Bias and Inconsistency

Bias is a systematic tendency of a subject to provide low (negative bias) or high (positive bias) opinion scores with respect to the average perceived quality. Inconsistency instead is a measure of the inability of a subject to repeatedly provide accurate opinion scores while assessing the media quality. In fact, subjects are known for not being able to reproduce their first opinion score when rating the same media content several times.

The authors of Reference [27] argue that these characteristics, i.e., bias and inconsistency, can reasonably capture the behavior of an individual subject while scoring the quality of media content. In fact, denoted by r_{ij} the opinion score of subject j, asked to score the quality of content i, Reference [27] proposed the following model of the scoring behavior of a subject:

$$r_{ij} = q_i + b_j + N(0, \sigma_j), \tag{1}$$

where q_i is the ground-truth quality of content i, b_j and σ_j are, respectively, the bias and inconsistency of the subject j.

To put it simply, this model says that when it comes to assessing the quality of any content, the scoring behavior of each subject essentially depends on two elements, i.e., their bias and level of inconsistency. Despite the rather simple character of this model, it has proved to be very effective in the context of various applications in media quality assessment. For instance, in the ITU-T Recommendation P.913 [16] this model has been recommended as an effective one to recover the

subjective quality from noisy raw ratings. It is also worth noting that the model in Equation (1) represents one the of most outstanding tools for outlier analysis in media quality assessment, since it can point out participants with abnormal scores (subjects with particularly biased or inconsistent opinion scores).

Therefore, the quality assessment community agrees on the fact that bias and inconsistency are key aspects of individual behavior with regard to quality assessment. The question we want to address in this section is to understand whether a deep CNN can model and mimic these two characteristics by learning from the opinion scores of a subject.

3.2 Showcasing the Ability of a DNN to Mimic Subject Bias and Inconsistency

Our idea is that of using the model represented by Equation (1) to simulate the opinion scores of a subject with a specified bias and inconsistency, then train a deep CNN, i.e., an AIO, to predict these simulated opinion scores and finally check whether the trained AIO exhibits similar bias and inconsistency as the simulated subject.

We aim at simulating the opinion scores of subjects on a five point **Absolute Category Rating (ACR)** scale using the model in Equation (1). The five point ACR scale [15] is one of the most commonly used scales in QoE assessment, where participants rate the quality of a stimulus using discrete categories. It typically involves presenting the participant five possible opinion scores, i.e., "Excellent," "Good," "Fair," "Poor," and "Bad," and asking them to select the opinion score that best describes their perception of the stimulus quality. Each opinion score represents an absolute judgment of quality, independent of other stimuli or reference points. The five opinion scores are usually associated with integers ranging from 1 to 5 to compute the MOS. Unless otherwise specified, the five-point ACR scale will be considered the quality scale throughout this article.

We used the model defined in Equation (1) to simulate the ratings of 28 subjects for the images coming from the Release 2 of the LIVE IQA (R2-LIVE-IQA) [38] and KonIQ-10k [13] datasets. While the R2-LIVE-IQA dataset focuses on the individual distortions such as blur, noise, JPEG, and JPEG2000 compression, the KonIQ-10k dataset includes images whose quality is impaired by authentic distortions, i.e., quality impairments that very often result from a complex combination of the aforementioned distortions [13].

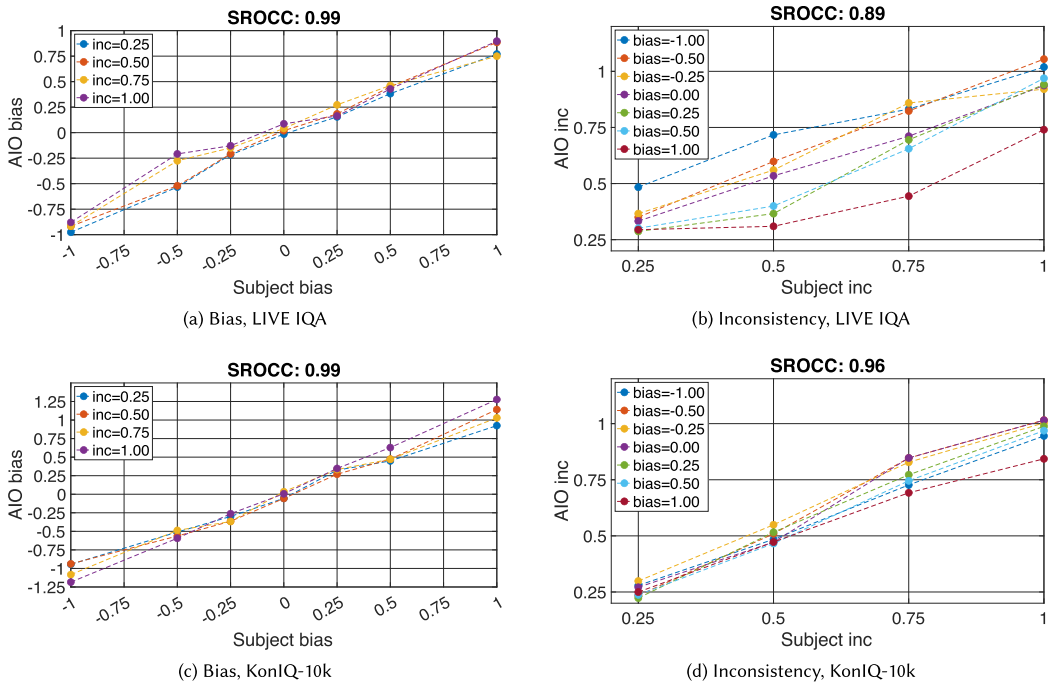
For each image i in both datasets, we denote by MOS_i the MOS for that image and consider it as the ground-truth quality. Then each simulated subject j was characterized by the couple of values $(bias_j,\ inc_j)$ that represents the exact bias and inconsistency of the simulated subject, respectively. As it can be noticed from the experiments presented in Reference [27], when rating the quality on a discrete scale ranging from 1 to 5, most of the subjects exhibit bias values mainly ranging from -0.5 to +0.5, and inconsistency values ranging from 0.25 to 1. To cover these ranges in our experiments, we employed the following set of values: {0.25, 0.50, 0.75, 1.00} for inconsistency (inc_j); and for bias ($bias_j$), in addition to {−0.50, −0.25, 0.00, 0.25, 0.50}, we also included +1 and −1 to model particularly biased subjects.

The opinion score r_{ij} of the simulated subject j for the image i was computed according to the model described in Equation (1) by the following formula:

$$r_{ij} = R_{ACR}(MOS_i + bias_j + inc_j \cdot z), \tag{2}$$

where z is a number simulated from a normal random variable with zero mean and standard deviation equal to 1, and $R_{ACR}()$ is a function that converts a real number into a label on the five point ACR scale. $R_{ACR}(x)$ returns: (i) 1 if $x < 1.5$; (ii) the closest integer to x if $1 \leq x \leq 5$; (iii) 5 if $x > 5$.

Once we obtained the opinion scores of each of the 28 simulated subjects for the images in both datasets, we used them as ground truth to train the AIO mimicking each of the 28 simulated



(a) Bias, LIVE IQA

(b) Inconsistency, LIVE IQA

(c) Bias, KonIQ-10k

(d) Inconsistency, KonIQ-10k

Fig. 1. Comparing the actual bias (left) and the actual inconsistency (right) of a subject to the overall bias and inconsistency estimated from the predicted ratings provided by the AIO mimicking that subject. It can be seen that the behavioral characteristics (bias and inconsistency) of the AIOs are strongly correlated with those of the subjects.

subjects. For the moment, we omit the details of the two-step learning approach used to train the AIO as they will be deeply discussed in Sections 4 and 6. Instead, we directly discuss the results presented in Figure 1, which compare the bias and inconsistency of the AIO to the actual bias and inconsistency of the corresponding simulated subject.

To estimate the overall bias and inconsistency of the AIO mimicking each simulated subject, the AIOs trained on the R2-LIVE-IQA dataset were used to predict the quality of 250,000 images with synthetic distortions (blur, noise, JPEG, and JPEG2000 compression) never seen during the training process. In the case of the KonIQ-10k dataset, which contains much more images than the R2-LIVE-IQA dataset (10.000 vs. 800), 80% of the dataset was used for the training and 20% for testing the ability of the trained AIOs to mimic the bias and inconsistency of the related subjects. For each image i in the test set, the softmax layer of the AIO mimicking the simulated subject j predicted five probabilities that we denote by $p_{ij}^t, \quad t = 1, 2, \ldots, 5$. p_{ij}^t represents the predicted probability that the simulated subject j score the quality of the image i as t.

To obtain the overall bias of the AIO of each simulated subject shown in Figures 1(a) and 1(c), following the definition of the bias provided in Reference [16], we computed the average of the deviations between the predicted opinion scores of the AIO and the mean of the predicted opinion scores by all the AIOs. The predicted opinion score $\hat{OS}_i^j$ of the AIO of the simulated subject j for the image i is the mode of the probability distribution determined by the probabilities $p_{ij}^t, \quad t = 1, 2, \ldots, 5$, i.e.,

$$\hat{OS}_i^j = \arg\max_t \{p_{ij}^t\}, \quad t = 1, 2, \ldots, 5. \tag{3}$$

The mean of the predicted opinion scores by the 28 AIOs for the image i (MOS_i^{AI}) was then defined as follows:

$$MOS_i^{AI} = \frac{\sum_{j=1}^{20} \hat{OS}_i^j}{28}. \tag{4}$$

The deviations between the predicted opinion score of the AIO of the simulated subject j and the mean of the predicted opinion scores for each image i was computed as follows:

$$dev_i^j = \hat{OS}_i^j - MOS_i^{AI}. \tag{5}$$

The overall bias shown in Figures 1(a) and 1(c) was therefore obtained by averaging the values of dev_i^j for each simulated subject j, over all the images in the test set. The inconsistency inc_i^j of the AIO of the simulated subject j on the quality of each image i in the test set was expressed as the variance of the probability distribution determined by the five predicted probabilities, i.e.,

$$inc_i^j = \sum_t t^2 \cdot p_{ij}^t - \left(\sum_t t \cdot p_{ij}^t \right)^2. \tag{6}$$

By averaging the values of inc_j^j over all the images in the test set, we obtained the overall inconsistency of each AIO shown in Figures 1(b) and 1(d).

