Research Article

Uncertainty-driven Sampling for Efficient Pairwise Comparison Subjective Assessment

Shima Mohammadi[1], João Ascenso[1]

1. Independent researcher

Assessing image quality is crucial in image processing tasks such as compression, super-resolution, and denoising. While subjective assessments involving human evaluators provide the most accurate quality scores, they are impractical for large-scale or continuous evaluations due to their high cost and time requirements. Pairwise comparison subjective assessment tests, which rank image pairs instead of assigning scores, offer more reliability and accuracy but require numerous comparisons, leading to high costs. Although objective quality metrics are more efficient, they lack the precision of subjective tests, which are essential for benchmarking and training learning-based quality metrics. This paper proposes an uncertainty-based sampling method to optimize the pairwise comparison subjective assessment process. By utilizing deep learning models to estimate human preferences and identify pairs that need human labeling, the approach reduces the number of required comparisons while maintaining high accuracy. The key contributions include modeling uncertainty for accurate preference predictions and for pairwise sampling. The experimental results demonstrate superior performance of the proposed approach compared to traditional active sampling methods. Software is publicly available at: shimamohammadi/LBPS-EIC

Corresponding authors: Shima Mohammadi, shima.mohammadi@lx.it.pt; João Ascenso, joao.ascenso@lx.it.pt

I. Introduction

Image quality evaluation plays a vital role by quantifying quality degradations according to human perception, which is especially required in many visual processing tasks, such as compression, super-

resolution, denoising and others. There are two primary methods for evaluating image quality: subjective assessment tests and objective quality metrics. Subjective assessment, considered the most accurate, involves gathering quality scores from human evaluators using a reliable methodology and graphical interface. However, despite its precision, this approach can be impractical in many scenarios, e.g. during the optimization or training of image compression algorithms or for continuous quality monitoring. Consequently, objective metrics, which automate the quality assessment process, are frequently used.

Depending on number of stimuli that are present to the subjects, several subjective assessment methodologies can be used to obtain quality scores, including single stimulus and double stimulus approaches. Single and double stimulus methodologies like Absolute Category Rating (ACR) and Double Stimulus Impairment Scale (DSIS) are even standardized[1][2]. These methodologies require a panel of human subjects to score each visual stimulus on a predefined category scale. The average of these scores is then reported as mean opinion scores (MOS) or differential MOS (DMOS) relative to a reference. However, despite the widespread use, it lack sensitivity when the quality differences between stimuli are small and also suffer from different types of biases, past experiences, or expectations. For example, the interpretation of the category scales can vary among subjects, the perception of "Excellent" quality might differ among subjects in different contexts, or, they may change their ratings during the test, leading to inconsistent scores. These limitations can result in less reliable test outcomes.

A promising alternative to address the aforementioned issues is to simplify the subjective evaluation process by replacing category rating with simple ranking. Ranking approaches compare image pairs or triplets to determine which one is of higher quality. Typically, it is much easier (and leads to an increase sensitivity) for human subjects to express a preference between two images rather than attribute a score for each. Pairwise comparison (PC) is an example of a ranking methodology, where subjects are asked to select which image of a pair they find more preferable. Typically, the preferences are obtained from all subjects and for all possible pairs and quality scores can be inferred using models such as Thurstone[3] or Bradley-Terry[4]. This methodology leads to more accurate and reliable results than category rating methodologies[5][6][7]. In such case, subjects in pairwise comparison tests are better able to detect small quality differences and are less affected by changes in viewing conditions. This approach is becoming rather popular nowadays for assessing high-quality to nearly visually lossless images and is the primary focus of the JPEG AIC ad-hoc group which is dedicated to standardizing both subjective and objective quality metrics for image compression.

Although the pairwise comparison test is rather accurate, it requires a large number of pairs, making it costly in most cases, especially because to assess the quality of n distorted images against a reference image, $n(n-1)/2$ pairs are needed. Thus, number of pairs grows quadratically $\mathcal{O}(n^2)$, with the number of distorted images. Consequently, when using many reference images and their distorted versions, the pairwise comparison test becomes impractical or very expensive. One solution to this problem is to sample only a subset of pairs for the pairwise comparison test. A straightforward approach is to randomly select a set of pairs for evaluation, but, several more elegant pairwise sampling methods were proposed already. The most promising approach is to select pairs based on past decisions made by subjects and according to a measure of the impact of the pair on the final scores. However, this requires running the sampling procedure after each comparison (trial) for every pair, making it impractical in many scenarios. Moreover, this sampling relies on past decisions and can be compromised by malicious subjects, leading to inefficient sampling. Therefore, it is preferable to sample a subset of pairs beforehand, independent of the subjective test results. Therefore, the key idea in this paper is significantly different from previous state-of-the-art methods: some of the ranking decisions can be automatically predicted using a deep learning framework, while human subjects will make the remaining decisions.

In this context, the objective of this paper is to propose uncertainty-based sampling to reduce the subjective test (following a pairwise comparison procedure) duration without compromising the accuracy of the results. The proposed uncertainty models will estimate human preferences between image pairs and identify which pairs require active human labeling. Unlike previous work, this approach will be implemented using a deep learning model for the first time. The novel concept is that pairs with unreliable preference estimations will be deferred to human subjects for accurate preferences, while the estimated preferences will be directly used for the remaining pairs.

Therefore, the main contributions of this paper are summarized below.

- **Modeling uncertainty for accurate preference prediction:** Propose a data (or aleatoric) uncertainty estimator to obtain an accurate prediction of the probability of preference within a pairwise comparison methodology. By leveraging deep learning models, the goal is to obtain reliable estimates of the underlying quality of both images in a pair, thereby modeling human subjects' preferences. The final estimates are used as replacement of the subjective scores for some of the pairs reducing the length of the subjective test.

- **Modeling uncertainty for accurate pairwise sampling:** Propose a model (or epistemic) uncertainty procedure to identify instances (pairs) where the deep learning model cannot reliably estimate the

probability of preference. In such cases, the pair of images must be subjectively evaluated. This contribution allows to select the pairs of images for which the deep learning model's predictions are uncertain and request, for these cases, the help from human subjects.

The rest of the paper is organized as follows: Related work is described in Section II. The uncertainty models are described in Section III. The quality estimation and pairwise sampling procedure is detailed in Section IV. In Section V, the performance evaluation including the training procedure and experimental results are described. Finally, the conclusion is provided in Section VI.

II. Related Work

This section provide an overview of relevant related work considering the objectives of this paper, first, pairwise sampling methods are reviewed followed by uncertainty estimation methods in objective quality metrics.

A. Pairwise Sampling

To address the challenging issue of pairwise sampling on subjective tests, several techniques have been proposed in the literature. These approaches can be classified into several types: 1) random selection: involves randomly selecting a subset of pairs[8] for human evaluation, 2) sorting-based selection: treats pairwise comparisons as a sorting problem, where each subject attempts to sort the visual stimuli[9],[10], 3) active-based selection: iterative process where selected pair(s) are evaluated one by one or in batches by one or more subjects, their preferences are recorded and used to determine the next pairs for evaluation in the next iteration, 4) predictive sampling: involves performing a subjective test with a pre-selected short list of pairs. For the remaining pairs, the pairwise preference is automatically predicted.

