POLITECNICO DI TORINO
Repository ISTITUZIONALE

Predicting individual quality ratings of compressed images through deep CNNs-based artificial observers

Original
Predicting individual quality ratings of compressed images through deep CNNs-based artificial observers / FOTIO TIOTSOP, L., Servetti, A., Barkowsky, M., Pocta, P., Mizdos, T., Van Wallendael, G., Masala, E.. - In: SIGNAL PROCESSING-IMAGE COMMUNICATION. - ISSN 0923-5965. - STAMPA. - 112:(2023). [10.1016/j.image.2022.116917]

Availability:
This version is available at: 11583/2974692 since: 2023-01-17T09:09:08Z

Publisher:
Elsevier

Published
DOI:10.1016/j.image.2022.116917

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
Elsevier postprint/Author's Accepted Manuscript

© 2023. This manuscript version is made available under the CC-BY-NC-ND 4.0 license
http://creativecommons.org/licenses/by-nc-nd/4.0/.The final authenticated version is available online at:
http://dx.doi.org/10.1016/j.image.2022.116917

(Article begins on next page)

03 November 2026

Predicting Individual Quality Ratings of Compressed Images through Deep CNNs-based Artificial Observers

Lohic Fotio Tiotsop[a,*], Antonio Servetti[a], Marcus Barkowsky[d], Peter Pocta[b], Tomas Mizdos[b], Glenn Van Wallendael[c], Enrico Masala[a]

[a] *Control and Computer Engineering Department, Politecnico di Torino, 10129 Torino, Italy*
[b] *Department of Multimedia and Information-Communication Technology, University of Zilina, Zilina, Slovakia*
[c] *Ghent University - imec, Ghent, Belgium*
[d] *Deggendorf Institute of Technology (DIT), Deggendorf, Germany*

Abstract

Unlike traditional objective approaches aimed at MOS prediction, subjective experiments provide individual opinion scores that allow, for instance, to estimate the distribution of users' opinion scores. Unfortunately, the current literature is lacking objective quality assessment approaches that simulate the process of a subjective test. Therefore, this work focuses on modeling an individual subject through a deep CNN that, once trained, is expected to mimic the subject in terms of quality perception; for this reason, we call it "Artificial Intelligence-based Observer" (AIO). Several AIOs, modeling subjects with different characteristics, can be derived and used to simulate the process of a subjective test, thus yielding a more complete objective quality assessment. However, the training of the AIOs is hindered by two major issues: i) the lack of training sets containing a large number of individual opinion scores; ii) the noisy nature of individual opinion scores used as ground truth. To overcome these issues, we motivate a two-step learning approach. During the first learning step, the architecture of the well-known ResNet50 is appropriately modified and its initial weights are updated using a large scale synthetically annotated dataset of JPEG compressed images created for quality assessment purpose. This yields a new deep CNN called JPEGResNet50 that can accurately evaluate the qual-

*Corresponding author

ity of JPEG compressed images. The second learning step, conducted on a subjectively annotated dataset, refines the generic perceptual quality features already learned by the JPEGResNet50 to derive the AIO of each subject. Extensive computational experiments show the potential and effectiveness of our approach.

Keywords: image quality assessment, AI observer, deep neural network, transfer learning

---

1. Introduction

The last few decades witnessed a remarkable growth in the amount of multimedia data generated and exchanged every day over the Internet [1]. This resulted in a large interest in the research field of media quality assessment. In fact, accurate objective metrics, i.e. those capable of predicting visual quality as perceived by human observers, allow to optimize multimedia processing systems while guaranteeing high Quality of Experience (QoE) to the end users.

Following the success of Machine Learning (ML) models and algorithms for other tasks related to multimedia data processing, e.g. image segmentation [2], image classification [3], image denoising [4], video scene classification and segmentation [5], the research in the media quality assessment field has naturally adopted these methods [6, 7, 8, 9].

Even though the output of a subjective experiment provides more information than the simple Mean Opinion Score (MOS), in particular the ratings of each individual observer, the use of ML has been mainly restricted to predicting the MOS. Only very recently, ML approaches have been applied to the problem of modeling and predicting individual opinion scores in media quality assessment [10, 11].

Designing and training an ML-based model that can later be used as a substitute of an individual observer in order to automatically reproduce his/her perception of quality is a recent and very promising research direction [11]. Such a model is referred to as an Artificial Intelligence-based Observer (AIO) in

the rest of this work. [11]. Once many AIOs (one for each subject with specific characteristics) have been trained, they allow to perform an objective assessment of the quality that resembles more to a subjective test. In fact, the predictions of these different AIOs yield individual opinion scores from which important information can be derived, in addition to the MOS. For instance: i) service providers could use predicted individual opinion scores to accurately estimate the percentage of unsatisfied customers regarding the perceptual quality of a given processed video sequence (PVS); ii) based on the quality score predicted by each AIO, it would be possible to make inference on the characteristics of the customers who would not be satisfied with the quality of the content under evaluation; iii) the trained AIOs could be used to simulate subjective tests in order to investigate the possible presence of peculiarities in a dataset of stimuli before using it for an actual subjective test.

An important number of researchers within the media quality assessment community focuses on modeling and explaining subjects' behavior in subjective experiments and the statistical properties of stimuli [12, 13, 14]. In that research context, an aggregated measure such as the MOS is not relevant as an input for the analysis. Instead, individual opinion scores coming from subjective tests are required. As the subjective tests are resource demanding, the possibility to automatically generate individual opinion scores with AIOs would definitely ease the development of new and potentially more effective models and tools.

While the approach of mimicking individual observers with AIOs yields more advantages as compared to MOS prediction-based approaches, its implementation in practice is hindered by two main issues: i) the lack of subjectively annotated datasets that include a large number of opinion scores expressed by the same subject; ii) the noisy nature of individual opinion scores caused by the subject inconsistency, i.e., the subject's inability to systematically repeat his/her first opinion score when asked for many ratings of the same stimuli [12].

One of the reasons why the literature has long been focusing on predicting the MOS rather than individual opinions is the noisy nature of the individual ratings. In fact, the arithmetic mean operation that leads to the MOS is aimed

precisely at mitigating the effect of that noise. Unfortunately, when it comes to training AIOs, one cannot get rid of that noise. This leads to learning tasks with noisy labels, which are known to be particularly challenging, especially if one does not have an abundance of training samples.

In a recent journal paper [11], to train the AIOs, we employed neural networks (NNs) with a very simple architecture in order to not overfit the few available training samples and thus not to learn the noise that affects the individual opinion scores. We trained a feedforward NN with no more than three hidden layers and few neurons for each observer. Being very simple NNs, they take, as an input, few hand-crafted features. This approach that exploits the hand-crafted features to model individual observers however suffers the following two main issues:

1. **Inaccuracy caused by reducing complex input signals to only a set of hand-crafted features.** The hand-crafted features are meant to provide concise information that approximates as accurately as possible the raw input content. This approximation step can however cause some inaccuracy that could be avoided if the AIOs would be trained and tested by feeding the model directly with the raw content as an input.
2. **Over-generalization of the hand-crafted features with respect to individual subject's characteristics.** When using the hand-crafted features, the same algorithms are used for extracting the features that model different subjects even if they do not have a similar interpretation of the same artifacts and hence do not express their opinion score based on the same reasoning. A good example in this case is the perception of blur, which may differ in between subjects, i.e., blurring of edges, blurring of texture, etc., but many objective algorithms only have a single (oversimplified) indicator such as the Gaussian blur.

To the best of our knowledge, this work is the first to consider the much more challenging task of designing deep Convolutional Neural Network (CNN)-based AIOs. By relying on deep CNNs, we get rid of the hand-crafted features and

feed the model directly with the raw content. Furthermore, the features that model each subject are directly extracted by the convolutional layers during the learning process. Therefore, the extracted features for each subject depend on his/her characteristics as they are computed based on his/her opinion scores that are used as ground truth labels.

To overcome the challenges caused by the lack of training samples and noisy nature of the learning task that yields the AIOs, we propose a two-steps learning approach whose contribution to advancing the state-of-the-art is threefold:

1. We show how a small scale subjectively annotated dataset can be leveraged to create a large scale synthetically annotated one useful for pre-training deep CNNs for image quality assessment and thus overcoming the challenges imposed by the lack of training samples.
2. The created large-scale dataset is used to train a deep CNN, that we name JPEGResNet50, with more than 50 convolutional layers, that can accurately assess the perceptual quality of the JPEG compressed images. We would like to recommend the use of our JPEGResNet50 as the basis of transfer leaning within the media quality assessment community instead of starting from pre-trained deep CNNs for the image classification task as done so far in the literature [15]. In fact, the JPEGResNet50 would represent a better starting point since it can already extract useful features for a perceptual quality prediction right from the beginning of the transfer learning process.
3. Starting from the JPEGResNet50, relying on a transfer learning approach, 19 AIOs, i.e., deep CNNs with the same architecture as that of the JPEGResNet50 but including different learned weights, are trained, each modeling an individual subject. These 19 deep CNNs-based AIOs and the JPEGResNet50 are made freely available for research purposes at `http://media.polito.it/AIobservers`.

In practice, starting from a small scale subjectively annotated dataset, i.e., the LIVE image quality assessment database [16], and the ImageNet compe-

tition dataset [17], we created a large scale dataset for the quality assessment task containing 500,000 synthetically annotated JPEG compressed images . We then designed a deep CNN architecture similar to that of the ResNet50 [18], except for the fully connected and softmax layers that were designed to output a five classes discrete probability distribution on the Absolute Category Rating (ACR) scale. We initialized the weights of the convolutional layers of such an architecture with those of the ResNet50 whereas the weights of the fully connected layer were randomly initialized. We then trained the network on the created quality assessment large scale dataset in order to obtain our JPEGResNet50. During this learning process, a small learning rate was used in order to progressively update the initial weights of the network and thus transform the previously learned object detection features into new ones, useful for predicting the perceptual quality of the JPEG compressed images.

The training process of the JPEGResNet50 constitutes the first learning step in the process that yielded the derivation of 19 deep CNNs-based AIOs. In fact, the JPEGResNet50 automatically identifies and extracts important features for image quality prediction, hence it can be considered as a suitable starting point for training other deep CNNs aimed at performing image quality assessment (IQA) of compressed images as an actual observer would do.

The second learning step was conducted on the data collected during the first phase of the "LIVE Multiply Distorted Image Quality" (LIVE-MD-ph1) experiment [19]. We modeled each observer from that experiment using a deep CNN with weights readjusted/refined from those of the JPEGResNet50 through transfer learning. By doing so, we obtained 19 deep CNNs, one for each observer. These deep CNNs take an image as input and predict the opinion score on the ACR scale that the corresponding observer would have expressed after evaluating the quality of that same image.

Extensive computational experiments have been conducted in order to assess the accuracy of the 19 trained deep CNNs-based AIOs as well as that of the JPEGResNet50. When compared to several state-of-the-art objective measures, it was observed that the JPEGResNet50 is particularly suitable to assess the

quality of the JPEG compressed images. Each AIO can mimic, with a rather good accuracy, actual observers yielding state-of-the-art performance in terms of MOS prediction while also providing an estimation of the distribution of users' opinion scores.