The results presented in Figure 1 clearly show that the behavioral characteristic of each AIO correlates well to those of the simulated subject it is mimicking. In fact, The AIOs of simulated subjects with a high/low bias and inconsistency preserve these characteristics when it comes to the quality prediction.

The results in Figures 1(b) and 1(d) are particularly interesting. In fact, for many classification tasks, it is known that measures of uncertainty directly derived from the output of the softmax layer are not suitable indicators of the humans' confidence when performing the same task [36, 56]. However, in this particular case of single subject quality perception modeling, as highlighted by the obtained results, the variance of the response of the softmax layer is strongly correlated to the subjects' inconsistency and thus to their confidence in repeating their first opinion score when rating several times the quality of the same media content. This represents a first step toward explaining why DNNs are expected to be a suitable tool for mimicking the quality assessment of a single subject.

The results in Figures 1(a) and 1(c) showcase the ability of DNN-based AIOs to mimic not only average participants (not biased subjects) but also participants with more skewed opinion scores (particularly positively or negatively biased subjects). In other words, DNNs-based AIOs can effectively mimic the fact that real observers have different expectations in terms of quality.

Although bias and inconsistency are fundamental aspects, they do not capture the whole complexity of a subject's scoring behavior. There are several other psychological factors that influence the opinion scores of a subject on the quality of given media content. The opinion scores simulated by the model described in Equation (1) clearly do not consider these additional factors that implicitly determine the opinion scores of a real subject. Thus, training a DNN to effectively predict the opinion scores of a real subject rather than the simulated subjects considered in this section might involve further complexity. Our goal in the next sections is to train deep CNNs-based AIOs that learn how to assess multiple image distortions from the opinion scores of real subjects. Achieving this goal is hindered by the fact that existing subjectively annotated datasets do not include, in general, reliable individual opinion scores for thousands of images as required for an effective training of a deep CNN with several hidden layers. Thus, the AIOs cannot be effectively trained directly on small-scale datasets available in the literature. To overcome this lack of training samples, we

will rely on the two-step learning approach introduced in Reference [53] and will generalize it for multiple image distortions assessment.

4 FIRST LEARNING STEP: THE MULTI-DISTORTION RESNET50 (MDRESNET50)

The primary objective of our first learning step, where we derive the MDResNet50 from the ResNet50, is to transform the ResNet50 into a novel network that eliminates semantic features in favor of learning new ones tailored for image quality assessment. The t-SNE maps shown in Figure 12 illustrate the successful achievement of this goal, as the MDResNet50 appears to classify the images based on the specific distortions affecting their quality, rather than relying solely on their semantic features. The features learned by the MDResNet50 will then be leveraged in Section 6 to derive the AIOs during the second learning step.

4.1 Large-scale Dataset with Synthetic Labels for the Training of the MDResNet50

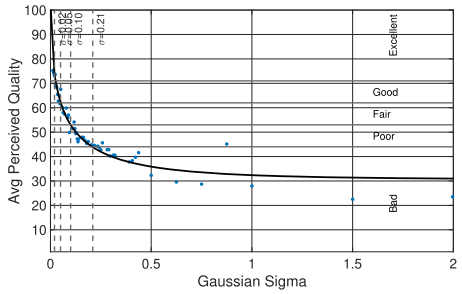
The goal of our first learning step is to train a deep CNN that can extract effective and general features for modeling the quality perception of an individual subject. It is therefore fundamental to perform such a learning step on a very large-scale dataset to avoid overfitting, and thus to end up with a deep CNN that can extract features having general validity.

Therefore, to train the MDResNet50, we created a dataset containing a very large number of images and synthetically annotated the quality of each image in the created dataset. Our rules to annotate the quality of each image are derived from the relationship between each considered distortion and the average quality perceived by subjects during the LIVE IQA experiment [38].

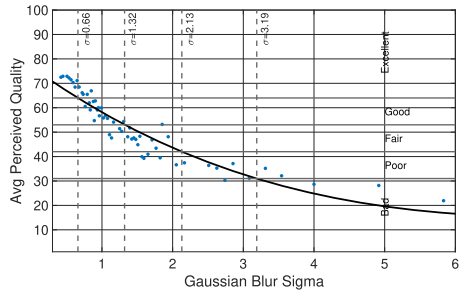
We considered four different types of distortion, namely, Gaussian noise, Gaussian blur, JPEG and JPEG2000 compression. In the R2-LIVE-IQA dataset, the level of each of these distortions is controlled by a single parameter. The noise and the blur are controlled by the standard deviations σ_{noise} and σ_{filter} of the added noise and the blur's filter, respectively. The JPEG compression is controlled by the quality parameter q. Finally, the level of JPEG2000 compression is controlled by the bit rate br of the version 2.2 of the kakadu codec [45].

For each type of distortion, we estimated a function linking the related parameter to the average subjectively perceived quality (average of the MOS values). The estimation was done by performing a least square fitting of the values of the distortion parameter to the average subjectively perceived quality. The curve showing the link between the average perceived quality and the distortion parameter for each type of distortion is shown in Figure 2.

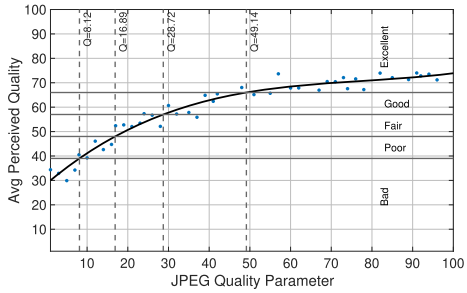
During the LIVE IQA experiment, the subjects provided opinion scores using a continuous scale ranging from 0 to 100. In an attempt to map this scale into the five point ACR scale, a natural choice would be to divide the scale ranging from 0 to 100 into five equal-sized intervals. Unfortunately, this approach disregards some contextual influence factors of the subjective experiments that can be taken into account if one proceeds differently. For instance, if subjects exhibit bias, then it is likely that not all of the quality scale will be utilized. Additionally, the quality of the stimuli selected for the test might not fully cover the quality scale based on the subjects' ratings. For these reasons, instead of dividing the entire 0 to 100 scale into five intervals with equal lengths, we firstly identified the part of the scale that was actually used by the subjects during the subjective test, i.e., the range of variation of the y-coordinates of the points displayed in Figure 2. Only this range was then divided into five intervals with equal lengths and mapped to the five quality labels in Table 1. As a result of this process, the intervals corresponding to "Bad" and "Excellent" appear to be wider than those of the other labels (see Figure 2 for more details), as they also include the extremes of the interval [0, 100], which were not utilized by the subjects during the test. This approach allows us to better account for the contextual influence factors of the subjective tests, such as the potential bias of the participants, in the synthetic labeling procedure.



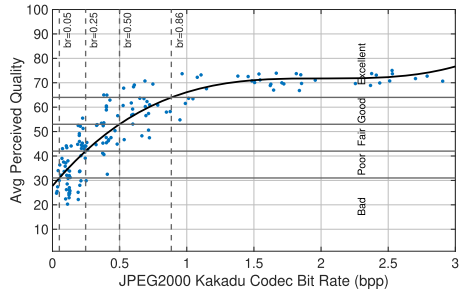
(a) $quality = 72.25 \cdot \exp(-3.70 \cdot \sqrt{\sigma}) + 30.59$



(b) $quality = -0.08 \cdot \sigma^3 + 2.28 \cdot \sigma^2 - 20.82 \cdot \sigma + 76.81$



(c) $quality = 0.0001 \cdot q^3 - 0.02 \cdot q^2 + 1.40 \cdot q + 28.62$



(d) $quality = 5.33 \cdot br^3 - 32.37 \cdot br^2 + 65.42 \cdot br + 27.74$

Fig. 2. Mapping the values of the parameters controlling each type of distortion to the average subjectively perceived quality (average of the MOS values). The curves were obtained by performing a least square fitting of the distortion parameter values to the average perceived quality.

Table 1. Rules to Synthetically Annotate Distorted Images

Distortion	Parameter	Distortion Levels and Corresponding Labels				
Noise	σ_{noise}	[0.21, 2.00]	[0.10, 0.21)	[0.05, 0.10)	[0.02, 0.05)	[0.00, 0.02)
Blur	σ_{filter}	[3.19, 6.00]	[2.13, 3.19)	[1.32, 2.13)	[0.66, 1.32)	[0.00, 0.66)
JPEG	q	[0.00, 12.00)	[12.00, 16.89)	[16.89, 28.72)	[28.72, 49.14)	[49.14, 100.00]
JP2K	br	[0.00, 0.05)	[0.05, 0.25)	[0.25, 0.50)	[0.50, 0.86)	[0.86, 3.00]
Quality labels		***Bad***	***Poor***	***Fair***	***Good***	***Excellent***

More precisely, for each type of distortion in Figure 2, we defined the range $R_{quality}$ as the one delimited by the smallest and the largest values of the average subjectively perceived quality in the dataset, i.e., the range of variation of the y-coordinates of the blue points in the graph. For instance, for the JPEG compression, the range $R_{quality}$ is the interval [30, 75]. We then split the range $R_{quality}$ for each type of distortion into five intervals with an equal size. These five intervals were then labeled, respectively, with "'Bad," "Poor," "Fair," "Good," and "Excellent," as shown in Figure 2. Finally, the interval corresponding to each of the five labels was projected on the curve linking the distortion to the average perceived quality, and then mapped to the distortion parameter to obtain the distortion range corresponding to each quality label (see Figure 2 for more details). Table 1 summarizes the final annotation rules derived from our analysis. We relied on these rules to synthetically annotate a dataset containing 2 million images.