Xu et al. in[11] proposes an active sampling approach based on information maximization to improve the sampling efficiency in the context of a crowd-sourcing scenario. The framework of HodgeRank is exploited: HodgeRank is a method used to decompose ranking data into consistent and inconsistent parts, providing a structured way to handle noisy and conflicting input. A supervised sampling method involving the labels already collected is proposed based on Bayesian information maximization which selects pairs with the largest information gain. Chen et al. in[12] proposes the Crowd-BT algorithm, which extends the Bradley-Terry (BT) model to include annotator quality in crowd-sourcing. An active learning strategy is proposed that involves learning annotator quality and pairwise preferences. The experiments

show that modeling annotator quality improves ranking accuracy and reduces labeling costs. Li et al. in[13] introduce a hybrid sampling strategy either based on minimum spanning tree or global maximum. It uses a Bayesian optimization framework along with the BT model to create a utility function, which is then used to determine the expected information gain for each pair. Mikhailiuk et al. in[10] propose a technique based on information gain maximization to select the most informative pairs for comparison. The algorithm updates the full posterior distribution at each iteration, offering improved performance over existing methods. Several techniques to reduce computational complexity were proposed, namely by computing an approximate (online) posterior estimation (ASAP-approx), selective expected information gain (EIG) evaluations and a Minimum Spanning Tree (MST) mode for Batch Mode. Fan et al. in[14] combines the reliability with informativeness into an active sampling framework. Informativeness is computed based on the uncertainty or the amount of information an image pair can provide while reliability is estimated to account for the human visual system's limitations, especially when the quality difference between two images is too subtle.

Our previous solution proposed a predictive sampling method, exploiting machine learning tools. Mohammadi et al. in[15] proposes to train a binary classifier with a support vector machine (SVM) to determine which pairs should be included in the subjective test and a predictor with support vector regression (SVR) to estimate the preference between the images of a pair. This solution uses as features, the scores of seven popular objective quality metrics and includes a labeling procedure to obtain the ground-truth data that is essential to train the classifier.

The approach followed in this work can also be classified as predictive sampling, where some of the pairs are selected, prior to the pairwise comparison subjective test. However, instead of handcrafted feature extraction, image pair labeling and explicit training of a classifier and predictor, it is used a deep neural network with explicit uncertainty models. This new solution proposed here avoids the data labelling approach, which does not consider the reliability (or confidence) in the prediction of the pairwise probabilities.

B. Objective Quality Metrics

Objective image quality metrics that exploit pairwise rankings during training are reviewed next due to its relevance with the proposed pairwise sampling procedure. Zhang et al. in[16] proposed a deep neural network model for blind image quality assessment. The new model is called UNIQUE and is trained on multiple image quality assessment databases simultaneously by computing the probability of preference

between random pairs in each database; it is capable of generalizing to different types of distortions. Moreover, UNIQUE is evaluated against existing blind and full-reference image quality assessment models, demonstrating its superior performance. Cao et al. in[17] design a computational framework that integrates data-centric (reliable human-annotated datasets) and model-centric (reliable quality metrics) approaches. Specifically, they propose a sampling-worthiness model that quantifies the difficulty and diversity of candidate images, identifying diverse and challenging samples for dataset construction. Ma et al. in[18] address the challenge of limited training data by automatically generating quality-discriminable image pairs (DIPs) from large-scale databases, reducing the need for subjective testing. They employ RankNet, a pairwise learning-to-rank algorithm, to learn an opinion-unaware BIQA (OU-BIQA) model from these DIPs, incorporating perceptual uncertainty into the training process. The uncertainty is measured for each candidate pair using objective quality scores from three quality metrics. Liu et al. in[19] propose using ranking as a self-supervised auxiliary task to address the scarcity of labeled data in computer vision tasks, particularly for problems such as image quality assessment and crowd counting. In image quality assessment, they leverage on unlabeled data for training by generating distortions (adding noise, compression, etc) to reference images over a wide range of intensities; in such case, images can be ranked according to their quality since increasing distortion parameter implies decreasing image quality. Also, the network's certainty is evaluated by performing the self-supervised proxy task (ranking) multiple times. The certainty score reflects how confident the network is in its ranking predictions which is exploited during the training process.

The approach followed in this work is different from the reviewed objective quality metrics (mostly with blind image quality models), due to the exploitation of novel uncertainty models for the estimation of the probability of preference in a pairwise sampling subjective test and for the corresponding reliability of this estimation. These models are of upmost importance to reduce the number of required pairwise comparisons without compromising the accuracy of the global ranking of images.

III. Uncertainty Estimation

The main idea of this paper is to exploit uncertainty estimators with two main objectives: 1) achieve precise estimations of the probability of preference between two images in a pair, ensuring reliable results for the pairs that do not require human evaluation. 2) guide the pairwise sampling procedure to select the pairs that need human evaluation, thereby reducing the duration of a pairwise comparison subjective test.

The objective of this section is to introduce the proposed approach for uncertainty estimation. In this context, uncertainty estimation involves quantifying the uncertainty associated with the predictions made by a deep learning model. There are different ways to perform uncertainty estimation, which will be detailed in the following sections.

A. Data Uncertainty

Aleatoric uncertainty, also referred to as data uncertainty, is an intrinsic characteristic of the input data distribution that denotes the inherent noise or variability, and it can be directly inferred from the data[20]. This type of uncertainty is irreducible unless the data collection process is repeated. In the context of pairwise comparison assessment, the only method to mitigate data uncertainty (specifically, the variability in human decisions) is to directly request human observers to label the corresponding pair.

As in previous works[16][17], it is assumed that the true perceptual quality Q of an image follows a Gaussian distribution[3] with parameters μ and σ, which represent the mean and standard deviation of the quality, respectively. In the context of a pairwise comparison test, it is also assumed that the quality of the two images A and B of a pair, $Q(A) \sim \mathcal{N}(\mu_A, \sigma_A)$ and $Q(B) \sim \mathcal{N}(\mu_B, \sigma_B)$ are independent of each other and thus the quality difference $Q(A) - Q(B)$ distribution is also Gaussian as defined in (1).

$$Q(A) - Q(B) \sim \mathcal{N}(\mu_{AB}, \sigma_{AB}) \quad \text{s.t.}$$
$$\mu_{AB} = \mu_A - \mu_B, \ \sigma_{AB}^2 = \sigma_A^2 + \sigma_B^2 \tag{1}$$

The selection of image A over image B can be obtained by randomly sampling the quality difference $Q(A) - Q(B)$ distribution, and if the value obtained is greater than zero, image A is selected over image B. This means that the positive area under the $Q(A) - Q(B)$ distribution indicates the probability of preferring A over B, i.e. image A has higher quality than image B. This could be quantified using the Gaussian cumulative distribution function (CDF) $\Phi()$ of the quality difference distribution, $Q(A) - Q(B)$ [3] as defined in (2).

$$\hat{Pr}(A \succ B) = \Phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) \tag{2}$$

where the $\succ$ symbol in (2) represent the preference and $\Phi(z)$ is CDF.

The data uncertainty, which only depends on the input images, corresponds to how spread is the quality difference around the mean μ_{AB} and thus can be represented by the variance (σ_{AB}^2) of this distribution[21], i.e. a high variance (σ_{AB}^2) represents a large dispersion and thus an increased level of

uncertainty and lower variance indicates a distribution more centered around the mean, suggesting less uncertainty.