The remainder of the paper is organized as follows. Section 2 presents related work while highlighting the relevance and innovativeness of our approach. In Section 3 the training process of the deep CNN-based AIOs is described in detail. Computational experiments and the related results are presented in Section 4, while conclusions are drawn in Section 5.

2. Related Work

The MOS obtained from subjective experiments has long been considered a highly reliable measure of perceptual quality. Objective metrics have therefore usually been developed for predicting the MOS as accurately as possible. Recently, however, several papers appeared in the literature underlining the limits of the MOS as a comprehensive measure of QoE while proposing more complete approaches for an effective objective evaluation of the perceptual quality as assessed by the end users [20, 21, 22, 23]. In [24], the authors argued that for a given objective quality score, the corresponding subjective quality should be considered a probability distribution. In other words, the MOS, being a single value, does not capture all the aspects that contribute to measuring QoE. Following the same direction, the authors of [25] illustrated the need to evaluate quality by referring to the entire Distribution of the Opinion Scores (DOS) expressed by observers rather than limiting the evaluation to the MOS. Some authors have attempted to predict the DOS [26, 27, 23]. Although the DOS provides more information than the MOS, it is still an aggregated measure and thus less informative than a measure that acts at the subject level, i.e., a measure that predicts individual opinion scores.

According to the ITU [28] and also the Qualinet white paper [29], a measure of QoE must be able to give indications on the level of satisfaction of the end

user while taking into account his/her personality and expectations. The fact that the user's personality and expectations are mentioned above implies that any measure of QoE should ideally perform at the single subject level. Subjective experiments satisfy such a requirement, since during the test, each single observer independently expresses his/her opinion scores based on his experience and background. It has been shown that the analysis of individual opinion scores coming from subjective tests allows to gain important information on the behavior of human subjects when rating the perceptual quality, and also to measure the intrinsic ability of a stimulus to confuse viewers with respect to its perceptual quality [12, 13].

Despite the usefulness of individual opinion scores, the design of objective approaches that allow to measure the quality at the single subject level is still in its early stage [10, 11]. This work aims at advancing the state-of-the-art in this direction. Objectively measuring the quality at the level of the single subject means being able to train a model that can mimic the perception of the quality of an individual. Unfortunately, the factors that influence or determine the perception of the quality of an individual are numerous, complex and even subject to uncertainty [30]. This makes it difficult to design an exhaustive set of hand-crafted features that can accurately model the quality perception of any individual subject. So, a natural solution in this case is to rely on deep CNNs that extract useful features directly from the data during the training process.

Deep CNNs have been largely used in media quality assessment despite the lack of large scale subjectively annotated datasets necessary for an effective training process [31, 32, 33, 34, 35, 36]. To overcome the lack of training samples, many authors relied on transfer learning approaches [37, 38] and data augmentation methods [39, 40, 41, 42]. The computer vision community has developed a number of data augmentation approaches [43]. Most of these approaches represent an implementation of a set of rules that, applied to an entity of the training set (image, video, audio), creates an additional entity that is expected to have the same label. For example, in an image classification task, a translation, rotation, and scaling of the object in an image does not change

its content and therefore its label.

As it can be seen from the latter example, the typical data augmentation approaches adopted in the computer vision community mainly affect the geometry of the elements present in multimedia content. While this type of modification can generate particularly challenging samples from the point of view of the computer vision tasks, they may not constitute a significant added value for the training of a model aimed for predicting perceptual quality. In fact, a modification of geometrical shape of the objects alone keeps unchanged features such as contrast, resolution, spatial and temporal activity and also quantity of motion (in case of video), which are important for visual quality assessment. The model could therefore perceive these new samples as substantially equivalent to the initial one from which they were generated.

For this reason, alternative approaches for generating more data in order to effectively train ML based models in the media quality assessment has been proposed. In [39, 40], in addition to the subjectively annotated training samples, the authors created new training samples for which they computed objective measures to be used as a substitute of the MOS. The authors in [41, 42], instead, proposed an approach to combine different subjectively annotated datasets into a single larger one, thus overcoming the issues stemming from the different contexts in which the subjective experiments have been conducted. In particular, the MOS values have to be realigned to take into consideration the context influence factors that may affect the result of each experiment in several different ways [44]. Therefore, the MOS values in the newly created dataset are, in practice, only estimates of the ones that would be expected while running a single large subjective experiment.

Nevertheless, using these approximated MOS values as ground truth data does not preclude the possibility of obtaining an effective model as long as such approximation turns out to be a highly probable realization of the actual subjective scores. This can be seen from the fact that accurate objective metrics can still be constructed if the annotated training set is augmented with data not necessarily collected during a single subjective experiment. This is because the

MOS computed from the ratings of a few subjects, as typically done in practice, is itself a value affected by a random measurement error [45].

Similar to most of the papers reviewed above, this work relies on deep CNNs and leverages the transfer learning concept and a data augmentation approach to cope with the lack of training samples. However, it differs from the previously published papers by two main aspects:

1. To the best of our knowledge, this is the first work in the media quality assessment that focuses on training a deep CNN with more than 50 hidden layers that can mimic an individual observer. Because of the lack of training samples and noisy nature of the individual opinion scores caused by the subject inconsistency [12, 13, 45], the considered learning task is much more challenging and data demanding in comparison with the training process of models aiming at MOS prediction.
2. Unlike the previous data augmentation approaches designed for media quality assessment that combine different subjectively annotated datasets [41, 42] or annotate the stimulus with the scores predicted by objective metrics [39, 40], this work exploits a small scale subjectively annotated dataset to figure out an annotation rule that is then used to synthetically create a large number of training samples. It is worth noting here that such a number is several decades larger than what can be subjectively annotated.

3. A Two-step Learning Approach to Train Deep CNNs-based AIOs

3.1. Introduction and Motivation of our Approach

Let us assume that we want to train a deep CNN that can mimic the quality perception of one specific subject that we will call Bob. To this aim, Bob is invited to a subjective test during which he is asked to watch and rate N stimuli. In practice, due to time constraints and Bob's fatigue, N is usually not too large. Therefore, a dataset that includes only the N stimuli subjectively rated by Bob is typically not large enough to effectively use it to train, from

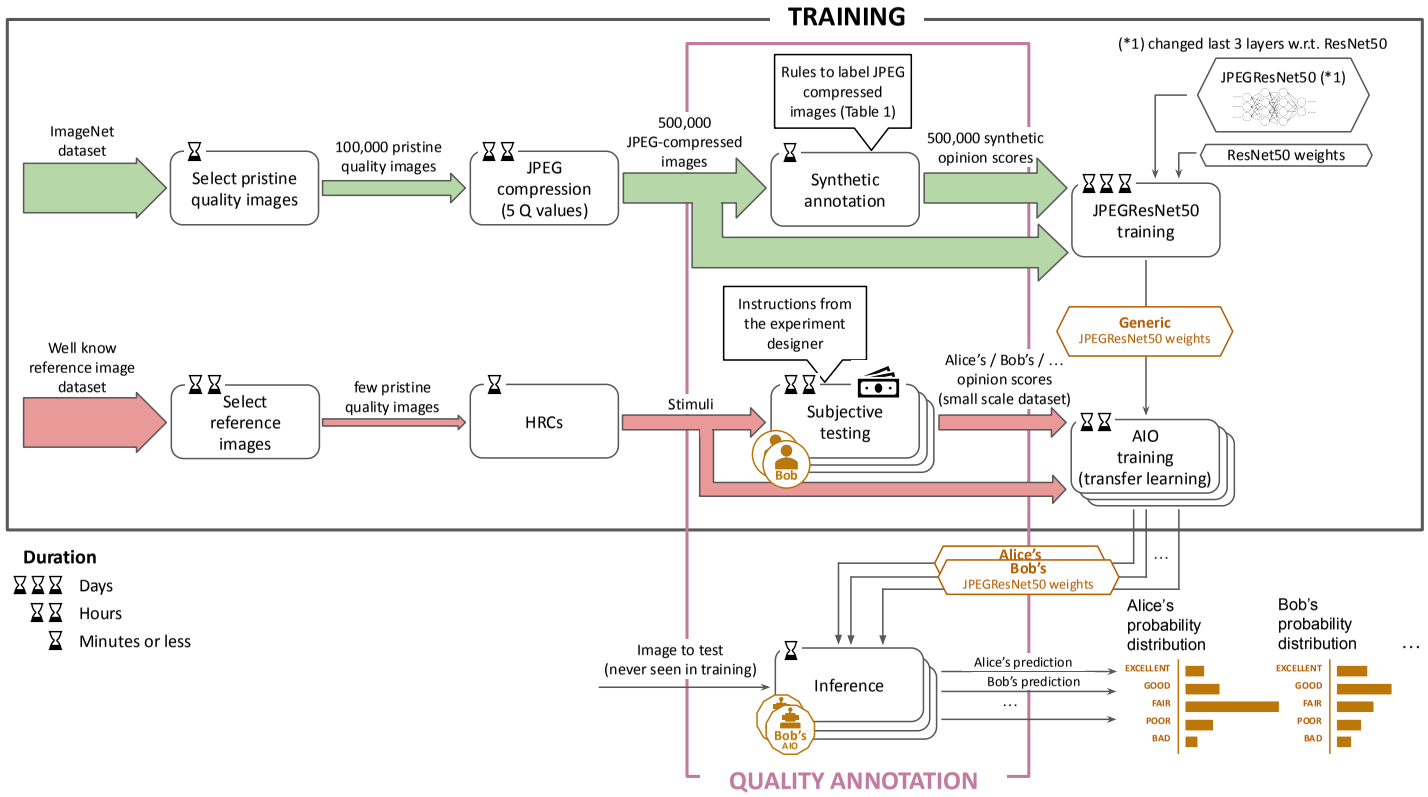


Figure 1: The diagram summarizes our proposed approach to train deep CNNs-based AIOs. The training process involves two learning steps. During the first learning step (see green arrows), the JPEGResNet50 is trained to extract complex features for perceptual quality prediction on a large scale dataset of the JPEG compressed images with synthetic labels. The JPEGResNet50's architecture and trained weights are then used during the second learning step (see red arrows) as the starting point to train, on a small scale subjectively annotated dataset, the deep CNN-based AIO of each individual subject. Once trained, during the inference phase, the AIO of a subject predicts the five probabilities with which he/she would choose any of the five alternatives on the ACR scale when asked to rate the quality of the input image. The difference in size between the green arrows and the red ones highlights the fact that the number of synthetically annotated training samples used for the first learning step is far more larger than the subjectively annotated one available for the second learning step.

scratch, a deep CNN that learns, from Bob's ratings, the way he perceives and rates quality.

In the literature this lack of training samples is very often addressed by relying on the transfer learning concept. For example, in the MOS prediction task the authors typically choose a deep CNN pre-trained for the image classification task on one of the large-scale datasets available within the computer vision community, e.g. the ImageNet dataset [17]. A single learning step, that starts from the weights of the chosen pre-trained network (a transfer learning), is usually performed on a small scale subjectively annotated dataset to transform the object detection features of the pre-trained network into perceptual quality features useful for MOS prediction.

However, we experimentally observed that such a single learning step, that is sufficient to obtain accurate models for MOS prediction, is not enough when modeling the quality perception of a single subject relying on the small number of training samples available in the existing state-of-the-art subjectively annotated datasets (see the results in Table 2 for more detail).