To obtain these 2 million images, we selected 100,000 high quality images from the ImageNet dataset [23]. From each of these 100,000 images, we generated 20 new images to include in our

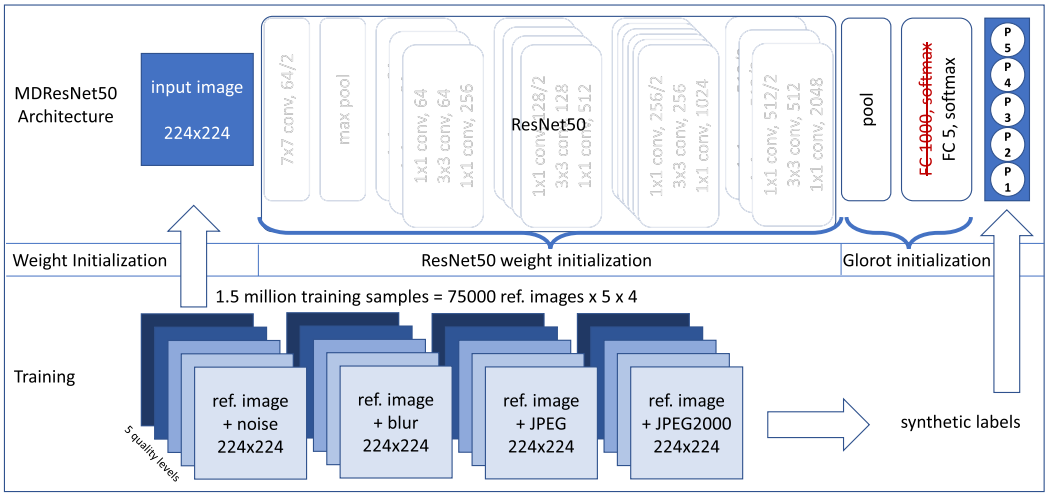


Fig. 3. The diagram details the training process of the MDResNet50. The architecture of the MDResNet50 differs from that of the ResNet50 by the use of a new fully connected and softmax layers. The weights of the convolutional layers were initialized with those of the ResNet50, while the weights of the newly added fully connected layer were initialized with random numbers generated by the Glorot's network initialization method. The MDResNet50 was trained with 1.5 million synthetically annotated images whose quality was impaired by four different distortions. Once trained, the MDResNet50 predicts a five-class discrete probability distribution after receiving an image as an input.

dataset. The 20 images were obtained by applying five different levels of each of the four distortions considered in this work to the initial image selected from the ImageNet dataset. For each distortion type, one level of distortion corresponds to a random choice of a distortion parameter value in one of the five ranges presented in Table 1. In this way, the quality of each generated image was annotated with the label corresponding to the range in which the distortion parameter used to generate the image was selected. It is worth noting here that the created synthetically annotated large-scale dataset cannot be considered as accurate as a dataset annotated by human subjects. However, as shown later in Section 5, the MDResNet50 that learned from such a large-scale dataset can extract very effective and general features for quality prediction.

4.2 Training Process of the MDResNet50

Figure 3 details the actions executed to train the MDResNet50. The architecture of the proposed MDResNet50 is strongly inspired by that of the well known ResNet50 [10]. We chose a **Residual Network (ResNet)** architecture for two primary reasons, i.e., training efficiency and empirical evidence. Given that our approach involves multiple training processes (one DNN for each subject), an architecture that ensures efficient training is essential. The authors of Reference [10] demonstrated that DNNs with a ResNet architecture train faster compared to the plain networks. Moreover, various authors in the media quality assessment community [53, 55, 61] have successfully employed the ResNet architectures, providing the empirical evidence for their suitability for quality assessment tasks. In a preliminary analysis, we tested all the three ResNet architectures commonly used in the literature, i.e., the ResNet18, ResNet50, and ResNet101. The ResNet50 and ResNet101 performed similarly, while the ResNet18 achieved a lower initial performance. We chose the ResNet50 over the ResNet101 as it has fewer parameters and thus lower computational demand.

We slightly customized the original ResNet50 architecture to create the MDResNet50. These adjustments ensure that the MDResNet50, once trained, can serve as a starting point to develop a DNN that can mimic individual subject quality assessments using the standard five-point ACR scale, without requiring additional architectural modifications. In particular, to obtain the architecture of the MDResNet50, as shown in Figure 3, we replaced the fully connected and softmax layers of the original ResNet50 with a new fully connected layer containing five neurons and a softmax layer that outputs five values. These values represent, for a given input image, the choice probability of each opinion score on the five point ACR scale, i.e., "Bad," "Poor," "Fair," "Good," and "Excellent." The MDResNet50 is therefore a deep CNN with 50 hidden layers. As it has been empirically observed for many computer vision tasks, we expect that a network with such a large number of hidden layers, when trained on a very large-scale dataset for quality assessment, learns very effective and general features characterizing the quality assessment task that can be later slightly fine-tuned to capture an individual quality perception.

We initialized the weights of the MDResNet50 as follows. For the convolutional layers, we kept the weights of the ResNet50. For the newly added fully connected layer, the initialization occurred with random numbers generated by Glorot's artificial network weights initialization method [9].

Before starting the training process of the MDResNet50, we applied two main transformations to our large-scale training set to obtain a network whose learned features have a robustness to noise and a sensitivity to image color modification that are consistent with what one would expect from human subjects.

In particular, for 33% of the images in the training set, we scaled the pixel values from the interval [0 255] to the interval [0 1] and added a Gaussian noise with a standard deviation randomly sampled in the interval (0 0.001). This noise basically corresponds to the quantity of noise that would be introduced by the quantization of the original signal using 8 bits. Thus, it is expected to be not perceptible by a human. By performing this transformation and keeping the synthetic label of the original image, we aimed at obtaining a network that can extract features that account for this aspect of the human visual system.

As shown later in the results section (see Figure 11(a) for more details), the state-of-the-art AIOs trained to recognize and score degradation caused by compression and blur predict a decrease in the quality of an image that is neither compressed nor blurry, solely because the image is presented with desaturated colors. To address this issue and ensure that the MDResNet0 is trained in a way to effectively extracts features characterizing the four considered distortions independent of whether the image is colored or grayscale, we removed the hue and saturation information from 33% of the images in the training set, converting them to grayscale images. We assigned these grayscale images the same synthetic labels as to the initial colored images. In fact, the proposed synthetic labeling procedure to create the training set of the MDResNet50 simulates subjective tests that are conducted using a single-stimulus ACR method. In our specific case, subjects are tasked with scoring their perception of mild to severe quality degradation caused by blur, noise, and compression artifacts. Therefore, when rating grayscale images, subjects do not have a colored reference image for comparison (given the single-stimulus nature of the test). Additionally, they are not trained to recognize grayscale conversions as potential degradations. Instead, their training focuses on identifying quality degradation caused specifically by blur, noise, and compression. Consequently, grayscale images appear to be undistorted to them unless any of the four degradations, which they are trained for, is applied. Therefore, we foresaw that presenting a given image in grayscale or in color scale would not significantly alter a subject's opinion score, as what really matters in this case is the perception of the four degradations under assessment.

For the training, we split the transformed dataset of 2 million synthetically annotated images into two different sets. We used 1.5 million images for the training and 0.5 million for the

validation and testing. The training was done with the stochastic gradient descent with momentum algorithm. The training settings were as follows: the learning rate was initially set to 10^{-4}, and it was multiplied by 0.1 each five epochs. The momentum parameter was kept at 0.9 for the whole training process. A five-class cross-entropy function was used as the loss function to guide the training process. The training process lasted 8 days for a total of 20 epochs. The training was performed on a computer running an Intel Core i9-10900X CPU with a clock speed of 3.7 GHz and 64 GB of RAM; and a GPU with an NVIDIA GeForce RTX 3090 GPU with 24 GB of RAM.

As it can be seen from Figure 3, once trained, the MDResNet50 receives as an input a 224*224 image patch and predicts five values, i.e., the five probabilities computed by the softmax layer. During the training process of the MDResNet50, we resized the images before feeding them to the network. Hence, to make inference with the trained network, an input image also needs to be downscaled to match the required input size.

5 EFFECTIVENESS OF THE MDRESNET50 AS A STARTING POINT FOR TRANSFER LEARNING FOR INDIVIDUAL MEDIA QUALITY ASSESSMENT

This section aims to clarify why the trained MDResNet50 is a good choice as a starting point for training AIOs. This is achieved by highlighting the following three main facts through observations and numerical experiments: (i) the MDResNet50 extracts features for blind IQA that generalize well to different contexts; (ii) the quality predictions provided by the MDResNet50 correlate to the individual opinion scores with a statistical significance; (iii) unlike the other DNN-based models, there is no need to modify the MDResNet50 architecture to use it as a baseline to train an AIO to mimic an individual perception of quality on the widely used five-point ACR scale.

5.1 General Validity of the MDResNet50 Features

Given an input image i, let us call $\mathcal{F}_i$ the set of features extracted by the convolutional layers of the MDResNet50 from the image i. The five probabilities predicted by the softmax layer of the MDResNet50 clearly depend on the features $\mathcal{F}_i$. Therefore, let us denote by $p_t(\mathcal{F}_i)\ \ t = 1, 2, \ldots, 5$, the predicted probability that the quality of the image i is scored as t. Then, the average quality of the image i estimated by the softmax layer of the MDResNet50 can be expressed as a function of $\mathcal{F}_i$ as follows:

$$MOS_{MDResNet50}(\mathcal{F}_i) = \sum_{t=1}^{5} t \cdot p_t(\mathcal{F}_i). \tag{7}$$

The effectiveness and general validity of the features $\mathcal{F}_i$ extracted by the MDResNet50 for a generic image i can be assessed by evaluating for instance the correlation between the predicted quality $MOS_{MDResNet50}$ and the MOS in numerous testing conditions. Table 2 shows the values of the **Spearman Rank-order Correlation Coefficient (SROCC)** between the MOS and the $MOS_{MDResNet50}$ for 15 different testing conditions. For comparison sake, we included in the analysis seven other recent DNN-based blind IQA approaches to predict the MOS, i.e., NIMA [44], PAQ-2-PIQ [61], DB-CNN [63], CNNIQA [20], MANIQA [60], HyperIQA [43], and the average quality predicted by the **Fuzzy Theory-based approach to predict the Opinion Score Distribution (FTOSD)** proposed in Reference [7]. The FTOSD was trained using the code available at Reference [8]. For the NIMA model, we used the implementation available in Reference [24]. For the other five measures, we employed the trained models made available in the PyTorch IQA tool [3].