Figure. 1 provides a visual representation for quality difference distribution, where image A exhibits higher quality than image B. Notably, the positive area under the quality difference distribution $Q(A) - Q(B)$ exceeds the negative area, indicating higher preference for selecting A over B.
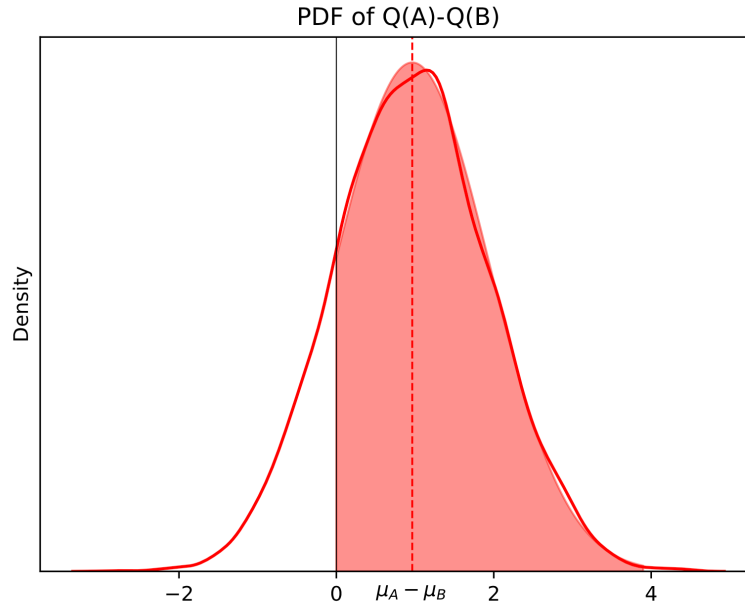

Figure 1. Quality difference, $Q(A) - Q(B)$, distributions. The shaded area under the probability density function (PDF) curve of $Q(A) - Q(B)$ represents the probability of preference $\hat{Pr}(A \succ B)$.

B. Model Uncertainty

Epistemic uncertainty, also known as model uncertainty, is an intrinsic characteristic of the model itself and arises from a lack of knowledge or data[20]. This type of uncertainty can be mitigated by incorporating additional data or enhancing the model architecture and training process. Model uncertainty can influence the selection of pairs for human evaluation. When the model exhibits low confidence in its prediction of the preference between two images in a pair (denoted as $\hat{Pr}$ in (2)), model uncertainty is high. In such instances, the correct preference must be determined by human subjects, and the model's prediction should be disregarded.

Monte Carlo dropout[22],[23], referred to as MC-dropout, serves as an effective approximation of probabilistic Bayesian models within deep Gaussian processes. This technique involves applying dropout during training to mitigate overfitting by randomly deactivating some neurons in each training iteration. Consequently, a different subset of the network architecture is used each time. With MC-dropout, multiple passes are also executed during inference where neurons are disabled in each pass based on the outcome of a Bernoulli random variable (0 or 1) with a probability of p(3). As a result, through n forward passes of the same input, varying subsets of network architectures produce distinct outputs. Consider a dropout layer present in both training and test that is applied to the previous layer which has weights W_i, thus obtaining a new set of weights D_i (after dropout) according to (3).

$$
\begin{aligned}
D_i &= W_i \cdot diag([z_{ij}]_{j=1}^{k_i}) \\
z_{i,j} &\sim Bernoulli(p_i) \text{ for } i = 1, \ldots, L, j = 1, \ldots, k_{i-1}
\end{aligned}
\tag{3}
$$

The binary variable $z_{i,j}$, used as input to layer i, indicates if unit j in layer $i-1$ from tensor W_i being dropped. p_i is the activation probability for layer i and is usually called dropout ratio and can be learned or set manually.

By repeatedly applying the same input, the model outputs can be interpreted as samples from a Gaussian distribution, providing an important characterization of the prediction's variability. In fact, quantifying this distribution will allow uncertain inputs to be treated differently. The predictive Gaussian distribution P_m is defined by parameters μ_m and σ_m, which are computed as $\mu_m = \frac{1}{n} \sum_{i=1}^{n} (\hat{Pr}(A \succ B))$ and $\sigma_m^2 = Var(\hat{Pr}(A \succ B))$, respectively. In the context of pairwise sampling, the model uncertainty is therefore characterized by distribution P_m serving two purposes:

1. Enhance the robustness of predictions by averaging multiple stochastic forward passes. Thus, μ_m will be used as an improved estimation of the probability of preference.

2. Quantification of the lack of confidence in the predicted preference between the two images of a pair. Higher values of σ_m^2 indicate greater dispersion around the mean, representing less reliable probability of preference estimation by the model. Thus, σ_m^2 will be used as a measure of reliability of the prediction.

IV. Proposed Pairwise Sampling Procedure

This section outlines the primary contribution of the paper: the comprehensive learning-based pairwise sampling procedure that leverages uncertainty models to achieve precise preference estimations and

identify the pairs that require subjective evaluation. To accomplish these goals, it is essential to detail the novel pairwise sampling framework (Section IV-A), the designed deep neural network model (Section IV-B) and the pairwise selection algorithm (Section IV-C).

A. Pairwise Sampling Framework

The proposed framework for pairwise sampling is summarized in the following steps:

1. **Quality distribution estimation:** first, the underlying quality distributions of both distorted images of a pair are calculated with a deep learning model presented in Section IV-B. More precisely, two differentiable learned functions $f_\mu()$ and $f_\sigma()$, calculate the mean quality score and associated standard deviation using both the reference and each distorted image of the pairs. This step receives as input a pair of distorted images, denoted as A and B and the corresponding reference images Ref. The output is the quality distribution parameters for each of the distorted images of the pair. i.e. μ_A, σ_A, μ_B and σ_B.

2. **Data uncertainty estimation:** the data uncertainty associated to the input images of the pair is then calculated according to (1). This step receives as input all parameters calculated in step 1, and outputs the data uncertainty associated to the pair represented by σ_{AB}^2. In Section III-A, data uncertainty was described in detail.

3. **Preference estimation:** the probability of preferring image A over image B in a pair denoted as $\hat{Pr}(A \succ B)$ is calculated with (2). This step receives as input μ_A and μ_B calculated in step 1 and the data uncertainty estimation of the previous step, i.e. σ_{AB}^2. The probability of preference $\hat{Pr}(A \succ B)$ is the output of this step. In Section III-A, this step was described in detail.

4. **Model uncertainty estimation:** the model uncertainty associated to the deep learning model is characterized by the Gaussian distribution P_m. This step receives as input a pair of images, repeatedly applies the same pair to the model (MC-dropout estimate) and computes the probability of preference (step 1 to 3). By running n forward passes of a network with dropout in both functions $f_\mu()$ and $f_\sigma()$, n predictions of $\hat{Pr}(A \succ B)$ are obtained since multiple estimates of μ_A, σ_A, μ_B and σ_B are provided by the neural network. This multiple passes allow to compute distribution P_m and thus characterise the model uncertainty. The μ_m and σ_m^2 parameters of P_m are used as a more reliable estimation of $\hat{Pr}(A \succ B)$ and as the dispersion (and thus confidence) in the model predictions, respectively. In Section III-B model uncertainty was described in detail.

5. **Pair selection procedure:** this procedure identifies the pairs for which the model is unreliable and should thus be evaluated by humans. The above steps are repeated for all possible pairs to calculate the data and model uncertainty associated with each one. All pairs are then sorted based on some specific criteria in descending order, and the top $X\%$ of the pairs are selected according to available budget. Three criteria are considered: two of those are rather straightforward, which is to sort according to σ^2_{AB} or σ^2_m, i.e. from higher to lower data or model uncertainty, respectively. However, these two criteria do not consider the impact on the final scores of choosing a pair for subjective assessment. Therefore, a new criteria for pair selection based on expected information change (EIC) is proposed (refer to Section IV-C for details). In this step, the budget directly defines the length of the subjective test.