For this reason, to model individual subjects instead of the MOS, in this work we propose to rely on a two-step learning approach, which is summarized in the diagram depicted in Figure 1. In particular, according to the proposed approach, the Bob's AIO is derived as follows:

First learning step (depicted/represented by the green arrows in Figure 1):

We created a large scale synthetically annotated quality assessment dataset of JPEG compressed images and used it to train a deep CNN that we named JPEGResNet50. The JPEGResNet50 is a deep CNN, with more than 50 convolutional layers, trained to score the perceptual quality of JPEG compressed images on the five point ACR scale. Therefore, thanks to its large number of convolutional layers, it is expected to extract detailed perceptual quality features useful as the basis for transfer learning when designing a model to predict the quality of compressed images.

Second learning step (depicted/represented by the red arrows in Figure 1):

We fine-tuned the weights of the JPEGResNet50 but this time performing the training on a small scale dataset annotated by Bob. During this last learning phase, by leveraging Bob's opinion scores, the generic perceptual features already learned by the JPEGResNet50 are progressively updated and refined to yield new ones that allow to model Bob's perception of quality. The final deep CNN that has the same architecture as the JPEGResNet50 but different trained weights is what we called Bob's deep CNNs-based AIO.

Our two-step learning approach summarized above can be motivated as follows. In our first learning step, we start from the ResNet50 that has been trained to perform an image classification task. This network is therefore designed to extract low- and high-level features that characterize the objects included in its training set. We believe that these features might be not suitable for modeling the perceived quality of images if not opportunely updated. For instance, in an image classification task, it is expected that the first layers of the network capture as much as possible the fact that the presence, in the image, of some defects such as noise, blur and blocking artifacts due to compression should not change the prediction of the network. Indeed, networks trained for image classification are sometimes compared on the basis of their robustness to the artifacts present in the input image. However, this expectation from the first layers of an image classification network is totally opposite to what a network designed to model the perception of quality is supposed to do. In fact, networks for perceptual quality modeling are supposed to learn that the presence of artifacts impacts the final prediction.

Therefore, when starting from a network trained for image classification to train a new one for quality assessment, the training set must be large enough to allow the network to effectively learn from and thus progressively transform object detection features into quality assessment ones. Some researchers in the media quality assessment community, e.g. [46, 27], successfully performed a single learning step on the ResNet architecture to reach models capable of

predicting the MOS without relying on an additional learning step as we did in this work. However, in all these papers the authors relied on datasets containing thousands of training samples and considered a learning task with different characteristics than the one studied in this paper.

For our learning task, i.e. modeling individual subjects, we unfortunately did not find any publicly available dataset in which the same subject rated thousands of pictures. As we have already highlighted in [11], how to design large scale subjective tests tailored to the training of the AIOs is still an open research issue. A related question is, for instance: how to manage the subject's fatigue during such experiments, since the same subject is supposed to rate several stimuli? Moreover, the training of models for the MOS prediction is performed with less noisy labels than those used in our considered application. In fact, the arithmetic mean operation mitigates the noise in individual opinion scores. By learning from individual scores, we are considering a learning task that is more demanding in terms of training samples. In fact, as highlighted in this paper [47], learning tasks performed with noisy labels require more complex model architectures but also more training samples.

To overcome these challenges posed by the considered learning task, i.e. the limited size of the training sets and the noisy nature of the labels, a possible solution is to perform a preliminary learning step on a synthetically annotated quality assessment large-scale dataset. This first learning step retrains the ResNet50 on a large-scale quality assessment dataset, i.e. our synthetically annotated dataset of JPEG compressed images, after its original training on the ImageNet dataset. This helps, first of all, in changing the high level representations (in the last layers) of the ResNet50 towards the representations that we need for our purpose, i.e. image quality assessment rather than object detection, but it also allows a progressive update of some aspects of the low-level features (in the first layers of the network), which reflect the image classification task but do not characterize at all the perceptual quality assessment task, e.g. the lack of sensitivity to the defects in the input image. After this first learning step, one can then use a second learning step to simply fine-tune the trained

network with the few subjectively annotated samples in the available training sets. That is the logic behind our two-steps learning approach.

In short, the first learning step updates the whole ResNet50, transforming it into a new network, i.e. our JPEGResNet50, that can assess the quality of JPEG compressed images and thus readily extract perceptual quality-aware features. Therefore, the JPEGResNet50 is much more convenient as starting point in a second learning step to derive the AIOs considering that their training occurs with few samples having noisy labels.

We note that in order to transform the ResNet50 into the JPEGResNet50 and thus get a network tailored for the quality assessment task, we made use only of the JPEG compressed images during this preliminary learning step. This was done to make sure that the resulting training set will not be too noisy. In fact, we decided to consider only a well-known and well-researched type of distortion, i.e. JPEG compression, for which we could easily figure out from the prior art whether the rule used to synthetically annotate the created large scale dataset was reasonable. At first glance, this could be perceived only as a limit of our approach. However, incorporating several types of distortions at the first learning step, and not being able to find an accurate degradation-to-quality mapping could have yielded a much noisier training set; and hence, the trained network at the first learning step could have required much more subjectively annotated data to be perfectly fine-tuned at the second learning step. This would have been a serious issue, since, as already mentioned, we could not find any large-scale subjectively annotated dataset to be used for the second learning step.

3.2. Creating a large-scale synthetically annotated training set

We now describe our approach to create a large scale synthetically annotated dataset from a small scale subjectively annotated one. This approach represents the first contribution of this work as highlighted in the introduction.

We considered the data gathered during phase 1 of the first release of the LIVE image quality assessment (LIVE-IQA-r1-ph1) experiment [16]. Since we

could not access the JPEG quality parameter value Q used to create the images in the original dataset, the following procedure has been used to estimate it. For each distorted image used during that experiment we computed its PSNR score s, then we compressed the source image using many different JPEG quality parameters Q, each time computing the PSNR value. Finally, we chose the Q value for which the obtained PSNR is the closest to s. In this way we obtained, for each subjectively evaluated image, the JPEG quality parameter Q that corresponds to its MOS.

Figure 2 reports the average perceived quality for each value of the JPEG quality parameter. The average perceived quality represents the mean of the MOS values of all stimuli sharing the same JPEG quality parameter Q. The black curve in the figure was obtained by performing a least square fitting of the Q values to the quality scale using a third order polynomial function. This curve provides indications on how different levels of the JPEG compression can be mapped to the quality scale.

Looking at the corresponding figure, it can be noticed that the viewers did not use the whole quality scale ranging from 0 to 100, as it typically happens

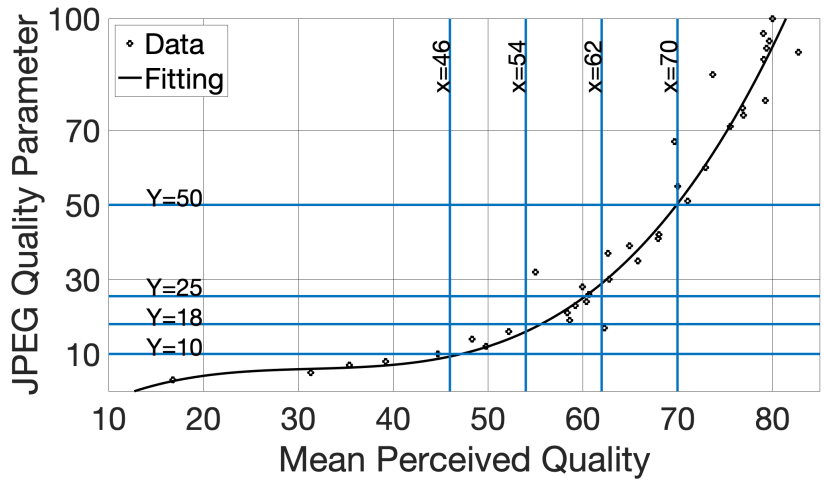


Figure 2: Least square fitting of the JPEG quality parameter to the MOS on the LIVE JPEG image dataset using a third order polynomial function.

in a subjective test, see [44] for more detail. In this case, in which a Double Stimulus Continuous Quality Scale type of experiment has been used, an average quality of about 45 is observed for images compressed in the very low JPEG quality parameter range of 0 to 10. In order to obtain a mapping to the MOS scale ranging from 1 to 5, a clipping is often used for the boundaries and the remaining part is linearly mapped. In particular, the original continuous quality scale ranging from 0 to 100 was converted to the five point ACR scale as follows: any quality score lying in [0, 46] was mapped to "Bad" (1); the interval [46, 70] was divided into three equally large intervals corresponding respectively to "Poor" (2), "Fair" (3) and "Good" (4); finally, any quality score in [70, 100] was considered as "Excellent" (5).

Using the curve depicted in Figure 2, the five attributes of the ACR scale were mapped to the five JPEG quality parameter ranges, yielding the annotation rule reported in Table 1.

It is very important to note here that although the annotation rule in Table 1 derives from an analysis conducted on the results of a subjective test, it cannot be as accurate as a subjective test. However, we are not primarily concerned with its accuracy in predicting quality, but rather with the fact that a deep CNN trained on a large-scale dataset (which includes a large diversity of the content, synthetically annotated by such a rule) would be expected to extract relevant generic quality degradation features that can later be refined by deploying a subjectively annotated dataset if necessary. For instance, as it will be seen later

JPEG Quality parameter interval	**Opinion score**	**Image label**
[2, 10]	1	Bad
[11, 18]	2	Poor
[19, 25]	3	Fair
[26, 50]	4	Good
[51, 100]	5	Excellent

Table 1: Mapping JPEG Quality parameter intervals to the opinion score.

in the results section, the JPEGResNet50 trained on the data annotated by this rule can extract features that are useful to predict the quality of the JPEG compressed images.

Based on the annotation rule defined in Table 1, we created a large-scale synthetically annotated dataset starting from the images available in the ImageNet competition dataset [17] that contains over a million images dedicated to the training and evaluation of deep neural network models for image classification. The steps yielding to the creation of the synthetically annotated dataset of 500,000 JPEG compressed images used as a training set for the JPEGResNet50 are shown in the diagram depicted in Figure 1. We started by selecting 100,000 pristine quality images from the ImageNet dataset. For each of these images, we generated five distorted images by compressing the original image using five different values of the JPEG quality parameter. The five values of the JPEG quality parameter were selected by randomly choosing one value in each one of the five intervals described in Table 1. The quality of each generated image was then annotated with the opinion score associated to the interval to which the related JPEG quality parameter belongs. In the end, we obtained a dataset containing 500,000 annotated images.

3.3. JPEGResNet50 architecture and training process

We used the synthetically annotated large-scale dataset derived in the previous section to train our JPEGResNet50 whose architecture is shown in Figure 3.

For our models, we were first of all looking for existing neural network architectures that have already proven to be effective for predicting the media quality as perceived by human subjects. That is the case for the ResNet architecture that has recently been successfully used by several authors [46, 27]. Furthermore, in the following paper [18], in which the ResNet architecture was described for the first time, the computational experiments provide evidence on the fact that, by relying on such an architecture, the training process is expected to converge much more efficiently. This makes such an architecture suitable for our case, since our approach is highly demanding in terms of training time. In

fact, one needs not only to train a generic deep CNN at the first learning step, but also to train as many deep CNNs as the number of subjects to be modeled at the second learning step.