Our trained DNN, i.e., the MDResNet50, excels in feature extraction, showcasing very good performance and generalization capabilities. This is shown by the results presented in Table 2, where the models were deployed to predict quality across diverse testing conditions, encompassing numerous contexts distinct from the training data. Notably, when tested against the state-of-the-art

Table 2. Spearman Rank-order Correlation Coefficient Between the MOS and the Prediction of the Quality Provided by Different DNN-based No-reference Blind IQA Approaches

	CSIQ				LIVE-R2				TID2013				VCL-FER		
Measures	**Gblur**	**Gnoise**	**JP2K**	**JPEG**	**Gblur**	**Gnoise**	**JP2K**	**JPEG**	**Gblur**	**Gnoise**	**JP2K**	**JPEG**	**Gblur**	**JP2K**	**JPEG**
CNNIQA [20]	0.70	0.03	0.67	0.23	0.76	0.38	0.52	0.21	0.82	0.05	0.85	0.53	0.75	0.61	0.02
NIMA [44]	0.61	0.44	0.75	0.64	0.58	0.30	0.76	0.74	0.81	0.56	0.89	0.78	0.60	0.66	0.37
DB-CNN [63]	0.73	0.61	0.87	0.25	0.77	0.70	**0.91**	0.47	**0.89**	0.64	**0.92**	0.66	0.74	0.79	0.52
HyperIQA [43]	0.82	**0.70**	**0.92**	**0.83**	0.87	**0.82**	**0.92**	0.86	**0.91**	**0.69**	0.90	**0.81**	0.81	**0.90**	**0.77**
MANIQA [60]	0.59	**0.70**	0.82	0.82	0.74	0.67	0.89	**0.87**	0.87	0.57	0.86	0.80	0.67	0.77	0.75
PAQ-2-PIQ [61]	**0.84**	0.24	0.79	0.63	**0.88**	0.55	0.80	0.68	0.85	0.31	0.88	0.71	**0.83**	0.68	0.61
FTOSD [7]	**0.87**	**0.77**	**0.91**	**0.92**	**0.96**	**0.96**	**0.90**	**0.93**	**0.91**	**0.85**	**0.92**	**0.95**	**0.94**	**0.85**	**0.90**
MDResNet50 [Our]	**0.91**	0.67	**0.91**	**0.94**	**0.93**	**0.96**	0.87	**0.90**	0.88	**0.82**	**0.91**	**0.92**	**0.90**	**0.83**	**0.91**

The top three metrics are highlighted in bold.

approaches, our model delivered the highest performance in five testing conditions out of 15 and was among the top three metrics in 12 testing conditions out of 15. This highlights the capability of the trained MDResNet50 to extract features that adapt and perform well in varying scenarios.

It is very interesting to observe that learning from the synthetic labels, as demonstrated by the MDResNet50, can yield competitive performance against highly accurate state-of-the-art models such as HyperIQA and FTOSD, which were trained on subjectively annotated datasets. However, it is crucial to emphasize here that our primary focus lies in training AIOs, i.e., DNNs capable of mimicking individual perception of quality, rather than predicting the MOS. Therefore, the analysis presented in Table 2 is mainly aimed at assessing the suitability of the MDResNet50 as a foundational model for training an AIO to replicate individual quality perception on the five-point ACR scale. The results unequivocally indicate that the MDResNet50 is well-equipped for this task, as it extracts features enabling state-of-the-art performance in quality prediction. Furthermore, a model like the MDResNet50, trained without subjective annotations, encounters a broader diversity of the content during the training, leading to features with lower content and context dependency. Consequently, we believe the MDResNet50 serves as an excellent starting point for transfer learning tasks aiming at the derivation of AIOs

5.2 Correlation Between MDResNet50 Output and Individual Opinion Scores

The correlation between the opinion scores simulated by the softmax layer of the MDResNet50 and the opinion scores of the 19 subjects whose AIOs will be trained in the next section was also studied. The predicted opinion score on the five point ACR scale of the MDResNet50 on the quality of an input image i can be simulated as the mode of the distribution determined by the five probabilities $p_t(\mathcal{F}_i)\ \ t = 1, 2, \ldots, 5$, i.e.,

$$OS_{MDResNet50}(\mathcal{F}_i) = \arg\max_t \{p_t(\mathcal{F}_i),\ \ t = 1, 2, \ldots, 5\}. \tag{8}$$

The SROCC between the opinion scores of each of the 19 subjects and the $OS_{MDResNet50}$ was computed. The obtained SROCC values varied from a minimum of 0.37 to a maximum of 0.69. All these correlations were tested to be different from 0 with a statistical significance, since all the p-values were strictly smaller than 0.05. Thus, with more than a 95% of confidence, the hypothesis that the features learned by the MDResNet50 do not capture some aspects of the individual quality assessment of each of these 19 subjects can be rejected. This, from our point of view, is an additional motivation to use the MDResNet50 as a baseline to derive the AIOs of these 19 real subjects.

5.3 Suitability of the MDResNet50 Architecture for the Training of AIOs

Unlike the other DNN-based IQA models tailored for a MOS prediction, the architecture of the MDResNet50 can be used as it is to train an AIO that mimics an individual subject on the five-point ACR scale. The MDResNet50 was customized to output five probability values, simulating a

subject's choices on the five-point ACR scale. It is worth noting here that modifying an architecture typically involves an introduction of new layers and, consequently, an optimization of weights of those new layers from scratch. Therefore, having an architecture that requires no modification offers significant advantages by avoiding the need to train new weights from scratch on a small-scale dataset, so reducing the risk of overfitting during the derivation of AIOs.

6 SECOND LEARNING STEP: DERIVING AIOS FOR MULTIPLE IMAGE DISTORTION ASSESSMENT FROM THE MDRESNET50

We describe in this section the process implemented to train 19 AIOs starting from the MDRes-Net50.

6.1 Small-scale Training Set

To train the AIOs, we relied on the data gathered during the first part of the **Multiply Distortion LIVE Image Quality Assessment (MD-LIVE-IQA)** subjective experiment [19] for which the raw opinion scores of 19 subjects were freely available. We aimed therefore at training the AIO for each of these 19 subjects.

The first part of the MD-LIVE-IQA dataset contains a total of 225 distorted images. These images were generated from 15 reference images. The quality of each reference image was impaired by using three different levels of the JPEG compression and blur but also considering the application of both blur and JPEG compression. The generated distorted images were shown to 19 subjects. The subjects expressed their opinion scores on a 0 to 100 continuous quality scale.

To train the AIOs so that they mimic the quality scoring process of the related subjects on the well-known and widely used five point ACR scale, we linearly mapped the raw opinion scores to the interval [1 5] and rounded them to integers.

6.2 From the MDResNet50 to the AIOs

It is worth noting here that 225 training samples are rather limited to effectively train from scratch a deep CNN with 50 hidden layers. Instead of starting from randomly generated weights to be optimized from scratch, we started the training of the AIOs from the weights of our MDResNet50 that can readily extract useful features for quality prediction. So, the weight optimization process mainly consisted in slightly modifying the features already learned by our MDResNet50 to obtain new ones that characterize each individual quality perception. In this way, the weight optimization process of the AIOs did not last that long and the risk to overfit the few available training samples was minimized, thus preserving the effectiveness of the AIOs also on content not seen during the training process.

Figure 4 summarizes all actions to derive an AIO from theMDResNet50. As it can be seen from Figure 4, no action is required to derive the architecture of an AIO from that of the MDResNet50. In other words, to train the AIO, we kept the architecture of the MDResNet50 and just performed a second learning step to opportunely update the weights of the MDResNet50 and thus to transform them into new ones that can extract features modeling the quality perception of each individual subject to be modeled.

For each AIO, as it can be seen from Figure 4, we initialized the weights with those of the MDResNet50. Then, we performed the training of the AIO of each of the 19 subjects using as ground truth their 225 opinion scores taken from the training dataset described in the previous section.

Before the training process of the AIO of each subject, a high percentage of the convolutional layers of the MDResNet50 was frozen, i.e., the weights of these convolutional layers were never modified during the training process yielding the AIO. This was done to preserve the high-level

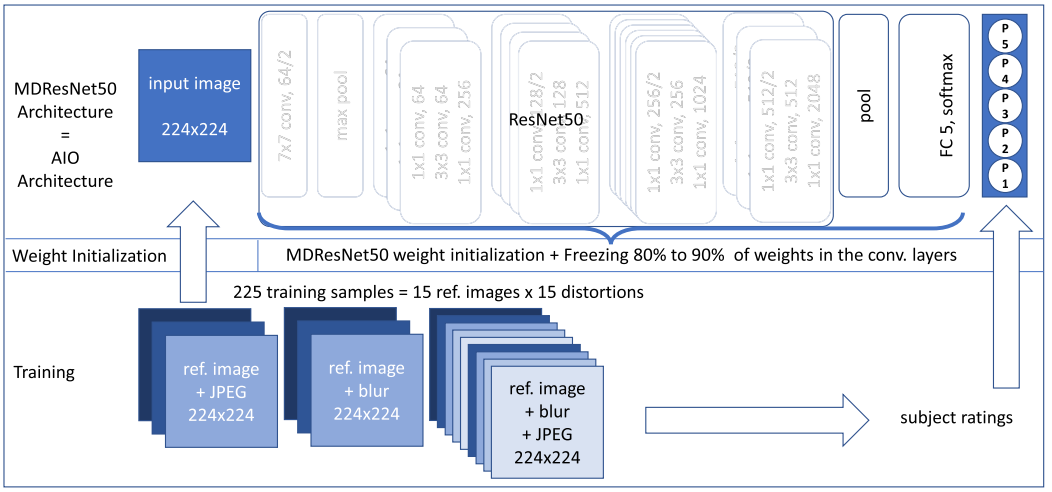


Fig. 4. The diagram details the training process of an AIO starting from the MDResNet50. The architecture of each AIO is the same as that of the MDResNet50. All the weights of each AIO were initialized with those of the MDResNet50. Each AIO was trained using the opinion scores of the related subject on the 225 images representing the training set. Once trained, each AIO predicts a five class discrete probability distribution after receiving an image as an input.

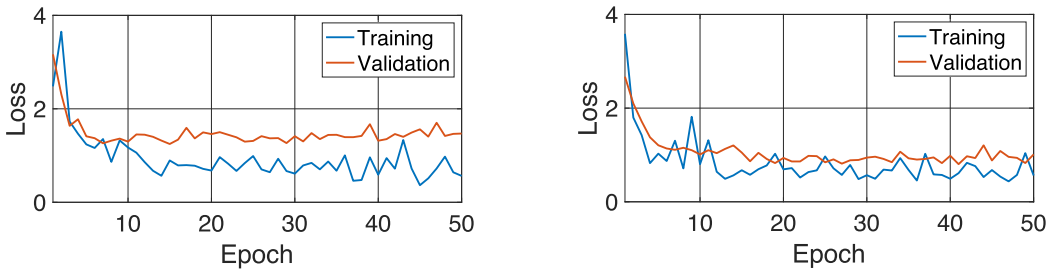
features learned by the MDResNet50 from the created large-scale dataset. The exact percentage of frozen layers varied from one subject to another.