6. **Subjective assessment test:** finally, the pairs selected in the previous step undergo a subjective assessment test following a pairwise comparison methodology. Each paired comparison is evaluated by several subjects and thus the relative preference between two images can be obtained, i.e. probability of preference $Pr(A \succ B)$.

7. **Preference aggregation:** A pairwise comparison (PC) matrix is created and populated with μ_m of step 4 ($\hat{Pr}(A \succ B)$ estimate) and the $Pr(A \succ B)$ of the previous step (subjective score). Finally, the PC matrix is converted to scores using the Bradley-Terry preference aggregation model. The BT model parameters are estimated using maximum likelihood estimation and represent the preference strength of each image relative to the others. Images that are consistently preferred over others will receive higher scores. The output of this step is a quality score per image, which quantifies the relative attractiveness or preference for all distorted images.
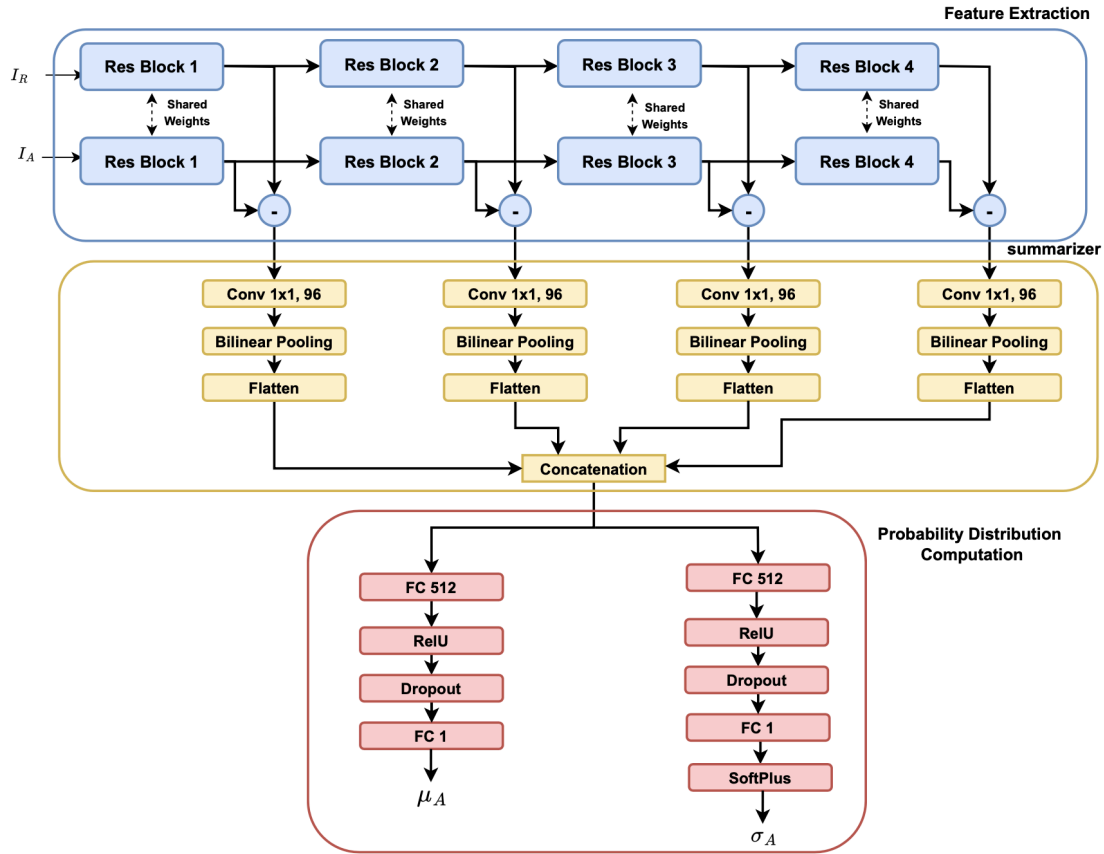
Figure 2. Proposed architecture of the deep learning framework ($f_{\mu}()$ and $f_{\sigma}()$).

B. Network Architecture

The function $f_{\mu}()$ and $f_{\sigma}()$ estimate the underlying quality distribution of each image of the pair using a deep network model. The proposed network architecture is shown in Fig. 2 for image I_A, leveraging ResNet34[24] as the backbone for feature extractor. The input is a distorted image (which can be I_A or I_B) and the corresponding reference, I_R, and the output corresponds to the quality probability distribution parameters, μ and σ, of the distorted image relative to the reference.

The weights of the feature extraction network for the distorted image are shared with the feature extraction network for the reference image. Residual maps are computed as the distance between the distorted image features and the reference image features after each residual block which facilitate the flow of gradient in back–propagation and leverage both high–level and low–level features (different scales of the input image) for the probability distribution computation. Each residual activation map then goes

through a 1×1 convolution to enhance cross-channel interaction and obtain a smaller number of channels, which is important to reduce the number of parameters and computational cost of the fully connected layers at the end of the model. Note that features are obtained in different levels of the feature extractor network backbone (and thus different resolutions) but the number of output channels is the same for all 1×1 convolutions. After, bilinear pooling[25] is applied to capture second-order statistics and pairwise feature interactions according to (4), resulting in more detailed and discriminative set of features, which is essential for accurately estimate the probability distribution parameters for each image. Before bilinear pooling, spatial, s, and then channel, c, dimensions are flattened to obtain a vector $z \in \mathbb{R}^{s \times c}$. The bilinear pooling is defined in (4).

$$\hat{Z} = z^T z \tag{4}$$

After flattening and concatenation, the features are passed through function $f_\mu()$ which is represented by a fully connected layer with 512 units and the ReLU non-linear activation. Following the model uncertainty approach described in III-B, a dropout layer was introduced which is applied both during training and inference. In this case, dropout randomly "drops out" (sets to zero) a fraction of the neurons in the previous layer. This means that these neurons are temporarily removed from the network, along with all their incoming and outgoing connections. The dropped neurons are selected at random with a probability p, known as the dropout rate. After, another fully connected layer with a single unit is used to generate the mean, μ, of the Gaussian distribution parameter.

Regarding, function $f_\sigma()$, the concatenated feature maps are passed through a fully connected layer with 512 outputs and a ReLU activation, followed by dropout as in function $f_\mu()$. Since both functions $f()$ have dropout layers, the probability of preference is estimated in multiple passes using different values of μ and σ; this enables the computation of model uncertainty. Then, another fully connected layer with a single output and a SoftPlus activation is used to constrain the values of σ to be positive while avoiding the problem of having zero gradients for negative inputs (also being more smooth than ReLU around zero). Moreover, in Section II of the supplementary material, all details of the deep learning network architecture are presented.

C. Pair Selection Procedure

This section introduces a novel procedure to select pairs aimed at enhancing the accuracy of scores obtained after preference aggregation. An expected information change criterion is proposed, measuring the impact on the final subjective scores when a pair is chosen for subjective evaluation. The sampling

algorithm relies on the computation of the posterior distribution over the scores obtained from selecting a pair for subjective assessment. This criterion is based on the typical utility function of[10][13], where the expected information gain (EIG) is computed by considering the two possible choices when a subject evaluate a pair. The posterior distribution can help to select the next comparison by accounting for possible deviations in the estimated probability of preference according to the model uncertainty.