Therefore, the architecture of our JPEGResNet50 is strongly inspired by that of the well known ResNet50 [18] as both architectures share the same convolutional layers and differ only on the fully connected and softmax layers.

From the JPEGResNet50 architecture shown in Figure 3, it can be seen that the network is designed to receive a $224 \times 224 \times 3$ image patch as an input. Such an input then goes through 52 convolutional layers that are meant to progressively extract more and more detailed perceptual quality features. Once such features are obtained, they are mapped through the fully connected and softmax layers to five values representing the probability with which the quality of the input image will be assessed by an average observer as "Bad" (1), "Poor" (2), "Fair" (3), "Good" (4) or "Excellent" (5). We assume the prediction of the JPEGResNet50 resembles that of an average viewer since the annotation of the training set, as discussed before, is based on the rule defined in Table 1 that maps the JPEG compression levels to the average perceptual quality.

To train the JPEGResNet50, the label of each image i in the artificially created large-scale dataset, was encoded as a binary vector V_i whose entries are defined as follows:

$$V_i(t) = \begin{cases} 1 & \text{if t is the opinion score of image } i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $t = 1, 2, \ldots, 5$.

Denoting by

- I the total number of images in the training set,
- β a vector containing all the weights of the JPEGResNet50 that are to be computed,
- $p_i^t(\beta) \;\; i = 1, 2, \ldots, I, \;\; t = 1, 2, \ldots, 5$ the predicted probability with which the perceptual quality of the image i will be rated as t, given the weights defined in β

the optimization problem guiding the training process of the JPEGResNet50 was formulated as follows:

$$\min_{\beta} \sum_{i=1,2,\ldots,I} \sum_{t=1,2,\ldots,5} -V_i(t) \log(p_i^t(\beta)) \tag{2}$$

$$\sum_{t=1,2,\ldots,5} p_i^t(\beta) = 1 \quad i = 1, 2, \ldots, I \tag{3}$$

$$p_i^t(\beta) \in [0,\ 1]; \quad i = 1, 2, \ldots, I; \quad t = 1, 2, \ldots, 5. \tag{4}$$

Eq. (2) expresses the minimization of the cross entropy, chosen as the cost function, whereas Eq. (3) and (4) establish the fact that the JPEGResNet50 outputs a probability distribution. Note that this constraint is implicitly imposed by the softmax layer inserted in the architecture.

To solve the problem described in Eq. (2)-(4) and thus to train the JPEGResNet50, we relied on the stochastic gradient descent with momentum (SGDM) optimization algorithm. The SGDM was deployed on a batch containing 90 images at each iteration, this was repeated for 60 periods, i.e. a total of $60 \cdot I/90$ iterations. The learning rate and momentum parameter of the SGDM were respectively fixed to 0.0001 and 0.9.

At the end of the training process all the weights, i.e. the entries of the vector β, are known. Therefore, when receiving an image i as an input, the JPEGResNet50 provides as an output the following five probability values: $p_i^t(\beta) \quad t = 1, 2, \ldots, 5$, that represent an estimate of the probability of each of the five possible opinion scores of the ACR scale. An estimation of the MOS of the image i using the JPEGResNet50 can then be expressed as follows:

$$MOS_{res}^i = \sum_{t=1}^{5} t p_i^t(\beta). \tag{5}$$

3.4. Deriving Deep CNNs-based AIOs from the JPEGResNet50

Once the training process of the JPEGResNet50 was completed, as it can be seen from Figure 1, its weights and architecture were used as a starting point for training the deep CNNs-based AIOs.

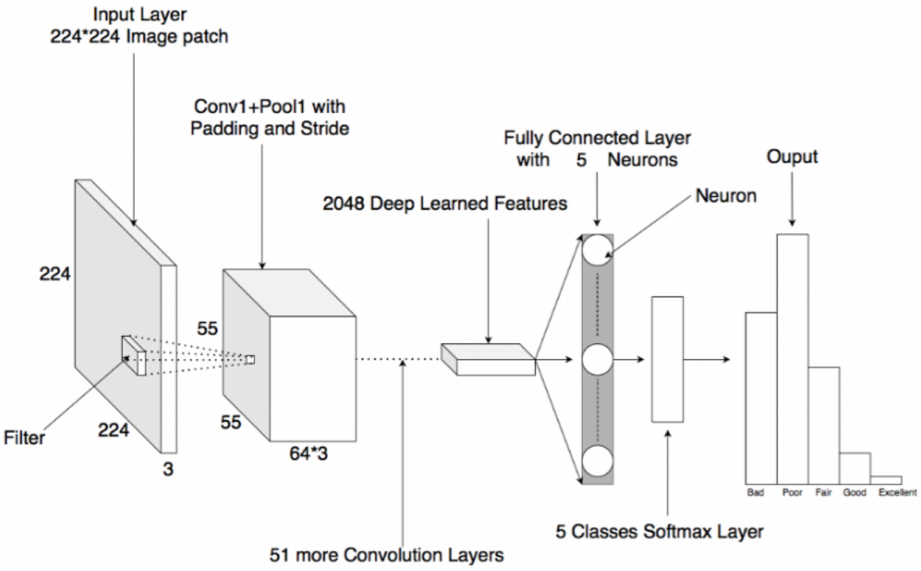


Figure 3: Architecture of the JPEGResNet50 as well as of the AIOs. The JPEGResNet50 receives as input a 224×224 color image and provides as output an estimation of probability with which an average viewer chose any of the five alternative of the ACR scale.

We considered the data collected during the LIVE-Multi-Distortion phase 1 (LIVE-MD-ph1) experiment [19], which includes 19 observers that rated the perceptual quality of 240 images distorted by JPEG compression and blurring artifacts. We would have preferred to perform our second learning step also on a dataset including only the compressed images in order to remain fully coherent with the first learning step. Unfortunately, we did not find any freely available subjectively annotated dataset of compressed images with a sufficient quantity of individual opinion scores that could allow us to effectively train the AIOs. That is the reason why we used a dataset involving another distortion (blur) never seen during the first learning step.

Starting from the JPEGResNet50, exploiting the ratings of each individual subject and a transfer learning approach, we derived 19 additional Deep CNNs, thus obtaining for each observer a model capable of predicting his/her choices in terms of the perceptual quality.

For each of the 19 subjects to be modeled, the transfer learning step was

performed as follows. We continued the training process of the JPEGResNet50 using, this time, as ground truth data, the ratings provided by that observer during the LIVE-MD-ph1 subjective experiment. In this way, the deep CNN modeling each observer directly takes advantage of the perceptual features previously learned during the training of the JPEGResNet50 on the synthetically annotated large-scale dataset. During this second training phase, the pre-learned features, i.e. those extracted by the JPEGResNet50, are further refined based on the ratings actually provided by each observer. This leads to a deep CNN, with different weights than those of the JPEGResNet50, that can extract a new set of features modeling of the observer quality perception.

In order not to overfit the small scale subjectively annotated training set, the deep CNN modeling each of the 19 observers was trained only for 10 epochs with a learning rate 100 times larger than the one used for the training process of the JPEGResNet50. All the 19 deep CNNs obtained at the end of this process possess the same architecture as the JPEGResNet50 (shown in Figure 3) but including different weights for each subject. These 19 networks represent our desired deep CNN-based AIOs.

In the prediction/inference phase (see Figure 1) each trained AIO predicts a probability distribution on the ACR scale just like the JPEGResNet50.

More formally, let us consider the deep CNN-based AIO mimicking the quality perception of the observer o: upon receiving, as an input, an image i, such an AIO provides as an output the following five probability values $p_{it}^o, \quad t = 1, 2, \ldots, 5$, that indicate with which probability the observer o would choose one of the five possible opinion scores of the ACR scale, when he/she would be asked to assess the quality of the image i. The predicted opinion score OS_i^o of the observer o for the image i can then be considered the one with the highest probability, i.e.

$$OS_i^o = \arg\max_t \left(p_{it}^o\right). \tag{6}$$

The MOS of each image i can therefore be estimated by the mean of the opinion scores predicted by the AIOs. We will refer to it as the MOS_{AI}.

As mentioned in the introduction, the modeling of individual observers has the advantage of allowing to estimate not only the MOS, but also, for instance, the expected distribution of users' opinion scores regarding the quality of a given image. We recall that such a distribution is especially important from a practical point of view. Given any image i, we are interested in determining the five probabilities $\alpha_i^t, \quad t = 1, 2, \ldots, 5$, i.e., the expected percentage of the end users that will rate the quality of i assigning t as the corresponding opinion score.

By exploiting the output of the AIOs modeling each of the 19 actual observers considered in this work, such percentages can be estimated as follows:

$$\alpha_i^t = \frac{1}{19} \sum_{o=1}^{19} p_{it}^o \quad t = 1, 2, \ldots, 5, \quad i = 1, 2, \ldots, I. \tag{7}$$

Please note that the proposed estimate of the distribution of the users' opinion scores is not just an empirical distribution derived from the 19 opinions scores predicted by the AIOs. Instead, it is derived from the probability values p_{it}^o. We have shown in our previous work [11] that the variance of the probability distribution derived from the probabilistic prediction of an AIO has the properties of a subject's inconsistency measure. By considering the subjects' inconsistency, we expect that the formula defined in Eq. (7) provides a robust estimate of the distribution of users' opinion scores.

We observe that, during the first learning step, the JPEGResNet50 could have been trained to model the third order polynomial function in Figure 2 instead of classifying compressed images into five classes. However, we preferred to discretize the quality scale directly during the training of the JPEGResNet50 in order to obtain a pre-trained network that already has a similar architecture as that of the AIOs and hence minimizes the amount of fine-turning actions required during the transfer learning process yielding the AIOs. In fact, if the JPEGResNet50 was trained for a regression task, in order to perform the transfer learning steps and get the desired AIOs, one would have needed to change the last layers of the architecture in order to switch from a regression task to a classification one. Therefore, some weights (those associated with the

newly added layers) should have been trained from scratch. We wanted to avoid such a situation, since, from our point of view, it makes the transfer learning step less efficient and potentially less effective, considering the fact that one would be learning from scratch on a limited size dataset.

4. Results

To assess the effectiveness of our approach, we conducted extensive computational experiments. These experiments and the related results are presented and commented in this section.

Please note that the JPEGResNet50 receives a 224×224 color image as an input; thus, the AIOs input is the same, since they share the same architecture. Therefore, in our experiments, to feed the JPEGResNet50 and the AIOs with a generic image, the central part of the image was first cropped to obtain a 400×400 image, which was then downscaled to a 224×224 one. By proceeding in this way, we implicitly made the assumptions that the perceptual quality of an image is predominantly determined by its central part and that by downscaling the image by a factor of 2 the visibility of artifacts is not reduced. We are aware that such a basic approach is not the most effective one and a more sophisticated way to pass an image as input to our trained models could yield better results. However, for simplicity's sake, we left this option for future research.

Finally, it is worth noting here that we have included the performance of our models also on the training set in this section. This is done in order to highlight and/or discuss potential cases of overfitting or underfitting of the training set.