More precisely, the optimal percentage of the convolutional layers to be frozen for each subject was identified through computational experiments. In particular, when training the AIO of each subject, we used 80% of their 225 opinion scores as the training set and 20% as the validation set. We trained the AIO several times by progressively increasing the percentage of the frozen convolutional layers of the MDResNet50. The optimal percentage for each subject was the one for which the AIO showed the highest performance on the validation set. As reported in Figure 4, the percentage of the frozen convolutional layers ranged from a minimum of 80% to a maximum of 90%.

After freezing the optimal percentage of the convolutional layers, the training of each AIO was performed with the stochastic gradient descent with momentum algorithm. The learning rate and the momentum parameter were set to 10^{-4} and 0.9, respectively, and kept fixed during the training process. A five-class cross-entropy function was used as the loss function to guide the training process. The optimal number of training epochs, i.e., the one guarantying the highest accuracy on the validation set, varied from one subject to another. The minimum number of the required training epochs was 10, while the maximum was 50. Figures 5(a) and 5(b) show, as an example, the training and validation curves for two AIOs. The training process of the AIO displayed in Figure 5(a) was interrupted after 10 epochs, since it can be seen that after that epoch the network starts overfitting the training set. However, the training process of the AIO presented in Figure 5(b) lasted for the full 50 epochs. In any case, the derivation process of an AIO from the MDResNet50 never resulted in a training process lasting more than 13 minutes on a computer running an Intel Core i9-10900X CPU with a clock speed of 3.7 GHz and 64 GB of RAM; and a GPU with an NVIDIA GeForce RTX 3090 GPU with 24 GB of RAM.

6.3 Making Inference with AIOs

Given an image i, to predict its quality with the AIO of the subject s, this image is first resized to match the required 224*224 input size. The resized image can then be provided as an input to the



(a) Early stop: (10 epochs) (b) Stopped after 50 epochs

Fig. 5. Illustration of the training progress of two AIOs. (a) The training process was stopped after 10 epochs to prevent overfitting.

AIO. Clearly, resizing an input image might affect its perceptual quality. For instance, a change in image aspect ratio during resizing may emphasize or mitigate the visibility of certain impairments. Unfortunately, similarly as in this article, several previous works in media quality assessment have disregarded that issue, since approaches to effectively and efficiently train deep CNNs for IQA that allow arbitrary size for the input image are still at their early stage. We therefore see this limit of our current work as a point to be addressed in a future contribution.

As it can be seen from Figure 4, when receiving an image i as an input, the softmax layer of the AIO of the subject s predicts five probabilities that we denote by $p_s^{it} \quad t = 1, 2, \ldots, 5$. From these probabilities the average perceived quality Q_s^i of the AIO of the subject s for the image i can be expressed as

$$Q_s^i = \sum_{t=1}^{5} t \cdot p_s^{it}, \tag{9}$$

while the predicted opinion score OS_s^t of the AIO of the subject s for the image i on the five point ACR scale is instead defined as

$$OS_s^i = \arg\max_t \{p_s^{it} \quad t = 1, 2, \ldots, 5\}. \tag{10}$$

It is worth noting here that while the large-scale dataset used for training the MDResNet50 considers each distortion individually, our second learning step to derive the AIOs is executed on the LIVE multiply distorted image quality dataset, where subjects also rated the quality of the images degraded by a combination of blur and JPEG compression. This enables the proposed AIOs to learn how to map features of individual distortions to the subjectively perceived quality of these images and thus to learn how to make inference also when distortions are jointly applied.

7 BENCHMARKING THE PERFORMANCE OF THE PROPOSED AIOS

Before presenting the computational experiments and the obtained results, we remind the interested reader of the fact that the training of AIOs is rather a recent research direction. In particular, to the best of our knowledge, the state-of-the-art on DNN-based AIO for blind IQA is mainly represented by this very recent paper [53]. So, we will benchmark the performance of the proposed AIOs with those proposed in Reference [53] but also with real observers.

7.1 Performance of AIOs in Assessing Multiple Image Distortions

The experiments involved evaluating the correlation between the predicted opinion scores generated by the AIOs and the subjectively perceived quality (i.e., the MOS) under various testing

Table 3. Spearman Rank-order Correlation Coefficient (SROCC) Between the Opinion Scores of Each One of the Mimicked Real Subjects and the MOS on the Training Set, i.e., the MD-LIVE-IQA Dataset

Subjects	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
SROCC	0.84	0.82	0.91	0.83	0.88	0.88	0.83	0.83	0.84	0.87	0.85	0.93	0.85	0.60	0.85	0.81	0.93	0.84	0.82

conditions. Subsequently, this correlation was compared with the correlation between the opinion scores of real subjects and the MOS.

Let us denote by OS_i the set of opinion scores given by the subject i to all stimuli in the dataset and by OS_{AIO_i} the opinion scores predicted by the AIO of the subject i. We compared the following correlations: $corr(OS_i, MOS)$ and $corr(OS_{AIO_i}, MOS)$. The value of $corr(OS_i, MOS)$ is reported in Table 3 for each subject. The correlation $corr(OS_i, MOS)$ serves as a benchmark or a reference to assess the performance of the AIO of subject i. In fact, we believe that the AIO of the subject i can be considered highly accurate when its quality predictions closely align with the MOS, much like the correlation achieved by the mimicked real subject, i.e., if $corr(OS_{AIO_i}, MOS)$ is close to $corr(OS_i, MOS)$.

Excluding subject #14 that clearly seems to be an outlier, the SROCC values between the ratings of the real subjects and the MOS belong to the interval $I_{benchmark} = [0.81, 0.93]$. This range can be used to analyze the performance of the proposed AIOs. Figure 6 presents a heatmap showing SROCC values between the AIO predictions (computed as defined in Equation (9)) and the MOS for the various datasets and image distortion types. Notably, in many testing conditions, the AIO predictions exhibited correlations exceeding 0.81, falling within $I_{benchmark}$. Therefore, in these cases the proposed AIOs competed well with real subjects in assessing the considered distortions. Some AIOs even achieved a remarkable 0.9 correlation with the MOS, despite being trained only on 225 samples. This demonstrates the potential of our two-step learning approach.

It should be noted here that some of the AIOs struggled with predicting the quality of the noisy and JPEG2000-compressed images. For example, the AIO of subject #12 achieved a low correlation (0.16) with the MOS for the images including noise coming from the TID2013 dataset, while the AIO of subject #8 showed an SROCC of 0.31 with the MOS for the JPEG2000-compressed images coming from the VCL-FER dataset. This difficulty is not very surprising, given the fact that the noise and the JPEG 2000, as distortion, are present only in our large-scale dataset used during the first learning step. In the second learning step, the best subjectively annotated dataset we found for our purpose only includes the distortions caused by the JPEG compression, blur and the combination of both. So, it seems that some of the AIOs probably slightly overfitted the training set of the second learning step that included only the distortions induced by the blur and JPEG compression and unfortunately therefore lost the MDResNet50 capability to characterize the artifacts caused by the noise and JPEG2000 compression.

The last row of the heatmap presented in Figure 6 shows the correlation between the **mean of the prediction of the AIOs (MOS-AI)** and the MOS. Although the MOS-AI did not deliver the highest performance for each individual testing condition, it showed a higher stability over all the testing conditions as compared to the AIOs of the individual subjects. In fact, for almost all the AIOs, the lowest correlation observed is smaller than the worse correlation between the MOS and the MOS-AI.

In Figure 7, we compared the overall performance, over all the 15 testing conditions (see Figure 6 for more details), of the proposed AIOs to that of the AIOs recently published in Reference [53]. It is worth noting here that the AIOs published in Reference [53] were trained on the same dataset as the ones proposed here, i.e., the MD-LIVE-IQA dataset. So, they mimic the same 19 subjects that we are modeling also in this article. This makes the comparison possible. The overall performance of each AIO across all the 15 testing conditions shown in Figure 6 was computed by applying

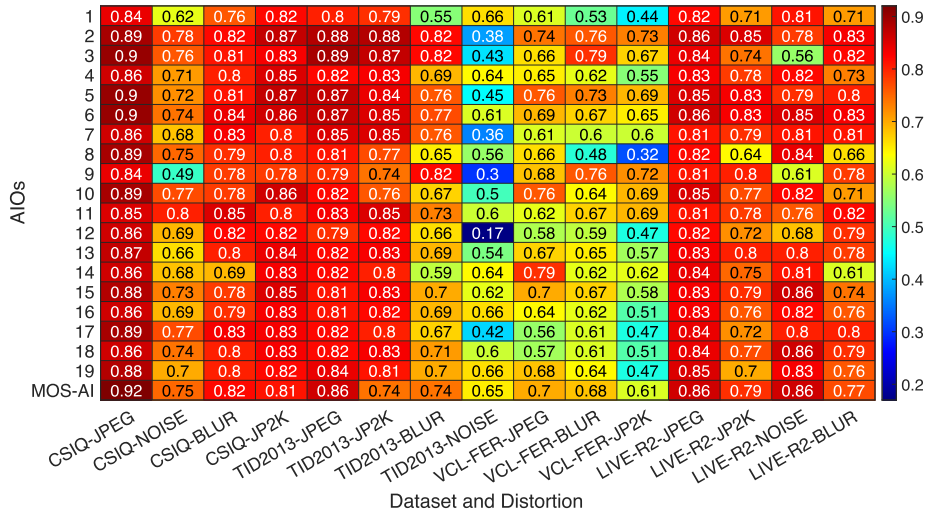


Fig. 6. SROCC between the prediction of each AIO and the MOS for the different datasets and distortions.