Initially, a pairwise comparison matrix M is created and populated with the mean predicted preferences μ_m obtained from the model uncertainty step. This is the initial state and corresponds to the probability of preference estimated from the score distributions obtained by the deep network model presented in Section IV-B. Next, the prior distribution of quality scores $p(S|M)$ are derived from M using the Bradley–Terry algorithm. To account for potential discrepancies, the posterior distribution $P_{ij}^{\pm}(S|M,A)$ is computed for each pair (i,j) by updating the predicted preferences of M with A_{ij} which is calculated according to (5).

$$A_{ij} = \text{Clip}(M(i,j) \pm \max(\delta, \sigma_m^2(i,j)), 0, 1) \tag{5}$$

Since predicting the outcome of selecting a pair is not possible, two possible cases are considered for pair (i,j): i) increasing or ii) decreasing $M(i,j)$ by some amount. This amount should reflect the model's uncertainty about its probability of preference estimation. In this work, it was found that the update provided by (5) works best, i.e. adjusting the preference by $\pm\delta$ or the model uncertainty $\sigma_m^2(i,j)$, which is first normalized using min–max normalization. When adjusting the preferences of M, the results are clipped with $\text{clip}(x,a,b) = \max(a, \min(x,b))$ to ensure the new values remain valid probabilities between zero and one, as specified in (5). Next, the Expected Information Change (EIC) for each pair is calculated according to (6). In this context, the EIC corresponds to the Kullback–Leibler (KL) divergence D_{KL} between the prior and posterior scores, which are modeled as multivariate normal distributions.

$$\begin{aligned} EIC(i,j) = &D_{KL}(p(S|M)||P_{i,j}^{+}(S|M,A)) + \\ &D_{KL}(p(S|M)||P_{i,j}^{-}(S|M,A)) \end{aligned} \tag{6}$$

This process is repeated for every pair of images (i,j) separately, and pairs expected to yield the maximum EIC are selected, i.e. pairs are sorted from the maximum to the minimum EIC. This approach ensures that the most impactful pairs are selected, improving the accuracy and robustness of the final quality score predictions.

V. Performance Evaluation

In this section, the proposed solution is thoroughly evaluated. First, the training procedure of the proposed deep learning model is presented, followed by the dataset used and the test conditions (for inference). Then, the experimental results are presented, namely for the evaluation of the deep learning model, pair selection and comparison with state of the art. From now on, the proposed Learning-Based Pairwise Sampling is referred to as LBPS-Data for data-based sampling, LBPS-Model for model-based sampling, and LBPS-EIC for EIC-based sampling.

A. Training Procedure

The objective of the training procedure is to find network parameters θ, that given the input images A, B and Ref, can accurately characterize the quality of both images of a pair. The input images are first center cropped to 224×224 as typical when the ResNet network architecture is used. The network parameters are initialized with the weights trained on ImageNet[26]. The training hyper parameters were carefully selected to optimize performance, namely, 8 pairs of images, A, B and Ref, were chosen randomly for each batch, and the learning rate was initialized to $1E-5$. A scheduler reduces the learning rate by a factor of 0.1 after each 10 epochs to improve convergence and preventing over-fitting. The dropout ratio used for training but also for inference is 0.2.

The loss function, defined in (7), measures the similarity between two probability functions: namely $\Pr(A \succ B)$, the probability of image A being preferred over B from the ground truth dataset, and the corresponding probability $\hat{Pr}(A \succ B)$ estimated using (2), the distribution parameters estimated by the proposed neural network model.

$$
\begin{aligned}
Loss = & 1 - \sqrt{\Pr(A \succ B)\hat{Pr}(A \succ B)} \\
& - \sqrt{(1 - \Pr(A \succ B)(1 - \hat{Pr}(A \succ B))}
\end{aligned}
\tag{7}
$$

The Adam optimizer was used to adjust the network's parameters based on the computed gradients of the loss function.

B. Training and Test Datasets

The proposed architecture was trained using PieAPP dataset[27] which consists of three disjoint sets for train, validation and test. The training and validation set consist of 140, and 20 reference contents respectively with 123 degraded images associated with each reference. Specifically, the PC subjective test

was designed for two kinds of pairs: inter-pairs and intra-pairs, which represent different distortion scenarios for each pair of images. In an inter-type comparison, A and B have different types of distortions. However, in intra-type comparison, A and B have the same type of distortions. The training set comprises 77k image pairs, and the validation set includes 9k pairs. Additionally, the test set consists of 40 reference images, each subjected to 15 different degradation types. The degradation types present in the test set differ entirely from those in the training and validation sets. This dataset configuration ensures diverse training examples, and assesses the model's ability to generalize to unseen degradation types. Each example in the aforementioned sets is labeled with the probability of preference, $Pr(A \succ B)$ obtained through a crowd-sourcing based PC subjective test in which subjects were asked to select the image that is more similar to the reference image.

Additionally, the proposed LBPS solutions were evaluated using the Pairwise Comparison Image Quality Assessment (PC-IQA) dataset[28]. This dataset was collected through a crowd-sourced pairwise subjective test methodology and includes 15 reference images, each with 15 corresponding distorted versions. As a result, the dataset comprises 1,800 image pairs, with 120 pairs per reference image. While the dataset is comprehensive, comparing all possible pairs, it is imbalanced because each pair has been assessed a different number of times.

C. Test Procedure

To assess the proposed framework, a simple algorithm is employed that utilizes several pairwise comparison matrices (PCMs) $\mathbf{m}$. Each PCM captures the preferences of a set of images linked to a specific reference (or content)[29]. This set of images comprises the original image and its degraded versions. Thus, the test procedure begins with the creation of the $\mathbf{m}_i$ PCMs which are populated with the preferences of the ground-truth for reference i. After, $\hat{\mathbf{m}}_i$ PCMs are created and populated with estimated preferences for a particular method. For instance, for LBPS-Data, these preferences are $\hat{Pr}(A \succ B)$ and for LBPS-Model and LBPS-EIC, the preferences are μ_m. Moreover, some preferences in $\hat{\mathbf{m}}_i$ may have been obtained from ground truth scores if the pair was selected for human evaluation. After, the BT preference aggregation model is applied to each $\mathbf{m}_i$ to infer scores $\mathbf{s}_i$ and to each $\hat{\mathbf{m}}_i$ to obtain $\hat{\mathbf{s}}_i$. The scores $\mathbf{s}_i$ and $\hat{\mathbf{s}}_i$ for all the references i are concatenated and some metric is applied to compute the correlation between $\mathbf{s}$ and $\hat{\mathbf{s}}$. This allows for the assessment of the quality scores produced for the entire dataset and not just for one reference.

In this work, three correlation metrics are employed: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-Order Correlation Coefficient (SROCC), and Root Mean Square Error (RMSE). PLCC is utilized to measure the linear correlation between the predicted and the ground truth quality scores, indicating how closely the predicted scores align with the ground truth scores. SROCC assesses the strength and direction of the monotonic relationship between the predicted and ground truth quality scores, providing insight into the consistency of the ranking order produced by the model relative to the ground truth. RMSE, meanwhile, quantifies the average magnitude of the error between the predicted and actual quality scores providing a direct measure of prediction accuracy.

D. Anchors

The proposed LBPS-EIC solution was evaluated against several state-of-the-art methods to assess its performance and robustness, including our previous work, PS-PC. Random sampling is the most straightforward solution and serves as baseline. In such case, the initial preference is set to 0.5, indicating no preference, and pairs are selected randomly for comparison. The random sampling process is repeated 25 times, and the average performance is reported. Moreover, in this evaluation, PS-PC uses four classifier/predictor models, which results in fixed number of selected pairs were used. Other active sampling approaches were included as anchors, particularly those more efficient, such as ASAP[10], Crowd-BT[12] and Hybrid-MST[13], all with the highest performing configuration. Active sampling approaches function differently in that they do not select pairs directly. Instead, they choose a trial (a comparison) or a set of trials in each run of the algorithm. Consequently, it requires raw judgments from subjects for each pair, which are not available in some datasets. For instance, the PieAPP dataset only provides the probability of preference between all possible pairs of images, requiring the creation of binary judgments to be used in the evaluation of active sampling methods.