4.1. Simulating the Process of a Subjective Test with the AIOs

In this section, we use the AIOs to simulate the process of five subjective tests and study the correlation between the opinion scores simulated by the AIOs and the opinion scores of the subjects that actually participated in the five experiments whose process is simulated.

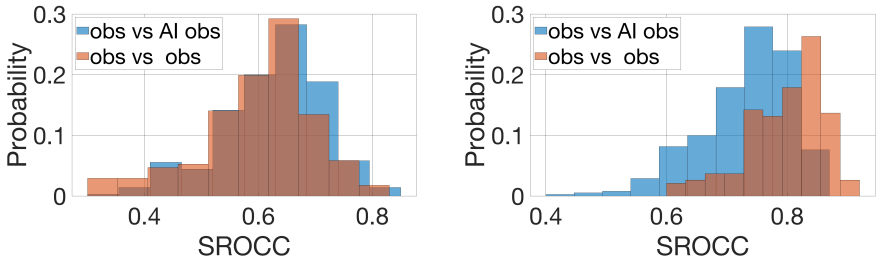
To show the effectiveness of the two-step learning approach discussed in this paper, the AIOs used to simulate the process of the five subjective tests

Datasets	**TLR**	**FW-TLR**	**Our approach**
LIVE-MD-ph1 (T)	[0.09, 0.70]	[0.00, 0.69]	[**0.34, 0.84**]
LIVE-IQA-r1-ph1	[-0.19, 0.27]	[-0.28 0.32]	[**0.42, 0.87**]
LIVE-IQA-r1-ph2	[-0.37 0.60]	[-0.33, 0.46]	[**0.79, 0.92**]
LIVE-MD-ph2	[0.04, 0.67]	[-0.05, 0.63]	[**0.21, 0.64**]
MICT	[-0.23 0.36]	[-0.28, 0.50]	[**0.30 0.76**]

Table 2: Effectiveness of our two-step learning approach. For each dataset, we considered all possible pairs made by an AIO and an actual subject. For each pair, we computed the Spearman Rank Order Correlation Coefficient (SROCC) between the ratings of the AIO and those of the actual subject. The table shows the minimum and the maximum SROCC obtained on each dataset for the different training approaches. TLR stands for single Transfer Learning step directly on the ResNet50 and FW-TLR stands for single Transfer Learning step on the ResNet50 after Freezing the Weights of 50% of the layers. (T) indicates the training set.

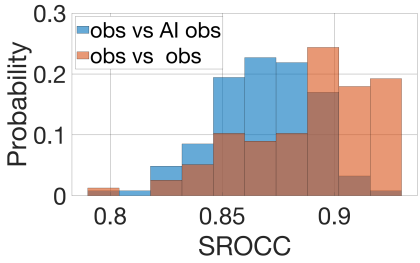
considered in this section were trained with three different approaches: i) by performing a single Transfer Learning step directly on the ResNet50 (TLR); ii) by freezing the weights of 50% of the layers of the ResNet50 and performing a single transfer learning step (FW-TLR); iii) by using our two-step learning approach described in the previous paragraphs.

The process of the five subjective experiments was then simulated by substituting the actual observers with the 19 AIOs trained by using each of the three approaches. We considered the following subjective experiments: the LIVE-MD-ph1 [19], the phase 1 and 2 of the first release of LIVE image quality assessment dataset, here abbreviated respectively as (LIVE-IQA-r1-ph1, LIVE-IQA-r1-ph2) [16], the MICT dataset [48] and finally the phase 2 of the LIVE Multiply Distorted Image Quality dataset (LIVE-MD-ph2) [19]. For each of these datasets, the opinion scores expressed by each actual observer that participated in the test were available. To simulate the process of these experiments, each image used in those experiments was given as an input to each one of the 19 AIOs derived from the three training approaches and the opinion scores of the AIOs were computed as indicated by Eq. (6).

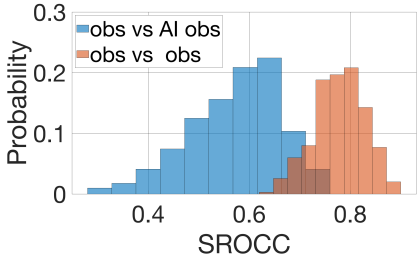


(a) LIVE-MD-ph1 (T)

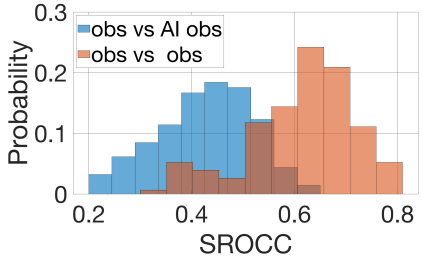
(b) LIVE-IQA-r1-ph1



(c) LIVE-IQA-r1-ph2



(d) MICT



(e) LIVE-MD-ph2

Figure 4: Comparing the distribution of correlation values observed between the ratings of pairs of actual observers to that of the correlation values observed between the ratings of pairs composed by an actual observer and an AIO. The higher the distributions overlap, the better. (T) stands for the training set.

To compare the AIOs to the actual observers, for each dataset, we considered all possible pairs made by an AIO and an actual subject. We then computed for each of these pairs the Spearman Rank Order Correlation Coefficient (SROCC) between the ratings of the AIO and those of the actual subject. Clearly, the higher this correlation is, the better it is.

The Table 2 presents, for each of the three training approaches, the range of values of the SROCC between the opinion scores of an actual subject and those of an AIO. For instance, it can be seen from the corresponding table that when training the AIOs by performing a single transfer learning step on the ResNet50 (TLR) and using these AIOs to simulate the process of the MICT experiment, the maximum correlation between the opinion scores of an AIO and an actual subject is 0.36, while the minimum is -0.23.

It can be seen from the obtained ranges of the correlation values reported in Table 2 that the proposed two-steps learning approach yields AIOs that better mimic the quality perception of the actual observers. In fact, the AIOs derived from the two-steps learning approach provided in all the cases opinion scores that correlates to those of the actual observers better than the ones obtained by the other investigated training approaches. Indeed, when adopting a single learning step, sometimes negative correlation coefficients are observed. This probably indicates, as mentioned before, that the networks did not have enough data in order to learn the main rules characterizing the quality assessment task.

The gap shown in Table 2 between the proposed two-steps learning approach and the other approaches was observed although we gave more training time to the approaches based on a single learning step. In fact, while in our two-steps learning approach 10 epochs were sufficient to derive each AIO from the JPEGResNet50, in the case of the two single learning step-based approaches, it took 30 epochs to start noticing that the network is overfitting the training set, i.e. that the accuracy on the validation set was no longer improving. This observation further supports the suitability of the JPEGResNet50 as compared to the ResNet50 as a stating point for transfer in the considered learning task.

Based on the results reported in Table 2, it seems clear that the AIOs derived

from the proposed two-steps learning approach outperform those obtained by a single learning step. Therefore, from now on, we will only consider the AIOs trained with our two-steps learning approach.

Figure 4 shows the histograms of the SROCC values between the ratings of a pair of actual observers and those of a pair made by an AIO and an actual observer. The SROCC values between the AIOs and the actual observers are quite similar to those obtained for any pair of the actual observers in the case of the LIVE-MD-ph1, LIVE-IQA-r1-ph1 and the LIVE-IQA-r1-ph2, since the histograms overlap well. This basically indicates that the choices of the AIOs are coherent with those of the actual observers, as expected.

For the MICT and LIVE-MD-ph2 datasets, less overlap was observed between the histograms, and lower SROCC values are observed in between the AIOs and actual observers (from 0.3 to 0.75 for the MICT dataset, from 0.2 to 0.65 for the LIVE-MD-ph2) than those obtained for the actual observers (from 0.6 to 0.9 for the MICT dataset, from 0.3 to 0.8 for the LIVE-MD-ph2). However, this result is not surprising. In fact, the LIVE-MD-ph2 contains images whose quality was impaired by adding noise artifacts. This type of artifacts was never seen by the AIOs during their training process. On the other hand, the MICT experiment, mentioned as the TOYAMA experiment in [49], involved a narrow range of JPEG quality degradation if compared to the much larger range of quality degradation considered in the LIVE-MD-ph1 experiment used to train the AIOs [49]. Therefore, the observers interpreted and used the quality scale in the MICT and LIVE-MD-ph1 experiments quite differently. This could explain why the AIOs trained on the LIVE-MD-ph1 did not succeed in simulating the rating process of the MICT experiment.

The results obtained on the LIVE-MD-ph2 and MICT datasets clearly confirm the fact that, as it happens with many other deep learning-based models, the trained AIOs should not be used beyond their design scope. When using our trained AIOs, the simulated individual ratings should be considered to have been gathered under the LIVE-MD-ph1 experimental setup, i.e., they are dependent on the influence factors involved in the LIVE-MD-ph1 subjective experiment,

DATASET	DISTORTION	BRISQUE	PIQUE	NIQE	PaQ-2-PiQ	PSNR	SSIM	$\mathbf{MOS_{res}}$	$\mathbf{MOS_{AI}}$
CSIQ [50]	JPEG	0.86	0.89	0.93	0.71	0.89	0.94	0.95	0.91
MICT [48]	JPEG	0.90	0.71	0.82	0.33	0.64	0.64	0.88	0.75
SDIVL [51]	JPEG	0.56	0.59	0.64	0.41	0.73	0.77	0.82	0.43
TID2013 [52]	JPEG	0.81	0.83	0.92	0.75	0.91	0.92	0.94	0.84
VCL-FER[53]	JPEG	0.76	0.66	0.80	0.50	0.57	0.82	0.93	0.76
LIVE-IQA-r1 [16]	JPEG	0.94	0.90	0.92	0.79	0.85	0.96	0.96	0.92
LIVE-IQA-r2 [54]	JPEG	0.96 (T)	0.83	0.79	0.70	0.95	0.92	0.91	0.86
MICT [48]	JP2K	0.87	0.79	0.84	0.35	0.84	0.84	0.46	0.69
LIVE-IQA-r1 [16]	JP2K	0.91	0.89	0.91	0.74	0.85	0.88	0.59	0.83
LIVE-MD-ph1 [19]	BLUR+ JPEG	0.12	0.75	0,66	0.70	0.37	0.36	0.25	0.83 (T)
LIVE-MD-ph2 [19]	BLUR+ NOISE	0.01	0.38	0.51	0.72	0.53	0.42	0.02	0.52

Table 3: PLCC value between the scores of each measure and the MOS separated per dataset and distortion type. It can be noticed that the proposed metrics, i.e. the $\mathbf{MOS_{res}}$ and the $\mathbf{MOS_{AI}}$, yield quite competitive PLCC values. (T) indicates that the dataset on which the metric is tested is a part of its training set.

since such subjective ratings are the ones used to train the AIOs.

4.2. Estimating the MOS

We evaluated the accuracy of the JPEGResNet50 as well as that of the AIOs in predicting the MOS of an image in this experiment.