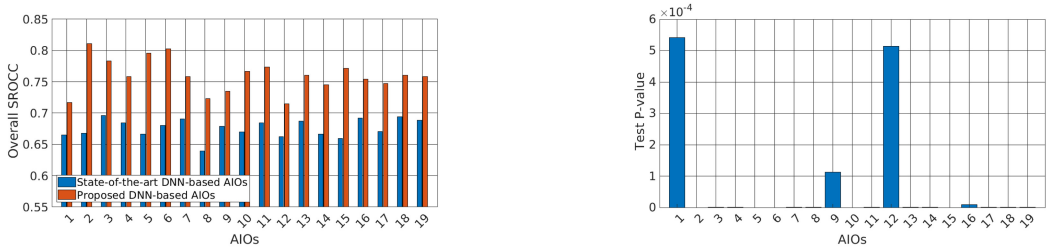


Fig. 7. (a) Comparison of the overall performance of the proposed AIOs to that of the AIOs published in Reference [53]; (b) p-values of statistical tests assessing the statistical significance of the superiority of the proposed AIOs over the AIOs coming from Reference [53].

the Fisher's Z transformation. It is important to note here that it is strongly recommended not to average correlation coefficients. Instead, it is recommended to first perform a Fisher-Z transform of the individual correlations, then calculate the average of these transformed values, and consider the inverse transform of this average as the overall correlation. This approach has been utilized by several authors in media quality assessment (e.g., References [1, 46]). Therefore, we employed the same method to derive the overall performance of each AIO.

As it can be seen from Figure 7(a), for each of the 19 subjects, the proposed AIO outperformed the state-of-the-art AIO published in Reference [53] in terms of the overall performance on the considered datasets and for all the investigated distortions. Considering all the 15 combinations of the datasets and distortions (see Figure 6 for more details), let us denote by ρ_i the overall correlation between the MOS and the predictions of the AIO, which we trained for the subject i in this work, i.e., the proposed AIO. Similarly, let γ_i represent the overall correlation between the MOS and the predictions obtained from the AIO published in Reference [53] for the subject i. To show that the superiority of the proposed AIOs occurs with statistical significance, we conducted a statistical test to examine the following null hypothesis: $\rho_i = \gamma_i$ against the alternative hypothesis: $\rho_i > \gamma_i$.

(a) std=0.05 (b) std=0.15 (c) std=0.25 (d) std=0.35 (e) std=0.45

Fig. 8. Visual effect of the Gaussian noise applied to the images in our sensitivity analysis. The level of degradation is controlled by the standard deviation (std) of the Gaussian noise. We expect that a subject tasked to rate distortion caused by noise, blur, JPEG and JPEG2000 compression perceives a decrease in the image quality when moving from panel (a) to panel (e).

As it is shown in Figure 7(b), all the p-values are smaller than $6 * 10^{-4}$. Thus, we can confidently reject the null hypothesis and favor the alternative hypothesis.

7.2 Sensitivity of the AIOs to Noise and Color Saturation

There are numerous aspects upon which a sensitivity analysis can be conducted. In this particular case, our objective is to verify whether the proposed AIOs address specific limitations of the state-of-the-art AIOs. As it will be elaborated upon later, these limitations include: (i) The state-of-the-art AIOs exhibit sensitivity to noise that deviates from what is expected from a real subject behavior. (ii) Despite being trained to evaluate the quality degradation caused by factors such as the signal compression and blur, the state-of-the-art AIOs predict a decrease in the quality of images that are neither compressed nor blurred, solely due to reduction in color saturation. The decision to examine the AIOs sensitivity to noise and color saturation was thus strongly motivated by the necessity to showcase the improvement of the proposed AIOs when it comes to the above-mentioned limitations of the state-of-the-art AIOs.

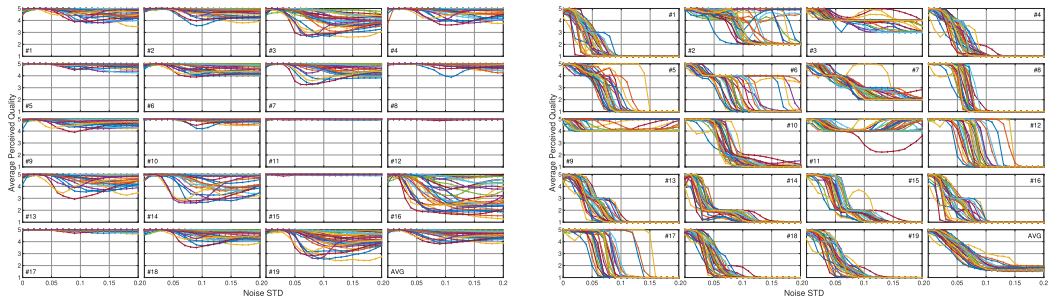
After adding the noise to an image that a subject considers as an image of pristine quality, it is expected that this subject perceives a lower quality. However, the reduction of the color saturation does not specially introduce any of the distortions studied in this work, and thus an accurate subject tasked to rate the annoyance of these distortions is expected to perceive the same quality after the reduction of color saturation. We provided modified images as an input to both the newly trained AIOs and the AIOs trained and released in Reference [53] to see, which AIOs better mimic the expected behavior of real subjects. We modified the images by either adding Gaussian noise or progressively reducing the color saturation. Figures 8 and 9 illustrate the applied modifications to one image.

For the experiment, we selected from the ImageNet dataset 30 images whose quality was judged as being pristine by all the AIOs, i.e., all the AIOs scored the quality of these images by predicting ”Excellent” as opinion scores. The selected images were never seen by the AIOs during their training process. We applied to each of these images different levels of noise or reduction of the color saturation. The obtained images were then given as an input to the AIOs and a curve representing the perceived quality of each AIO as a function of the applied modification was drawn.

Figures 10(a) and 10(b) show the average perceived quality of each AIO (computed as defined in Equation (9)) as a function of the standard deviation of the added noise. Each plot corresponds to one AIO and each curve of each plot represents the trend of the perceived quality for one image as the standard deviation of the added noise increases. As it can be seen from Figure 10(a), the AIOs trained in Reference [53] are not particularly sensitive to the noise. In fact, for several AIOs, i.e., #5, #9, #10, #11, #12, #15, and #17, the perceived quality remains constant when the standard

(a) perc=80% (b) perc=60% (c) perc=40% (d) perc=20% (e) perc=0%

Fig. 9. Visual effect of the color saturation reduction applied to the images in our sensitivity analysis. The modification is measured in terms of the percentage of the remaining color saturation with respect to the original image. We expect that a subject tasked to rate distortion caused by noise, blur, JPEG and JPEG 2000 compression does not perceive a significant difference in the image quality when moving from panel (a) to panel (e).



(a) AIOs published in [53] (b) Proposed AIOs

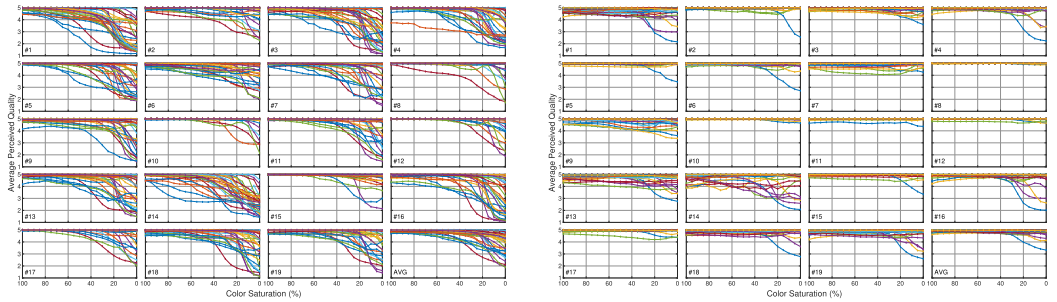
Fig. 10. The average perceived quality by the state-of-the-art AIOs published in Reference [53] (left) and the proposed AIOs (right) as a function of the standard deviation of the Gaussian noise added to the input image. Each plot (excluding the one labeled as "AVG") corresponds to one AIO and each curve in each plot shows the trend of the perceived quality by the AIO as the standard deviation of the added noise increases. The plot labeled as "AVG" shows the average of the trends exhibited by the 19 AIOs.

deviation of the noise increases. For the other AIOs, for some images, there is a slight decrease of the perceived quality when the standard deviation of the noise increases.

The behavior of the AIOs trained in this work is shown in Figure 10(b). Except for a few AIOs (3 out of 19), in general, as the standard deviation of the noise increases, the proposed AIOs clearly perceive a lower quality. The results show that although the proposed AIOs learned from the same dataset, which was used to train the AIOs published in Reference [53], they have a sensitivity to noise that is much more similar to that of real subjects. This can be explained by the fact that we froze the convolutional layers of the MDResNet50 that extract high level noise related features and allowed the AIOs, through the transfer learning process, to benefit from these features.

Figures 11(a) and 11(b) show the average perceived quality of each AIO (computed as defined in Equation (9)) as a function of the percentage of the color saturation. On the x-axis 100% corresponds to the original color saturation of the image, while 0% corresponds to a grayscale image, i.e., when the image color has been totally desaturated. It can be observed in Figure 11(a) that the perceived quality of several images by the AIOs trained in Reference [53] decreases when the color saturation is reduced by more than 60%. In fact, these AIOs seem to perceive a very low quality when a grayscale image is provided as an input.

Figure 11(b) presents instead the trend of the perceived quality provided by the proposed AIOs. In general, as expected from a real subject, the proposed AIOs do not significantly change their



(a) AIOs published in [53] (b) Proposed AIOs

Fig. 11. Average perceived quality by the state-of-the-art AIOs published in Reference [53] (a) and the proposed AIOs (b) as a function of the percentage of the color saturation of the input image. Each plot (excluding the one labeled as "AVG") corresponds to one AIO and each curve in each plot shows the trend of the perceived quality by the AIO as the image color saturation decreases. The plot labeled as "AVG" shows the average of the trends exhibited by the 19 AIOs.

perceived quality when the color saturation of the input image decreases. In fact, for each AIO, the curve associated with each image exhibits an almost constant trend. This result is particularly interesting, because although the proposed AIOs never saw images with a partial reduction of the color saturation during the training process, they are still able to mimic the fact that this kind of modification does not have a significant impact on the quality of the image when one is assessing the annoyance of impairments caused by blur, noise and compression. This is again an ability inherited from the MDResNet50, which through the transformation into luminance component of part of the training samples, has learned that the absence of color does not necessarily imply a quality degradation.