Therefore, for the PieAPP dataset, the observed judgments were simulated based on the probability of preference available for each pair of images. A Bernoulli distribution was used, assuming that individual trials are independent, where the success probability, p, represents the probability of preferring one image over the other. Sampling the Bernoulli distribution involves generating a random variable (using a uniform distribution) and choosing image A over image B if the generated random number is less than the ground truth preference. Due to inherent randomness in generating binary judgments, 25 iterations of the active sampling approaches were conducted, and the average is reported. Moreover, the PC matrix was initialized with a probability of preference 0.5 for the anchors.

Additionally, the proposed LBPS–EIC solution was evaluated on PC–IQA, which binary judgements are available in the dataset, to further assess the performance. It is important to note that anchors randomly select a judgment from all the available judgments. Since these decisions are obtained randomly, an average of 25 iterations is used.

E. Experimental Results

This Section presents the several experiments made to evaluate the main contributions of this work, namely the performance of the deep learning model (Section V–E1), the proposed pair selection method (Section V–E2) and the evaluation of the proposed solution with the highest performance, against the anchors previously defined (Section V–E3). A qualitative analysis with a focus on high uncertainty cases is shown in Section I of the supplementary material. Importantly, in all these experiments, 200 iterations was used for MC–Dropout for each of the datasets, and δ is configured to 0.3 in LBPS–EIC criteria.

1) Deep Learning Model Evaluation

In this Section, the probability of preference $\hat{Pr}(A \succ B)$ computed for all pairs from the quality distributions estimated by the deep neural network model is analyzed in two different experiments.

First, Fig. 3 illustrates the relationship between the predicted $\hat{Pr}(A \succ B)$ and ground truth preferences $Pr(A \succ B)$ for all pairs of the PieAPP test set, and PC–IQA dataset, providing a clear and comprehensive visual summary of the model's performance. As shown, the data points are clustered near the $y = x$ line which indicates the model's strong performance in estimating the probability of preference.
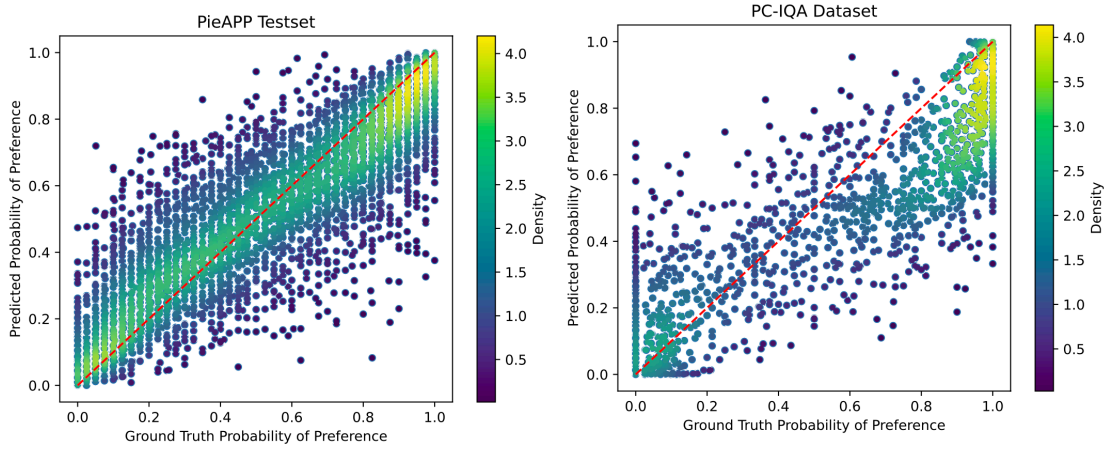
Figure 3. Predicted preference versus ground truth preference using a scatter plot combined with a heatmap. The heatmap overlays the scatter plot to illustrate the density of data points.

Fig. 4 illustrates the relationship between the predicted preferences, denoted as $\hat{Pr}(A \succ B)$, and the estimated model uncertainty, σ_m^2. As shown, model uncertainty is higher when the images of a pair have similar quality levels, which means probability of preference around 0.5. Moreover, when one of the images has significantly higher or lower quality than the other (probability of preference close to 1 or 0), model uncertainty decreases as expected.
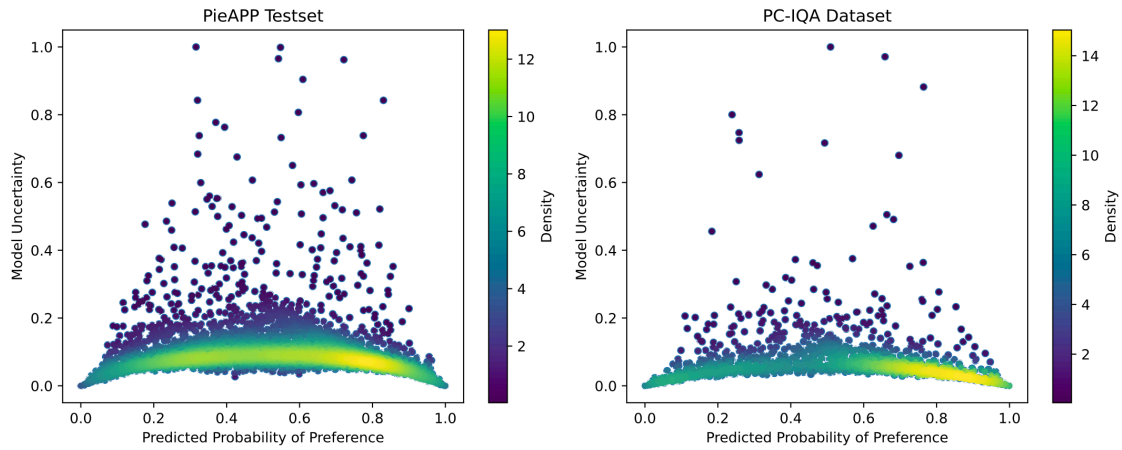

Figure 4. Model uncertainty versus predicted preferences.

2) Pair Selection Evaluation

In this Section, the LBPS–Data, LBPS–Model, LBPS–EIC pair selection methods are evaluated. Fig. 5 presents the experimental results for all methods using the three performance metrics PLCC, RMSE and SROCC following the test procedure of Section V-D. As expected, the overall performance improves with an increasing budget of selected pairs (denoted as $X\%$) for subjective assessment. When no pairs are selected (0%), the performance reflects the model's predictive capability, resulting in identical performance across all criteria.

However, when the budged increases, LBPS–EIC, which leverages both data and model uncertainties, achieves the highest performance. Specifically, selecting 10% of pairs yields a PLCC correlation above 0.92, which rises to nearly 0.99 at 50% selection. LBPS–Model demonstrates the second-best performance when fewer than 40% of pairs are selected, but its performance lowers compared to LBPS–Data as the number of deferred pairs increases. This is expected since when many pairs are selected, data uncertainity is the best performing driving to select pairs.