The results are summarized in Table 3, 4 and 5. For each image, we computed the PSNR, SSIM [55], BRISQUE [56], PIQUE [57], NIQE [58] and PaQ-2-PiQ [46] scores. The BRISQUE [56], PIQUE [57], NIQE [58] and PaQ-2-PiQ [46] are no reference metrics, similarly as the models proposed in this work, while the PSNR and SSIM are full reference metrics, which are therefore expected to provide a higher accuracy in terms of MOS prediction. We also computed the MOS_{res}, i.e. the estimation of the MOS by the JPEGResNet50 as indicated by Eq. (5), and finally the MOS_{AI}, i.e. the mean of the predicted opinions by the 19 AIOs, upon receiving as an input the corresponding image.

Before calculating the Pearson Linear Correlation Coefficient (PLCC) and the Root Mean Square Error (RMSE) shown in the Table 3 and Table 5, we

DATASET	DISTORTION	BRISQUE	PIQUE	NIQE	PaQ-2-PiQ	PSNR	SSIM	$\mathbf{MOS_{res}}$	$\mathbf{MOS_{AI}}$
CSIQ	JPEG	0.85	0.85	0.90	0.71	0.90	0.93	0.93	0.87
MICT	JPEG	0.92	0.69	0.81	0.35	0.60	0.66	0.87	0.75
SDIVL	JPEG	0.54	0.59	0.54	0.44	0.76	0.82	0.71	0.29
TID2013	JPEG	0.83	0.79	0.90	0.73	0.93	0.90	0.92	0.83
VCL-FER	JPEG	0.79	0.68	0.82	0.52	0.58	0.82	0.94	0.74
LIVE-IQA-r1	JPEG	0.92	0.87	0.89	0.81	0.93	0.94	0.92	0.85
LIVE-IQA-r2	JPEG	0.97 (T)	0.84	0.84	0.80	0.94	0.95	0.90	0.86
MICT	JP2K	0.90	0.80	0.81	0.37	0.88	0.88	0.52	0.67
LIVE-IQA-r1	JP2K	0.92	0.89	0.91	0.74	0.92	0.91	0.69	0.78
LIVE-MD-ph1	BLUR+ JPEG	0.12	0.76	0.64	0.71	0.37	0.36	0.27	0.83 (T)
LIVE-MD-ph2	BLUR+ NOISE	0.16	0.37	0.48	0.72	0.52	0.37	0.01	0.53

Table 4: SROCC value between the scores of each measures and the MOS separated per dataset and distortion type. It can be noticed that the proposed metrics, i.e. the $\mathbf{MOS_{res}}$ and the $\mathbf{MOS_{AI}}$, yield quite competitive SROCC values. (T) indicates that the dataset on which the metric is tested is a part of its training set.

have normalized all the metrics from their original scale to the MOS scale by performing a least square fitting using the following logistic function:

$$\widehat{MOS} = \beta_1 \left(0.5 + \frac{1}{1 + \exp \beta_2 (VQM - \beta_3)} \right) + \beta_4 \cdot VQM + \beta_5 \tag{8}$$

The PLCC, SROCC and RMSE values presented respectively in Table 3, Table 4 and Table 5 show that the proposed models are very competitive with respect to all the other metrics considered in this experiment in terms of the MOS prediction. The JPEGResNet50 is particularly accurate when estimating the quality of the JPEG compressed images. For instance, on the VCL-FER dataset, the MOS_{res} provided by the JPEGResNet50 yielded a PLCC of 0.93 and a SROCC of 0.94, while the PIQUE only achieved 0.66 and 0.68, respectively. In this case even the PSNR and SSIM yielded lower accuracy in comparison to the output provided by the JPEGResNet50. This is really interesting, if one takes into account the fact that the JPEGResNet50 has been trained using only synthetically generated data. We hypothesize that such accuracy is because the weights of the JPEGResNet50 are learned in such a way that the proba-

DATASET	DISTORTION	BRISQUE	PIQUE	NIQE	PaQ-2-PiQ	PSNR	SSIM	$\mathbf{MOS_{res}}$	$\mathbf{MOS_{AI}}$
CSIQ	JPEG	0.63	0.56	0.48	0.88	0.56	0.43	0.37	0.51
MICT	JPEG	0.51	0.82	0.67	1.10	0.89	0.90	0.55	0.76
SDIVL	JPEG	0.77	0.75	0.72	0.85	0.64	0.60	0.54	0.85
TID2013	JPEG	0.40	0.48	0.34	0.57	0.28	0.26	0.24	0.38
VCL-FER	JPEG	0.56	0.64	0.51	0.74	0.70	0.49	0.31	0.56
LIVE-IQA-r1	JPEG	0.33	0.40	0.35	0.56	0.49	0.25	0.26	0.35
LIVE-IQA-r2	JPEG	0.26 (T)	0.55	0.88	0.85	0.31	0.38	0.42	0.50
MICT	JP2K	0.60	0.74	0.65	1.12	0.64	0.65	1.06	0.87
LIVE-IQA-r1	JP2K	0.35	0.39	0.36	0.57	0.45	0.41	0.69	0.47
LIVE-MD-ph1	BLUR+ JPEG	0.49	0.34	0.47	0.35	0.45	0.46	0.47	0.27 (T)
LIVE-MD-ph2	BLUR+ NOISE	0.54	0.50	0.50	0.38	0.46	0.49	0.54	0.46

Table 5: RMSE value between the scores of each measure and the MOS separated per dataset and distortion type. It can be noticed that the proposed metrics, i.e. the $\mathbf{MOS_{res}}$ and the $\mathbf{MOS_{AI}}$, yield quite competitive RMSE values. (T) indicates that the dataset on which the metric is tested is a part of its training set.

bility values $p_i^t(\beta)$ in Eq. (5) take into account the potential imprecision that affects the labels in the synthetically generated dataset. Specific experiments are however needed to verify the validity of such a hypothesis.

The accuracy of the JPEGResNet50 is however strongly dependent on the type of distortion that affects the perceptual quality of the processed image. In fact, the JPEGResNet50 is not able to accurately process images whose quality is impaired by artifacts jointly caused by the blur and JPEG compression as well as the blur and noise. All the other considered no reference metrics performed better than our JPEGResNet50 on JPEG2000 compressed images. This was somehow expected, since the compression process used for generating the synthetic data deployed for the training of the JPEGResNet50 was only related to the JPEG quality parameter. This latter observation highlights the necessity to develop, in future work, approaches for artificially generating large-scale datasets suitable for training deep neural network models that can be deployed for a wider range of applications.

When looking at the prediction of the MOS through the mean of the opinion

	BRISQUE	**PIQUE**	**NIQE**	**PaQ-2-PiQ**	**PSNR**	**SSIM**	**$\mathbf{MOS_{res}}$**	**$\mathbf{MOS_{AI}}$**	**Total**
BRISQUE	—	4	1	7	4	2	3	4	25
PIQUE	1	—	0	8	2	0	3	3	17
NIQE	3	4	—	8	4	1	3	5	28
PaQ-2-PiQ	1	1	1	—	1	1	2	1	8
PSNR	3	3	0	8	—	1	4	4	23
SSIM	5	5	2	8	3	—	3	7	33
$\mathbf{MOS_{res}}$	5	6	3	6	5	2	—	7	34
$\mathbf{MOS_{AI}}$	2	1	0	7	2	0	3	—	15

Table 6: Results of the statistical test performed for comparing the PLCC values provided by the different metrics on all the datasets. The number in the i-th row and j-th column of the table represents the number of datasets on which the i-th metric performed significantly better than the j-th one with 95% of confidence. For instance, the $\mathbf{MOS_{res}}$ performed better than the **BRISQUE** with statistical significance on 5 datasets.

scores of the AIOs, i.e. the MOS_{AI}, one can observe that the MOS_{AI} predicts the quality of the JPEG compressed images with a lower accuracy than the JPEGResNet50. However, it does perform better when it comes to the assessment of the visual quality of the images distorted by the blur and noise artifacts.

From the results in Table 3, Table 4 and Table 5, one can notice that the PaQ-2-PiQ offered in general lower performance than all the other metrics on JPEG and JPEG2000 compressed images. On the other hand, it was more accurate than many other metrics when predicting the quality of images affected by more than a single type of distortion. For instance, it offers the greatest performance on the LIVE-MD-ph2 dataset that includes images distorted by jointly adding blur and noise. It also shows a good performance on the LIVE-MD-ph1 dataset in which the quality of the images was impaired by applying some blur and performing JPEG compression. This higher performance of the PaQ-2-PiQ on images with more than one distortion can be probably explained by the fact that this type of images is closer to the ones including "real" distortions used for the training of that metric. In fact, user-generated images sometimes contain some blur due to the non-stability of the camera, then they are compressed, thus introducing other artifacts.

The competitiveness of our proposal was further investigated by conducting statistical tests. More precisely, we compared from the statistical significance point of view each pair of metrics in terms of PLCC. In this experiment, for fairness, we excluded the LIVE-MD-ph1 and LIVE-IQA-r2 datasets on which our models and the BRISQUE were trained, respectively. The results are summarized in Table 6. As one can notice, none of the metrics was significantly more accurate than all the others on all the datasets. Indeed, the results show that the JPEGResNet50 can predict the quality of the JPEG compressed images with as a high accuracy as a full reference metric would offer. In fact, while SSIM was significantly more accurate in terms of the PLCC in 33 comparisons, the MOS_{res} was in 34 cases. On the other hand, the MOS_{AI} demonstrated a lower performance. However, in comparison to many metrics, it shows a greater robustness in predicting the quality of the images affected by multiple distor-

tions.

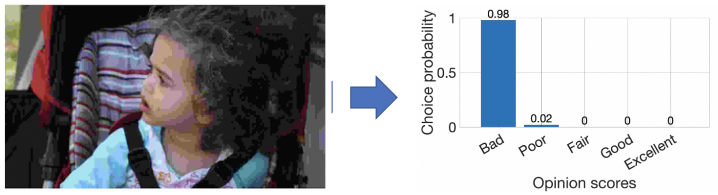
It is fundamental to notice that beyond the high competitiveness of the proposed metrics, i.e. the MOS_{res} and the MOS_{AI}, they offer in parallel a considerable advantage over the other metrics in terms of the MOS prediction. In fact, both the JPEGResNet50 as well as the model of each single AIO return a discrete probability distribution that can be used to estimate not only the MOS but also the distribution of the opinions of the end users on the quality scale. The results related to this advantage of our approach are presented in the following section.

4.3. Estimating the Distribution of Users' Opinion Scores

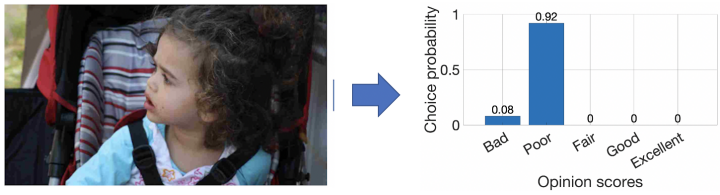
For each image, the computation of the five probability values that constitute the distribution of users' opinion scores on the ACR scale was performed according to the Eq. (7) after passing that image as an input to the 19 trained AIOs.

We start by an example to better illustrate this experiment. For this purpose, an image whose quality is progressively degraded by applying JPEG compression is considered. Starting from the original image, we have thus generated five compressed images, which have been given as an input to the 19 AIOs. The distribution of opinions was derived based on the output of each AIO. Figure 5 illustrates the results. One can notice that the support of the predicted distribution moves progressively to the right as the JPEG quality parameter increases. Furthermore, the predicted distribution shows a greater variance when the JPEG quality parameter is 35, while for small values of this parameter (5 and 15) the obtained distribution is almost totally concentrated on a single opinion score. This is a very interesting observation because it suggests that the AIOs can mimic the well-known ability of human subjects to evaluate low-quality content more consistently.