7.3 Opinion Score Distribution Prediction

In this section, we assess the accuracy of AIOs in predicting the **Opinion Score Distribution (OSD)**. The experiments were conducted on six datasets. In the experiments, the ground-truth OSD was predicted using: the proposed AIOs, the state-of-the-art AIOs published in Reference [53], NIMA [44] and FTOSD [7]. The following six datasets were considered: parts 1 and 2 of the LIVE multiply distortion experiment [19] (LIVE MD1 and LIVE MD2), sessions 1 and 2 of the first release of the LIVE image quality assessment experiment [39] (LIVE R1S1 and LIVE R1S2), the MICT [33] and the SJTU IQSD [7]. The effectiveness of each method was evaluated in terms of the average **Earth Mover's Distance (EMD)** [7] between the predicted distribution and the ground-truth one. The lower the EMD, the better it is.

Table 4 presents the results of this study. It can be noticed from this table that the trained AIOs outperform the state-of-the-art AIOs also in terms of predicting the OSD. When comparing the proposed AIOs to the approaches explicitly designed for the OSD prediction, we find that while the proposed 19 AIOs show competitive performance with NIMA, they generally lag behind FTOSD in terms of OSD prediction accuracy. Nonetheless, the results remain highly promising, particularly considering the fact that we attempt to predict the entire distribution with just 19 opinion scores. We believe that incorporating more AIOs, each modeling a subject with different characteristics, could further enhance the accuracy of AIOs in predicting the OSD. This is a point to investigate in future work. Additionally, this analysis underscores the potential of AIOs, demonstrating their competitiveness to predict the OSD while retaining information on individual preferences, a capability not available with approaches specifically tailored for the OSD prediction.

Table 4. Earth Mover's Distance (EMD) Between the Ground-truth Distribution of Opinion Scores and the Predicted One

	LIVE MD1	LIVE MD2	LIVE R1S1	LIVER1S2	MICT	SJTU IQSD
NIMA [44]	0.308	0.322	0.289	0.239	0.276	0.233
FTOSD [7]	0.189	0.173	0.206	0.213	0.297	0.109
AIOs from Reference [53]	0.120	0.294	0.369	0.372	0.335	0.300
Proposed AIOs	0.123	0.243	0.285	0.312	0.311	0.254

Table 5. Overall SROCC Between the MOS and the Prediction of the AIOs Trained With and Without Our First Learning Step, i.e., Using Directly the ResNet50, VGG16, and ViT to Derive the AIOs Instead of First Training the MDResNet50

AIOs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
ResNet50	0.17	0.03	0.08	0.24	0.06	0.19	0.21	0.14	0.18	0.10	0.19	0.33	0.16	0.00	0.21	0.11	0.01	0.17	0.22
VGG16	0.33	0.48	0.38	0.34	0.31	0.30	0.25	0.45	0.39	0.27	0.37	0.42	0.39	0.17	0.33	0.24	0.45	0.29	0.35
ViT	0.38	0.48	0.45	0.43	0.28	0.46	0.55	0.50	0.48	0.45	0.60	0.51	0.22	0.50	0.24	0.46	0.4	0.37	0.32
MDResNet50	**0.72**	**0.81**	**0.78**	**0.76**	**0.80**	**0.80**	**0.76**	**0.72**	**0.74**	**0.77**	**0.77**	**0.71**	**0.76**	**0.75**	**0.77**	**0.75**	**0.75**	**0.76**	**0.76**
MDResNet50 perf > ViT perf: Stat Test p-values	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$

7.4 Ablation Study: Importance of the MDResNet50

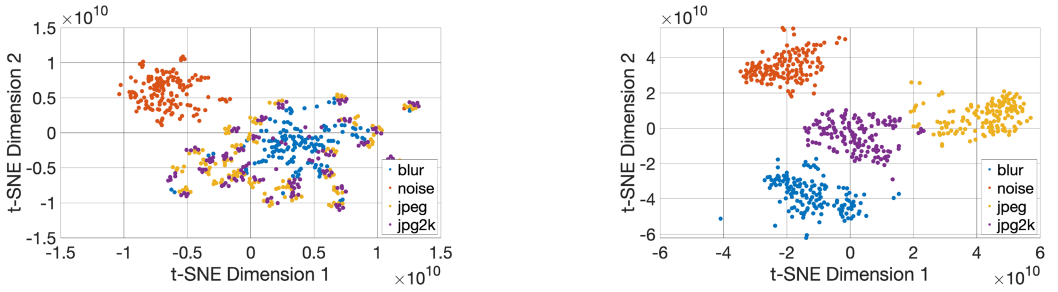
We investigated the impact of eliminating the initial learning step within our two-step learning approach on the performance of the AIOs. Specifically, we compared the results achieved when employing the ResNet50 [10], VGG16 [40], and a **Vision Transformer (ViT)** model [30] pretrained on the ImageNet dataset, as starting point for our second learning step, as opposed to our custom MDResNet50, which was trained during the first step of our training process.

The outcomes are summarized in Table 5. The calculation of the overall SROCC was performed similarly as in Section 7.1 and it encompassed all the 15 testing conditions, i.e., the datasets and distortions (see Figure 6 for more details). Notably, the results revealed the fact that the MDResNet50 serves as a superior foundation for the derivation of AIOs as compared to the ResNet50, the VGG16 and the considered ViT model. In fact, the overall SROCC values between the MOS and the quality predictions provided by the AIOs derived from the MDResNet50 outperformed those derived from a direct use of the ResNet50, the VGG16, and the ViT. We also conducted statistical tests to assess the significance of the superiority of the AIOs derived from the MDResNet50 with respect to those obtained from the ViT model. The corresponding p-values are reported in Table 5. All the p-values are smaller than 10^{-4}. Hence, the use of the MDResNet50 yields AIOs that perform better with statistical significance. It is interesting to note here that, despite the ViT did not perform better than the MDResNet50, it outperformed the original ResNet50. This suggests that performing the proposed two-step learning approach with a ViT, as underlying architecture, could yield AIOs with enhanced performance. This is coherent with the recent performance exhibited by the ViT architectures in IQA [62]. We will further investigate this aspect in the future as an extension of our current work.

Figure 12 shows a comparison of the 2D t-SNE maps of the features extracted by the ResNet50 and the MDResNet50 from input images whose qualities were degraded by different distortions. It can be clearly seen from that figure that the MDResNet50 better distinguishes among the different distortions compared to the original ResNet50, showing a distinct cluster for each type of distortion. This further highlights the suitability of the MDResNet50 as a starting point for transfer learning in image quality assessment compared to the original ResNet50 model.

8 CONCLUSIONS

In this work, we focused on the question of how to train a deep CNN that can effectively mimic the quality perception of an individual subject, i.e., an AIO. We proposed an approach to create a



(a) ResNet50 (b) MDResNet50

Fig. 12. 2D t-SNE maps of the features extracted by the ResNet50 (a) and the MDResNet50 (b). It can be noticed that the MDResNet50 better distinguishes among the different image distortions.

synthetically annotated dataset containing 2 million images whose quality was impaired by noise, blur, JPEG, and JPEG2000 compression. We then trained the MDResNet50 on the created large-scale dataset, ending up with a network that can extract effective features characterizing the four considered distortions. The AIO of each subject was then obtained by performing a transfer learning step on the MDResNet50.

The training procedure described in the previous paragraph was first applied to train the AIOs of 20 subjects whose ratings were simulated in a way that their bias and inconsistency were known *a priori*. We then showed that each trained AIO can accurately capture and mimic the bias and inconsistency of the simulated subject, which was trained to mimic. This shows that the AIOs can mimic these two fundamental aspects of the individual scoring behavior, i.e., the bias and inconsistency. This result constitutes a first step toward proving the ability of AIOs to mimic the certain well-known aspects of the individuals scoring behavior.

We then trained the AIOs of 19 real subjects and assessed their effectiveness. In particular, we have shown that the overall performance of each AIO in assessing the four considered image distortions is quite close to the one achieved by the related real subject, which is mimicked. Moreover, unlike the previously published deep CNN-based AIOs, the proposed AIOs have a sensitivity to the noise and image color saturation that is more similar to that of real subjects.

Future work will consider the use of very recent large language model-inspired architectures as well as ViT architectures to potentially enhance the AIOs performance as they have shown outstanding performance in IQA [62, 64].

ACKNOWLEDGMENT

Some of the computational resources were provided by HPC@POLITO (http://www.hpc.polito.it).

REFERENCES

[1] Christos G. Bampis, Zhi Li, and Alan C. Bovik. 2018. Spatiotemporal feature integration and model fusion for full reference video quality assessment. *IEEE Trans. Circ. Syst. Video Technol.* 29, 8 (2018), 2256–2270.

[2] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek. 2018. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.* 27, 1 (2018), 206–219.

[3] Chaofeng Chen and Jiadi Mo. 2022. IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment. Retrieved from https://github.com/chaofengc/IQA-PyTorch

[4] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J. Karam. 2011. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. Broadcast.* 57, 2 (2011), 165–182. https://doi.org/10.1109/TBC.2011.2104671

[5] Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* 25, 5 (2013), 845–869.

[6] Yixuan Gao, Xiongkuo Min, Wenhan Zhu, Xiao-Ping Zhang, and Guangtao Zhai. 2023. Image quality score distribution prediction via alpha stable model. *IEEE Trans. Circ. Syst. Video Technol.* 33, 6 (2023), 2656–2671. https://doi.org/10.1109/TCSVT.2022.3229839

[7] Yixuan Gao, Xiongkuo Min, Yucheng Zhu, Jing Li, Xiao-Ping Zhang, and Guangtao Zhai. 2022. Image quality assessment: From mean opinion score to opinion score distribution. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM, 997–1005.

[8] Yixuan Gao, Xiongkuo Min, Yucheng Zhu, Jing Li, Xiao-Ping Zhang, and Guangtao Zhai. 2022. Image quality assessment: From mean opinion score to opinion score distribution. Retrieved from https://github.com/YixuanGao98/Image-Quality-Assessment-From-Mean-Opinion-Score-to-Opinion-Score-Distribution

[9] Xavier Glorot and Y. Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res. Proc. Track* 9 (Jan. 2010), 249–256.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[11] Tobias Hoßfeld, Poul E. Heegaard, Martín Varela, and Sebastian Möller. 2016. QoE beyond the MOS: An in-depth look at QoE via better metrics and their relation to MOS. *Qual. User Exper.* 1, 1 (Sep. 2016), 1–23. https://doi.org/10.1007/s41233-016-0002-1

[12] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not enough! In *Proceedings of the 3rd International Workshop on Quality of Multimedia Experience (QoMEX'11)*. IEEE, 131–136. https://doi.org/10.1109/QoMEX.2011.6065690

[13] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.* 29 (2020), 4041–4056.