Regarding the SROCC metric, overall performance also improves with an increasing number of selected pairs for subjective assessment. However, all three criteria exhibit similar performance, indicating minimal change in the rank order of predictions or minor misordering. This suggests that while all uncertain estimation models enhance the fidelity of the scores, they still maintain the ranking of the stimuli. In contrast, RMSE shows substantial performance gains, indicating a closer alignment between predicted and ground truth values. In conclusion, these findings underscore the effectiveness of LBPS–EIC by incorporating both data and model uncertainties, particularly in scenarios involving a low to medium budget of selected pairs for subjective assessment.
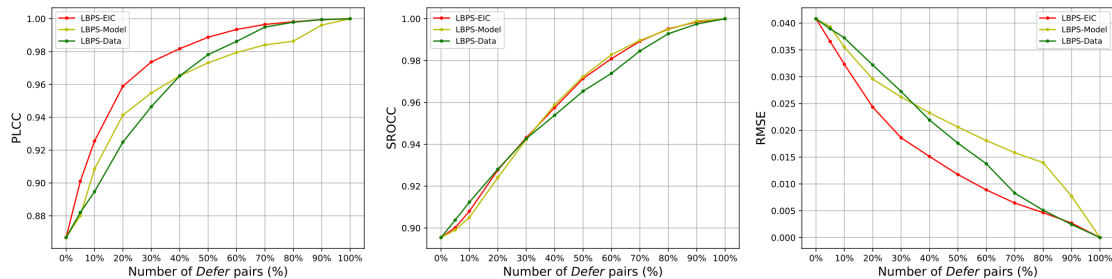

Figure 5. Evaluation of the pairwise selection method

3. State-of-the-art Comparison

In this section, the best-performing method, LBPS–EIC, is evaluated in comparison to the anchors, following the test procedure outlined in Section V-C. To facilitate this evaluation, a fixed number of trials (budget) was selected to evaluate all methods: 2.5%, 5%, 7.5%, 10%, 15%, 20%, 30%, 40% and 50%. Here, the budget defines the number of trials that can be conducted, expressed as a percentage. The maximum number of trials is set to $\frac{n(n-1)}{2} \times 15$, with 15 representing the number of subjects. For the anchors, the average over 25 runs is reported and the standard deviation over these runs is not significant and thus is not reported. The performance of the LBPS–EIC is compared with selected anchors using all test datasets. Experimental results for the PS-PC method were not included for the PC-IQA dataset since the training (used for training PS-PC) and test datasets are rather similar due to the use of the same degradation types.

As shown in Table I and Table II, LBPS–EIC demonstrates notable performance improvements compared to random sampling across both datasets. For a budget of 50% trials, LBPS–EIC achieves a PLCC correlation of 0.99/0.98 for the PieApp/PC-IQA datasets, compared to 0.83/0.66 for random sampling. This indicates that at 50% of the trials, the scores obtained after pair selection with LBPS–EIC are already very similar to the ground truth scores. Furthermore, as shown in Table I and Table II, LBPS–EIC consistently exhibits the highest PLCC correlation across all trial percentages compared to other methods for both datasets. For the PieApp dataset, at 10% of trials, LBPS–EIC achieves a PLCC correlation of 0.93, whereas the anchors yield lower correlations of 0.82, 0.79, 0.86, and 0.80 for ASAP, Crowd–BT, Hybrid–MST, and PS–PC, respectively. For the PC-IQA dataset a similar behaviour is observed, at 10% of trials, LBPS–EIC achieves a PLCC correlation of 0.91, whereas the anchors have much lower correlations between 0.67 to 0.86.

On the other hand, considering the SROCC correlation metric, Hybrid–MST has the highest correlation, albeit only slightly superior compared to LBPS–EIC. The difference between SROCC values is less than 0.03 for the PieApp dataset and 0.04 for the PC-IQA and only occurs for budgets greater than 10%. However, LBPS–EIC continues to show improvements in correlation performance as the trial budget increases and has almost the same performance when 50% of the pairs are selected. Note also, that for a budget less than 10%, LBPS–EIC achieves the best performance since its able to provide accurate estimation of the probability of preference and selects the most impactfull pairs. An important point to note is that active sampling methods that are used as anchor require running the sampling procedure after each trial (or a batch of trials), which is challenging in many scenarios (e.g., crowd–sourcing) where

many subjects evaluate pairs in parallel or the computational resources to select the pairs are insufficient. For instance, if four subjects are doing the test in parallel the selection performance of Hybrid–MST batch mode will be lower since all the data is only obtained (and thus can be used to compute the next pairs) after the four subjects finish their evaluations. In the proposed solutions, this procedure is much more simplified since the selection of pairs is done *a priori* of the subjective test.

Method	Metric	Trial percentage								
		2.5%	5%	7.5%	10%	15%	20%	30%	40%	50%
LBPS–EIC	PLCC	**0.89**	**0.90**	**0.92**	**0.93**	**0.94**	**0.96**	**0.97**	**0.98**	**0.99**
	SROCC	**0.89**	**0.90**	**0.90**	0.91	0.92	0.93	0.94	0.96	0.97
ASAP	PLCC	0.55	0.71	0.79	0.82	0.87	0.89	0.92	0.94	0.95
	SROCC	0.73	0.85	0.90	0.92	0.94	0.96	0.97	0.98	0.98
Crowd–BT	PLCC	0.53	0.67	0.74	0.79	0.83	0.86	0.89	0.90	0.91
	SROCC	0.64	0.78	0.85	0.89	0.93	0.94	0.95	0.96	0.96
Hybrid–MST	PLCC	0.67	0.78	0.83	0.86	0.90	0.92	0.94	0.95	0.95
	SROCC	0.77	0.87	0.90	**0.92**	**0.95**	**0.96**	**0.97**	**0.98**	**0.98**
PS–PC	PLCC	–	–	–	0.80	–	0.84	0.86	0.89	0.89
	SROCC	–	–	–	0.44	–	0.49	0.53	0.60	0.62
Random	PLCC	0.30	0.42	0.49	0.55	0.61	0.66	0.74	0.78	0.83
	SROCC	0.31	0.48	0.6	0.67	0.75	0.81	0.87	0.91	0.93

Table I. LPBS–EIC performance evaluation compared to anchors for the PieAPP test set.

Method	Metric	Trial percentage								
		2.5%	5%	7.5%	10%	15%	20%	30%	40%	50%
LBPS–EIC	PLCC	**0.83**	**0.87**	**0.89**	**0.91**	**0.93**	**0.95**	**0.96**	**0.98**	**0.98**
	SROCC	**0.83**	**0.84**	0.85	0.85	0.87	0.9	0.92	0.95	0.96
ASAP	PLCC	0.45	0.60	0.70	0.74	0.79	0.83	0.88	0.92	0.94
	SROCC	0.73	0.82	0.85	0.86	0.88	0.89	0.91	0.92	0.93
Crowd–BT	PLCC	0.44	0.56	0.62	0.67	0.73	0.78	0.83	0.86	0.88
	SROCC	0.62	0.74	0.80	0.83	0.85	0.87	0.89	0.9	0.90
Hybrid–MST	PLCC	0.66	0.78	0.83	0.86	0.91	0.94	**0.96**	0.97	0.99
	SROCC	0.79	0.85	**0.87**	**0.89**	**0.91**	**0.92**	**0.94**	**0.96**	**0.96**
Random	PLCC	0.23	0.32	0.35	0.40	0.46	0.50	0.57	0.63	0.66
	SROCC	0.31	0.44	0.54	0.58	0.67	0.71	0.77	0.81	0.83

Table II. LBPS–EIC performance evaluation compared to anchors for the PC-IQA test set.