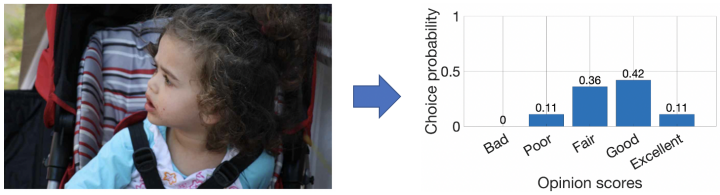
We then generalized this preliminary experiment by predicting the distribution of users' opinion scores for all the images included in the five annotated datasets. Figure 6 shows the estimated distribution for each image as a function



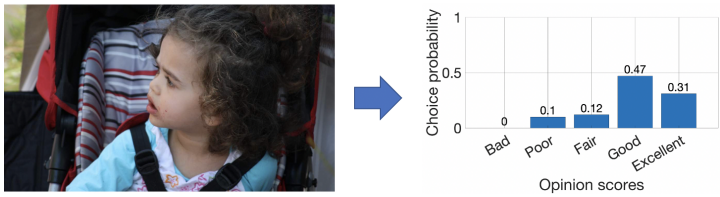
(a) JPEG Quality Parameter equal to 5



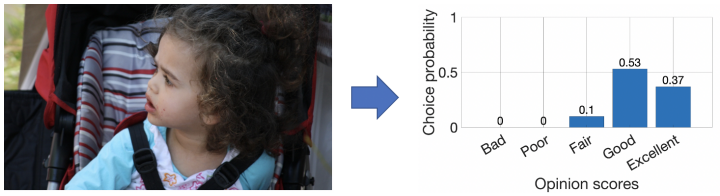
(b) JPEG Quality Parameter equal to 15



(c) JPEG Quality Parameter equal to 35

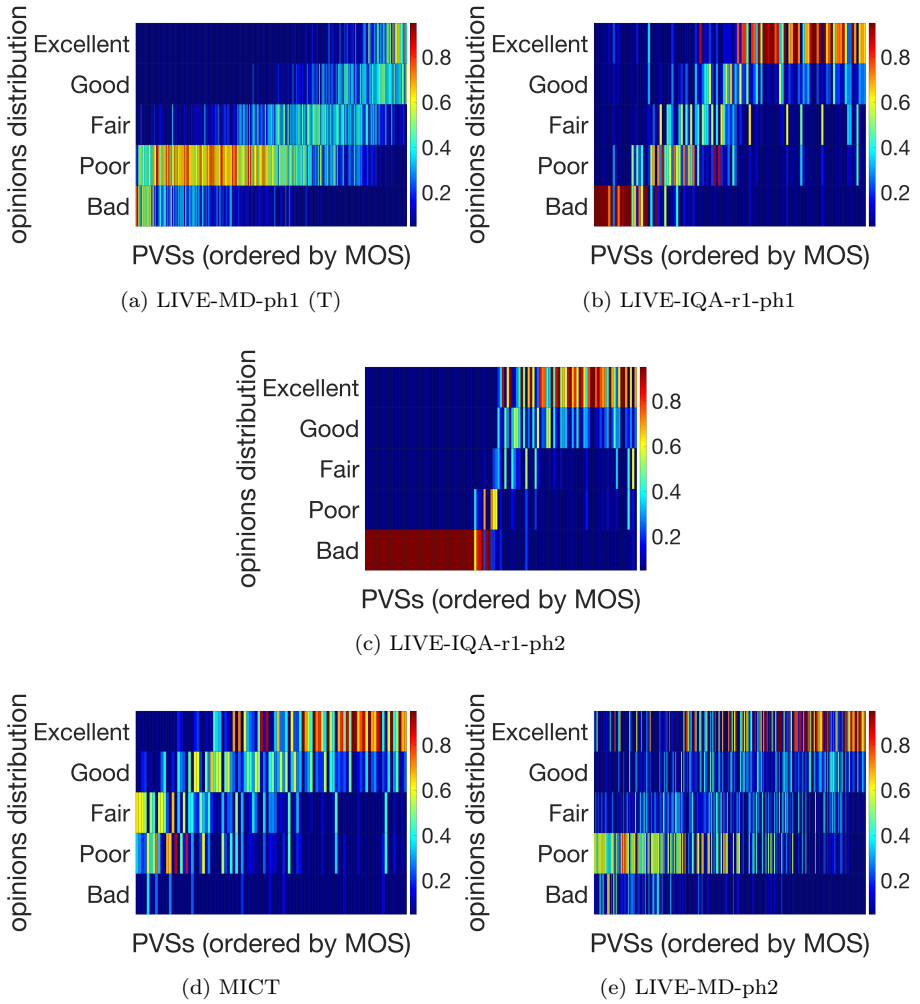


(d) JPEG Quality Parameter equal to 65



(e) JPEG Quality Parameter equal to 95

Figure 5: Showcasing the usage of the AIOs in practice. The figure shows the distribution of the user opinions as predicted by the AIOs. The quality of the image given as an input is progressively degraded by applying JPEG compression.



(a) LIVE-MD-ph1 (T)

(b) LIVE-IQA-r1-ph1

(c) LIVE-IQA-r1-ph2

(d) MICT

(e) LIVE-MD-ph2

Figure 6: The predicted distribution of the users' opinion scores for each image as a function of its MOS. It is worth noting here that the mode of the predicted distribution tends to increase as the MOS increases, as clearly expected. Furthermore, the distribution is concentrated around the mode in most of the cases. (T) stands for the training set.

of its MOS. It can be noticed, as expected, that as the MOS increases, higher probability values are progressively concentrated on the higher opinion scores. In fact, a positive correlation between the MOS and the mode of the distribution of the user opinions can be observed. It is also important to notice that, as it often happens in practice during subjective experiments, the support of the predicted distribution is in almost all the cases concentrated on consecutive opinion scores around the mode of the predicted distribution. This highlights the fact that the AIOs, during the training process, have been able to capture the ordinal nature of the ACR scale. This, however, was not trivial since none of the constraints of the optimization problem guiding the training process of the AIOs explicitly imposes that.

To better appreciate the effectiveness of our estimation of the distribution of users' opinion scores, we conducted statistical tests aiming at determining, for each image, whether the predicted distribution of the users' opinion scores is different from the empirical fractions observed during the subjective experiment, with a statistical significance. We relied on the Kolmogorov–Smirnov test in this case. The tests were performed with 95% of confidence. Table 7 reports, for each considered dataset, the percentage of the images for which the predicted distribution can be considered not different than the one observed during the subjective test with statistical significance. In all the cases such a percentage is greater than 50%. These results are really promising and show a high potential of the proposed approach for future research in QoE measurement scenarios.

4.4. Time Required to Simulate the Process of a Subjective Test with the AIOs

In this section we report the computational time required, on average, by an AIO to perform its evaluation of the quality of a given picture. This allows to figure out an estimate of the time that would by required by the trained AIOs to simulate the process of a subjective experiment.

The experiments were done on a computer having an Intel(R) Core(TM) i9-10900X CPU with a clock speed of 3.7 GHz and 64 GB of RAM. The used computer is also equipped with a GPU with an NVIDIA GeForce RTX 3090

Dataset	Percentage of images
LIVE-MD-ph1 (T)	100%
LIVE-IQA-ph1	66%
LIVE-IQA-ph2	90%
MICT	50%
LIVE-MD-ph2	68%

Table 7: Percentage of the images for which the predicted users' distribution of opinions is not statistically different from the empirically observed one. (T) stands for the training set.

GPU with 24 GB of RAM.

We selected at random 100 images from the ImageNet dataset. Then, we recorded the computational time needed by the PIQUE, NIQE, BRISQUE, PaQ-2-PiQ and the 19 AIOs to evaluate the quality of these images. For the PIQUE, NIQE, BRISQUE and PaQ-2-PiQ, the recorded computational time was then divided by 100 to get the average computational time needed to evaluate the quality of a single picture. The time for a single AIO is obtained by further dividing by 19, i.e. the number of the AIOs. The experiment was repeated 10 times and the range of the obtained values is reported. This is done in order to account for the potential variability of the computational time due to the need of the operating system to sporadically run some of its internal tasks/processes.

The average time required by a single AIO to predict the quality of a picture is ranging from 16 to 18 milliseconds (ms). This range of the computational times is quite similar to those of the BRISQUE, NIQE, PIQUE and PaQ-2-PiQ for which the observed ranges are respectively: 13-16 ms, 20-22 ms, 22-24 ms and 21-22 ms. It is important to note that the BRISQUE, NIQE and PIQUE are not deep neural network based measures. As such, their computation does not exploit the characteristics of the GPU.

Based on the reported ranges, we observe that, relying on the 19 trained AIOs, no more than 35 seconds (19 AIOs * 18 ms * 100 images) are required to simulate the process of a subjective experiment involving 100 images. This time

frame is negligible compared to what it would take to setup a real subjective experiment involving 19 subjects and 100 stimuli, conduct it, and screen the obtained result. However, it is important to note that, although AIOs offer greater efficiency, real subjects are expected to provide a more accurate evaluation of the quality in general.

5. Conclusion

In this work we focused on the issue of modeling the quality perception of an individual observer with a deep CNN. The purpose of our study was to create models being able to replicate the choices of a real observer with a high accuracy. To cope with the difficulties related to the training of deep neural networks on small-scale annotated datasets, we propose to synthetically annotate a large-scale dataset by mapping progressive levels of the JPEG compression to the five-point ACR quality scale. Using this dataset, we trained our JPEGResNet50, a deep neural network with up to 52 hidden convolutional layers. The results demonstrate that the JPEGResNet50 can be readily used to accurately evaluate the quality of the JPEG compressed images. To obtain the desired deep CNN-based models of single observers, we relied on a transfer learning approach. The model that mimics the quality perception of each of the 19 observers considered in our study is obtained by continuing the training of the JPEGResNet50 on a dataset annotated by these observers. During this second learning phase, the perceptual features already learned by the JPEGResNet50 are further updated/fine-tuned based on the opinion scores expressed by the observer during the subjective test. This allows to obtain, for each observer, a set of features that can accurately model his/her quality perception. A total of 19 deep CNNs (one for each observer) have therefore been trained and released.

The experiments performed on several datasets highlighted the accuracy of these models in terms of the MOS prediction, while promising results were obtained when comparing the proposed models to the actual observers and estimating the distribution of the user opinion scores on the quality of a given

image.