[14] ITU-T. Rec. BT.500. 2012. Methodology for the Subjective Assessment of the Quality of Television Pictures. https://www.itu.int/rec/R-REC-BT.500

[15] ITU-T. Rec. P.910. 2008. Subjective Video Quality Assessment Methods for Multimedia Applications. https://www.itu.int/rec/T-REC-P.910

[16] ITU-T. Rec. P.913. 2021. Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment. https://www.itu.int/rec/T-REC-P.913

[17] L. Janowski and Z. Papir. 2009. Modeling subjective tests of quality of experience with a Generalized Linear Model. In *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEX'09)*. IEEE, 35–40. https://doi.org/10.1109/QOMEX.2009.5246979

[18] Lucjan Janowski and Margaret Pinson. 2015. The accuracy of subjects in a quality experiment: A theoretical subject model. *IEEE Trans. Multimedia* 17, 12 (2015), 2210–2224.

[19] Dinesh Jayaraman, Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. 2012. Objective quality assessment of multiply distorted images. In *Proceedings of the Conference Record of the 46th Asilomar Conference on Signals, Systems and Computers (ASILOMAR'12)*. IEEE, 1693–1697.

[20] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1733–1740.

[21] Jari Korhonen. 2019. Assessing personally perceived image quality via image features and collaborative filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 8169–8177.

[22] Jari Korhonen. 2019. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.* 28, 12 (2019), 5923–5938. https://doi.org/10.1109/TIP.2019.2923051

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. AIP, 1097–1105.

[24] Christopher Lennan, Hao Nguyen, and Dat Tran. 2018. Image Quality Assessment. Retrieved from https://github.com/idealo/image-quality-assessment

[25] Jing Li, Suiyi Ling, Junle Wang, and Patrick Le Callet. 2020. A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 3339–3347.

[26] Z. Li and C. G. Bampis. 2017. Recover subjective quality scores from noisy measurements. In *Proceedings of the Data Compression Conference (DCC'17)*. IEEE, 52–61. https://doi.org/10.1109/DCC.2017.26

[27] Zhi Li, Christos G. Bampis, Lucjan Janowski, and Ioannis Katsavounidis. 2020. A simple model for subject behavior in subjective experiments. *Electr. Imag.* 2020, 11 (2020), 131–1.

[28] Lixiong Liu, Tianshu Wang, Hua Huang, and Alan Conrad Bovik. 2020. Video quality assessment using space–time slice mappings. *Signal Process.: Image Commun.* 82 (2020), 115749.

[29] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. 2017. RankIQA: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, 1040–1049.

[30] MATLAB. 2023. Pretrained Vision Transformer (ViT) Neural Network. Retrieved from https://it.mathworks.com/help/vision/ref/visiontransformer.html

[31] Karan Mitra, Arkady Zaslavsky, and Christer Ahlund. 2015. Context-aware QoE modelling, measurement and prediction in mobile computing systems. *IEEE Trans. Mobile Comput.* 14 (May 2015), 920–936.

[32] A. Mittal, A. K. Moorthy, and A. C. Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 21, 12 (2012), 4695–4708.

[33] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. 2008. Which semi-local visual masking model for wavelet-based image quality metric? In *Proceedings of the 15th IEEE International Conference on Image Processing*. IEEE, 1180–1183.

[34] Da Pan, XueTing Wang, Ping Shi, and ShaoDe Yu. 2021. No-reference video quality assessment based on modeling temporal-memory effects. *Displays* 70 (2021), 102075.

[35] Judith A. Redi, Yi Zhu, Huib de Ridder, and Ingrid Heynderickx. 2015. *How Passive Image Viewers Became Active Multimedia Users*. Springer International Publishing, Cham, Switzerland, Chapter 2, 31–72.

[36] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Adv. Neural Info. Process. Syst.* 31 (2018), 1–11.

[37] M. Seufert. 2019. Fundamental advantages of considering quality of experience distributions over mean opinion scores. In *Proceedings of the 11th International Conference on Quality of Multimedia Experience (QoMEX'19)*. IEEE, 1–6. https://doi.org/10.1109/QoMEX.2019.8743296

[38] HR Sheikh. 2005. LIVE Image Quality Assessment Database Release 2. Retrieved from http://live.ece.utexas.edu/research/quality

[39] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. 2005. LIVE image quality assessment database. Retrieved from http://live.ece.utexas.edu/research/quality

[40] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. Retrieved from https://arXiv:1409.1556

[41] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* 1 (2022), 1–19. https://doi.org/10.1109/TNNLS.2022.3152527

[42] Robert C. Streijl, Stefan Winkler, and David S. Hands. 2016. Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives. *Multimedia Syst.* 22, 2 (Mar. 2016), 213–227. https://doi.org/10.1007/s00530-014-0446-1

[43] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE, 3664–3673. https://doi.org/10.1109/CVPR42600.2020.00372

[44] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE Trans. Image Process.* 27, 8 (2018), 3998–4011. https://doi.org/10.1109/TIP.2018.2831899

[45] David S. Taubman, Michael W. Marcellin, and Majid Rabbani. 2002. JPEG2000: Image compression fundamentals, standards and practice. *J. Electr. Imag.* 11, 2 (2002), 286–287.

[46] Lohic Fotio Tiotsop, Florence Agboma, Glenn Van Wallendael, Ahmed Aldahdooh, Sebastian Bosse, Lucjan Janowski, Marcus Barkowsky, and Enrico Masala. 2021. On the link between subjective score prediction and disagreement of video quality metrics. *IEEE Access* 9 (2021), 152923–152937. https://doi.org/10.1109/ACCESS.2021.3127395

[47] L. F. Tiotsop, E. Masala, A. Aldahdooh, G. Van Wallendael, and M. Barkowsky. 2019. Computing quality-of-experience ranges for video quality estimation. In *Proceedings of the 11th International Conference on Quality of Multimedia Experience (QoMEX'19)*. IEEE, Berlin, Germany, 1–3. https://doi.org/10.1109/QoMEX.2019.8743303

[48] Lohic Fotio Tiotsop, Tomas Mizdos, Marcus Barkowsky, Peter Pocta, Antonio Servetti, and Enrico Masala. 2022. Mimicking individual media quality perception with neural network based artificial observers. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 1 (2022), 1–25.

[49] Lohic Fotio Tiotsop, Tomas Mizdos, Enrico Masala, Marcus Barkowsky, and Peter Pocta. 2021. How to train no-reference video quality measures for new coding standards using existing annotated datasets?. In *Proceedings of the IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP'21)*. IEEE, 1–6. https://doi.org/10.1109/MMSP53017.2021.9733456

[50] Lohic Fotio Tiotsop, Tomas Mizdos, Miroslav Uhrina, Marcus Barkowsky, Peter Pocta, and Enrico Masala. 2020. Modeling and estimating the subjects' diversity of opinions in video quality assessment: a neural network based approach. *Multimedia Tools Appl.* 80 (2020), 1–19.

[51] L. Fotio Tiotsop, T. Mizdos, M. Uhrina, P. Pocta, M. Barkowsky, and E. Masala. 2020. Predicting single observer's votes from objective measures using neural networks. In *Proceedings of Human Vision and Electronic Imaging Conference (HVEI'20)*. Society for Imaging Science and Technology (IS&T), 130–1 – 130–8,.

[52] Lohic Fotio Tiotsop, Antonio Servetti, Marcus Barkowsky, and Enrico Masala. 2022. Regularized maximum likelihood estimation of the subjective quality from noisy individual ratings. In *Proceedings of the 14th International Conference on Quality of Multimedia Experience (QoMEX'22)*. IEEE, 1–4. https://doi.org/10.1109/QoMEX55416.2022.9900903

[53] Lohic Fotio Tiotsop, Antonio Servetti, Marcus Barkowsky, Peter Pocta, Tomas Mizdos, Glenn Van Wallendael, and Enrico Masala. 2023. Predicting individual quality ratings of compressed images through deep CNNs-based artificial observers. *Signal Process.: Image Commun.* 112 (2023), 116917.

[54] Lohic Fotio Tiotsop, Antonio Servetti, and Enrico Masala. 2020. Full reference video quality measures improvement using neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20)*. IEEE, 2737–2741. https://doi.org/10.1109/ICASSP40776.2020.9053739

[55] Domonkos Varga, Dietmar Saupe, and Tamás Szirányi. 2018. Deeprn: A content preserving deep architecture for blind image quality assessment. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'18)*. IEEE, 1–6.

[56] Vishal Thanvantri Vasudevan, Abhinav Sethy, and Alireza Roshan Ghias. 2019. Towards better confidence estimation for neural models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'19)*. IEEE, 7335–7339.

[57] Anish Kumar Vishwakarma and Kishor M. Bhurchandi. 2022. No-reference video quality assessment using novel hybrid features and two-stage hybrid regression for score level fusion. *J. Visual Commun. Image Represent.* 89 (2022), 103676.

[58] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (2004), 600–612.

[59] Qianqian Xu, Jiechao Xiong, Qingming Huang, and Yuan Yao. 2014. Online hodgerank on random graphs for crowdsourceable QoE evaluation. *IEEE Trans. Multimedia* 16, 2 (2014), 373–386. https://doi.org/10.1109/TMM.2013.2292568

[60] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 1191–1200.

[61] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE, Seattle, Washington, 3572–3582.

[62] Junyong You and Jari Korhonen. 2021. Transformer for image quality assessment. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'21)*. IEEE, Anchorage, Alaska, 1389–1393.

[63] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. 2018. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circ. Syst. Video Technol.* 30, 1 (2018), 36–47.

[64] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui et al. 2023. Q-Boost: On visual quality assessment ability of low-level multi-modality foundation models. Retrieved from https://arXiv:2312.15300

[65] Yi Zhu, Sharath Chandra Guntuku, Weisi Lin, Gheorghita Ghinea, and Judith A. Redi. 2018. Measuring individual video QoE: A survey, and proposal for future directions using social media. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 2s (2018), 1–24.

Received 17 November 2023; revised 28 March 2024; accepted 26 April 2024