VI. Conclusion

In conclusion, the uncertainty-based sampling methods introduced in this work represent a significant improvement for pairwise comparison subjective assessment tests by reducing their duration while maintaining accurate quality scores. The LBPS–EIC framework stands out as a practical solution applicable to various image processing tasks, ensuring high-quality assessments with less human effort and lower computational costs. Utilizing deep learning models to predict human preferences and guide the sampling of pairs requiring human evaluation has proven highly effective. This approach models both aleatoric and epistemic uncertainty to achieve accurate preference predictions and optimize pairwise sampling, outperforming traditional active sampling methods. These advancements mark a substantial step forward in balancing the efficiency and precision of subjective quality assessments, ultimately benefiting both benchmarking and the training of learning-based quality metrics. For future work, it is planned to integrate both learning-based and active sampling approaches with a

reinforcement learning framework, allowing an intelligent agent to learn to select pairs through interaction with human subjects.

References

1. ^Recommendation ITU-T P.913 (2016). "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment." International Telecommunication Union. ISSN 0923-5965.

2. ^{ITU-R Question 102-3/6} (2019). "Recommendation 500-19: Methodology for the subjective assessment of the quality of television pictures." ITU-R Recommendation BT.

3. a, b, cThurstone L. "A Law of Comparative Judgment." Psychology Review. **34**: 273-86, April 1994. doi:10.1037/0033-295X.101.2.266.

4. ^Bradley RA, Terry ME (1952). "Rank analysis of incomplete block designs: I. The method of paired comparisons." Biometrika. **39** (3/4): 324–345.

5. ^Pinson MH, Wolf S. "Comparing subjective video quality testing methodologies." In: Visual Communications and Image Processing. 2003. **5150**:573–582.

6. ^Tominaga T, Hayashi T, Okamoto J, Takahashi A (2010). "Performance comparisons of subjective quality assessment methods for mobile video." In: International Workshop on Quality of Multimedia Experience, Trondheim, Norway, June 2010, pp. 82-87. doi:10.1109/QOMEX.2010.5517948.

7. ^Mantiuk RK, Tomaszewska A, Mantiuk R (2012). "Comparison of four subjective methods for image quality assessment." Computer graphics forum. **31**(8): 2478–2491.

8. ^Xu Q, Huang Q, Jiang T, Yan B, Lin W, Yao Y (2012). "HodgeRank on Random Graphs for Subjective Video Quality Assessment". IEEE Transactions on Multimedia. **14** (3): 844-857. doi:10.1109/TMM.2012.2190924.

9. ^Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, Vozel B, Chehdi K, Carli M, Battisti F, et al. Image database TID2013: Peculiarities, results and perspectives. Signal processing: Image communication. **30**:57–77, 2015.

10. a, b, c, dMikhailiuk A, Wilmot C, Perez-Ortiz M, Yue D, Mantiuk RK. "Active sampling for pairwise comparisons via approximate message passing and information gain maximization." In: International Conference on Pattern Recognition; January 2021; Milan, Italy.

11. ^Xu Q, Xiong J, Chen X, Huang Q, Yao Y (2018). "Hodgerank with information maximization for crowdsourced pairwise ranking aggregation." In: AAAI Conference on Artificial Intelligence. **32**. Louisiana, USA.

12. a, b Chen X, Bennett PN, Collins-Thompson K, Horvitz E. "Pairwise ranking aggregation in a crowdsourced setting." In: ACM international conference on Web search and data mining; February 2013; Rome, Italy. p. 193–202.

13. a, b, c Li J, Mantiuk R, Wang J, Ling S, Le Callet P. "Hybrid-MST: A Hybrid Active Sampling Strategy for Pairwise Preference Aggregation." In: Advances in Neural Information Processing Systems. Vol. **31**. Montréal, Canada; 2018.

14. ^Fan Z, Jiang T, Huang T (2017). "Active Sampling Exploiting Reliable Informativeness for Subjective Image Quality Assessment Based on Pairwise Comparison." IEEE Transactions on Multimedia. **19** (12): 2720–2735. doi:10.1109/TMM.2017.2711860.

15. ^Mohammadi S, Ascenso J. "Predictive Sampling for Efficient Pairwise Subjective Image Quality Assessment." In: Proceedings of the 31st ACM International Conference on Multimedia, MM '23, New York, NY, USA, October 2023. p. 6676–6684. doi:10.1145/3581783.3612443.

16. a, b Zhang W, Ma K, Zhai G, Yang X (2021). "Uncertainty-aware blind image quality assessment in the laboratory and wild." IEEE Transactions on Image Processing. **30**: 3474–3486.

17. a, b Cao P, Li D, Ma K. "Image Quality Assessment: Integrating Model-Centric and Data-Centric Approaches." In: Conference on Parsimony and Learning; January 2024. p. 529-541.

18. ^Ma K, Liu W, Liu T, Wang Z, Tao D (2017). "dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs". IEEE Transactions on Imgage Processing. **26** (8): 3951–3964. doi:10.1109/TIP.2017.2708503.

19. ^Liu X, van de Weijer J, Bagdanov AD (2019). "Exploiting Unlabeled Data in CNNs by Self-Supervised Learning to Rank." IEEE Transactions on Pattern Analysis and Machine Intelligence. **41** (8): 1862–1878. doi:10.1109/TPAMI.2019.2899857.

20. a, b Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR, Makarenkov V, Nahavandi S (2021). "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." Information Fusion. **76**: 243–297. doi:10.1016/j.inffus.2021.05.008.

21. ^Tsukida K, Gupta M (2011). "How to Analyze Paired Comparison Data." UWEE Technical Report 206, 2011. January 2011.

22. ^Gal Y, Ghahramani Z. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In: Balcan MF, Weinberger KQ, editors. Proceedings of The 33rd International Conference on Machi

ne Learning. Proceedings of Machine Learning Research. Vol. 48. New York, USA; June 2016. p. 1050-1059.

23. ^*Gawlikowski J, Tassi C, Ali M, Lee J, Humt M, Feng J, Kruspe A, Triebel R, Jung P, Roscher R, Shahzad M, Yang W, Bamler R, Zhu X (2023). "A survey of uncertainty in deep neural networks." Artificial Intelligence Review. **56**: 1–77. doi:10.1007/s10462-023-10562-9.*

24. ^*He K, Zhang X, Ren S, Sun J. "Deep residual learning for image recognition." In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016 June; Las Vegas, NV, USA. p. 770-778.*

25. ^*Lin T-Y, RoyChowdhury A, Maji S (2015). "Bilinear CNN Models for Fine-Grained Visual Recognition." In: IEEE International Conference on Computer Vision (ICCV). pp. 1449-1457. doi:10.1109/ICCV.2015.170.*

26. ^*Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009). "ImageNet: A large-scale hierarchical image database." In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. pp. 248-255. doi:10.1109/CVPR.2009.5206848.*

27. ^*Prashnani E, Cai H, Mostofi Y, Sen P. "PieAPP: Perceptual Image-Error Assessment Through Pairwise Preference." In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018 June; UT, USA.*

28. ^*Xu Q, Huang Q, Yao Y (2012). "Online Crowdsourcing Subjective Image Quality Assessment." In: ACM International Conference on Multimedia, ser. MM '12, October 2012, p. 359–368. doi:10.1145/2393347.2393400.*

29. ^*Mohammadi S, Ascenso J. "Evaluation of Sampling Algorithms for a Pairwise Subjective Assessment Methodology." In: IEEE International Symposium on Multimedia (ISM); December 2022; Italy. pp. 288-292. doi:10.1109/ISM55400.2022.10040647.*

Declarations