We see several directions in which this work can be improved in future contributions. The deep CNN-based AIOs trained in this paper were designed for still image applications. A potential extension of the approach would be towards video content. Also, by conducting the first learning step only with JPEG compressed images, the ability of the trained AIOs to accurately assess a quality degradation caused by other types of artifacts could be questionable. Therefore, in the future, it would be interesting to start from a synthetically annotated dataset involving several different types of distortions. Another question of high interest for the design of more accurate AIOs is how to collect enough reliable subjective raw opinion scores in order to get rid of synthetic labels or transfer learning approaches. This can be related to our work published in [59]. In fact, new recommendations for the design of subjective tests aiming at the training of AIOs needs to be investigated. It is important to research the aspects of the human perception of quality that a deep CNN can really mimic. In other words, it would be interesting to understand whether a deep CNN, trained to predict the opinion scores of a human subject, attempts to simulate the mental process that guides human choices or implements a totally different approach that however yields the same prediction. A starting point in this direction could be a comparative analysis of the sensitivity of a human subject and that of his/her AIO to specific modifications to the input signal. Finally, we are aware of the fact that AIOs with an improved performance might be obtained by testing several other existing neural network architectures or by designing another one, from scratch, that is tailored to the considered learning task. However, this cannot be done without a very high computational training effort. For this reason, we see it as a point for a future contribution.

6. Acknowledgment

This work presented in this paper has been supported in part by PIC4SeR (`http://pic4ser.polito.it`). Some of the computational resources were pro-

vided by HPC@POLITO (`http://www.hpc.polito.it`).

References

[1] Cisco, Annual internet report: Growth in internet users (2018–2023) (2018).

[2] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, T. Vercauteren, Interactive medical image segmentation using deep learning with image-specific fine tuning, IEEE Transactions on Medical Imaging 37 (7) (2018) 1562–1573.

[3] B. Demir, S. Ertürk, Improving SVM classification accuracy using a hierarchical approach for hyperspectral images, in: 2009 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 2849–2852.

[4] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising, IEEE Transactions on Image Processing 26 (7) (2017) 3142–3155.

[5] E. Ong, S. S. Husain, M. Bober-Irizar, M. Bober, Deep architectures and ensembles for semantic video classification, IEEE Transactions on Circuits and Systems for Video Technology 29 (12) (2019) 3568–3582.

[6] C. G. Bampis, Z. Li, I. Katsavounidis, A. C. Bovik, Recurrent and dynamic models for predicting streaming video quality of experience, IEEE Transactions on Image Processing 27 (7) (2018) 3316–3331.

[7] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on Image Processing 21 (12) (2012) 4695–4708.

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

[9] F.-Z. Ou, Y.-G. Wang, G. Zhu, A novel blind image quality assessment method based on refined natural scene statistics, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 1004–1008. doi:10.1109/ICIP.2019.8803047.

[10] J. Korhonen, Assessing personally perceived image quality via image features and collaborative filtering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Long Beach, CA, USA, 2019, pp. 8169–8177.

[11] L. F. Tiotsop, T. Mizdos, M. Barkowsky, P. Pocta, A. Servetti, E. Masala, Mimicking individual media quality perception with neural network based artificial observers, ACM Transactions on Multimedia Computing, Communications, and Applications 18 (1) (2022).

[12] L. Janowski, M. Pinson, The accuracy of subjects in a quality experiment: A theoretical subject model, IEEE Transactions on Multimedia 17 (12) (2015) 2210–2224.

[13] Z. Li, C. G. Bampis, Recover subjective quality scores from noisy measurements, in: 2017 Data Compression Conference (DCC), 2017, pp. 52–61. doi:10.1109/DCC.2017.26.

[14] J. Li, S. Ling, J. Wang, P. Le Callet, A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3339–3347.

[15] S. Bianco, L. Celona, P. Napoletano, R. Schettini, On the use of deep learning for blind image quality assessment, Signal, Image and Video Processing 12 (2) (2018) 355–362.

[16] H. R. Sheikh, Z. Wang, L. Cormack, A. C. Bovik, Live image quality assessment database, http://live.ece.utexas.edu/research/quality (2005).

[17] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[19] D. Jayaraman, A. Mittal, A. K. Moorthy, A. C. Bovik, Objective quality assessment of multiply distorted images, in: 2012 Conference record of the forty sixth asilomar conference on signals, systems and computers (ASILOMAR), IEEE, 2012, pp. 1693–1697.

[20] K. Mitra, A. Zaslavsky, C. Ahlund, Context-aware QoE modelling, measurement and prediction in mobile computing systems, IEEE Transactions on Mobile Computing 14 (2015) 920–936. doi:10.1109/TMC.2013.155.

[21] T. Hoßfeld, P. E. Heegaard, M. Varela, S. Möller, QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS, Quality and User Experience 1 (1) (Sep 2016). doi:10.1007/s41233-016-0002-1.

[22] R. C. Streijl, S. Winkler, D. S. Hands, Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives, Multimedia Systems 22 (2) (2016) 213–227. doi:10.1007/s00530-014-0446-1.

[23] H. Zeng, L. Zhang, A. C. Bovik, A probabilistic quality representation approach to deep blind image quality prediction (2017). arXiv:arXiv:1708.08190v2.

[24] L. F. Tiotsop, E. Masala, A. Aldahdooh, G. Van Wallendael, M. Barkowsky, Computing quality-of-experience ranges for video quality estimation, in: Eleventh International Conference on Quality of Multimedia Experience (QoMEX), 2019, pp. 1–3. doi:10.1109/QoMEX.2019.8743303.

[25] M. Seufert, Fundamental advantages of considering quality of experience distributions over mean opinion scores, in: 2019 Eleventh International

Conference on Quality of Multimedia Experience (QoMEX), 2019, pp. 1–6. doi:10.1109/QoMEX.2019.8743296.

[26] L. Janowski, Z. Papir, Modeling subjective tests of quality of experience with a generalized linear model, in: International Workshop on Quality of Multimedia Experience (QoMEX), 2009, pp. 35–40. doi:10.1109/QOMEX.2009.5246979.

[27] D. Varga, D. Saupe, T. Szirányi, Deeprn: A content preserving deep architecture for blind image quality assessment, in: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, San Diego, CA, USA, 2018, pp. 1–6.

[28] ITU-T, Vocabulary for performance, quality of service and quality of experience (ITU-T Rec. P.10/G.100), International Telecommunication Union (ITU), Places des Nations 20, CH-1211 Geneva (2017).

[29] Qualinet white paper on definitions of Quality of Experience (2012), European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), P. Le Callet, S. Möller and A. Perkis, eds., Lausanne, Switzerland, Version 1.2 (2013).

[30] J. A. Redi, Y. Zhu, H. de Ridder, I. Heynderickx, How Passive Image Viewers Became Active Multimedia Users, Springer International Publishing, Cham, 2015, pp. 31–72.

[31] A. Bouzerdoum, A. Havstad, A. Beghdadi, Image quality assessment using a neural network approach, in: Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, 2004., 2004, pp. 330–333.

[32] P. Gastaldo, G. Parodi, J. Redi, R. Zunino, No-reference quality assessment of JPEG images by using CBP neural networks, in: International Conference on Artificial Neural Networks, Springer, 2007, pp. 564–572.

[33] P. Gastaldo, R. Zunino, I. Heynderickx, E. Vicario, Objective quality assessment of displayed images by using neural networks, Signal processing: Image communication 20 (7) (2005) 643–661.

[34] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-reference image quality assessment, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1733–1740.

[35] Y. Ding, R. Deng, X. Xie, X. Xu, Y. Zhao, X. Chen, A. S. Krylov, No-reference stereoscopic image quality assessment using convolutional neural network for adaptive feature extraction, IEEE Access 6 (2018) 37595–37603.

[36] X. Wang, X. Liang, B. Yang, F. W. Li, No-reference synthetic image quality assessment with convolutional neural network and local image saliency, Computational Visual Media 5 (2) (2019) 193–208.

[37] T. Lu, A. Dooms, A deep transfer learning approach to document image quality assessment, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1372–1377.

[38] W. Zhang, K. Ma, J. Yan, D. Deng, Z. Wang, Blind image quality assessment using a deep bilinear convolutional neural network, IEEE Transactions on Circuits and Systems for Video Technology 30 (1) (2020) 36–47.

[39] J. Kim, S. Lee, Fully deep blind image quality predictor, IEEE Journal of selected topics in signal processing 11 (1) (2016) 206–220.

[40] W. Liu, Z. Duanmu, Z. Wang, Blind quality assessment of compressed videos using deep neural networks., in: ACM Multimedia, 2018, pp. 546–554.

[41] L. Krasula, Y. Baveye, P. Le Callet, Training objective image and video quality estimators using multiple databases, IEEE Transactions on Multimedia 22 (4) (2020) 961–969.

[42] T. Mizdos, M. Barkowsky, M. Uhrina, P. Počta, Linking bitstream information to QoE: A study on still images using HEVC intra coding, Advances in Electrical and Electronic Engineering 17 (12 2019). doi:10.15598/aeee.v17i4.3625.

[43] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of Big Data 6 (1) (2019) 60.

[44] M. H. Pinson, S. Wolf, An objective method for combining multiple subjective data sets, in: Visual Communications and Image Processing 2003, Vol. 5150, International Society for Optics and Photonics, 2003, pp. 583–592.

[45] S. Pezzulli, M. G. Martini, N. Barman, Estimation of quality scores from subjective tests: beyond subjects' MOS, IEEE Transactions on Multimedia 23 (2020) 2505–2519.

[46] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, A. Bovik, 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), 2020, pp. 3572–3582.

[47] B. Frénay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE transactions on neural networks and learning systems 25 (5) (2013) 845–869.

[48] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, Which semi-local visual masking model for wavelet based image quality metric?, in: 2008 15th IEEE International Conference on Image Processing, 2008, pp. 1180–1183.

[49] S. Tourancheau, F. Autrusseau, Z. M. P. Sazzad, Y. Horita, Impact of subjective dataset on the performance of image quality metrics, in: 2008 15th IEEE International Conference on Image Processing, 2008, pp. 365–368.

[50] E. C. Larson, D. M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, Journal of Electronic Imaging 19 (1) (March 2010).

[51] S. Corchs, F. Gasparini, R. Schettini, No reference image quality classification for JPEG-distorted images, Digital Signal Processing 30 (2014) 86 – 100. doi:https://doi.org/10.1016/j.dsp.2014.04.003.

[52] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C. . J. Kuo, Color image database TID2013: Peculiarities and preliminary results, in: European Workshop on Visual Information Processing (EUVIP), 2013, pp. 106–111.

[53] A. Zarić, N. Tatalović, N. Brajković, H. Hlevnjak, M. Lončarić, E. Dumić, S. Grgić, VCL@ FER image quality assessment database, AUTOMATIKA 53 (4) (2012) 344–354.

[54] H. R. Sheikh, M. F. Sabir, A. C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Transactions on Image Processing 15 (11) (2006) 3440–3451.

[55] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612.

[56] A. Mittal, A. K. Moorthy, A. C. Bovik, No-reference image quality assessment in the spatial domain, IEEE Transactions on image processing 21 (12) (2012) 4695–4708.

[57] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, S. S. Medasani, Blind image quality evaluation using perception based features, in: 2015 Twenty First National Conference on Communications (NCC), 2015, pp. 1–6.

[58] A. Mittal, R. Soundararajan, A. C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal processing letters 20 (3) (2012) 209–212.

[59] L. F. Tiotsop, F. Agboma, G. Van Wallendael, A. Aldahdooh, S. Bosse, L. Janowski, M. Barkowsky, E. Masala, On the link between subjective score prediction and disagreement of video quality metrics, IEEE Access (2021).