

Full Length Article

Predicting image quality score distribution using deep neural networks[☆]Nikola Plavac^{ID*}, Seyed Ali Amirshahi^{ID}, Marius Pedersen^{ID}, Sophie Triantaphillidou^{ID}

Department of Computer Science, Norwegian University of Science and Technology, Gjøvik, 2815, Norway

ARTICLE INFO

Keywords:

Blind image quality assessment
Image quality score distribution
Deep learning
Transfer learning

ABSTRACT

Blind objective image quality assessment methods typically predict a single Mean Opinion Score (MOS) to represent perceived image quality. However, MOS discards valuable information about observer variability, as different score distributions can produce the same mean value. Predicting the full quality score distribution provides a richer and more realistic representation of perceptual quality. In this work, we propose an Image Quality Distribution Network (IQDN) in five configurations: a custom convolutional neural network trained from scratch and four transfer learning variants based on ResNet50, VGG16, Xception, and DenseNet121 backbones. The models were trained to predict five-bin normalized quality score histograms on three datasets: KonIQ-10k, CID2013, and a newly introduced NAP540 dataset. Performance was evaluated using multiple point-wise error and distribution similarity metrics. Results show that the Image Quality Distribution Network with Xception backbone consistently performs better than existing state-of-the-art approaches in terms of Earth Mover's Distance, capturing score distributions more effectively. Furthermore, a hybrid loss function combining Kullback-Leibler (KL) divergence and Huber loss improved performance across datasets. The predicted distributions also reconstruct MOS that closely correlate with ground-truth MOS. Finally, incorporating a quality-aware backbone demonstrates strong potential, particularly in cross-dataset testing scenarios.

1. Introduction

Image Quality Assessment (IQA) is a valuable tool for ensuring quality wherever imaging is used. IQA methods can be classified into two main groups: subjective and objective. In subjective IQA, observers rate the quality of an image based on specific criteria through psychophysical experiments [1]. The quality scores given to the image by all observers are then often averaged and expressed as a Mean Opinion Score (MOS). Objective IQA methods use mathematical techniques to produce quality scores, often referred to as image quality metrics. Depending on the availability of the reference image, objective metrics can be classified into three main groups: full-reference, reduced-reference, and no-reference (blind) quality metrics [2]. The goal of most objective IQA metrics is to provide a single score that corresponds to (or correlates with) the MOS assigned to the image through a subjective experiment. Although the MOS represents the opinion of an “average observer”, it contains no information about the individual scores assigned by each observer. As a result, images may have the same MOS values while having entirely different subjective score distributions (Fig. 1). In addition, the standard deviation of the observers' ratings, which is important to evaluate the agreement or disagreement between observers, is not taken into account when the image quality is expressed solely through MOS.

To address this matter, different studies [3–6] have proposed deep neural network architectures for predicting a distribution of quality ratings rather than MOS. Some of these studies assume that observers' ratings follow a specific (e.g., Gaussian) distribution [3,5], some apply several augmentation methods altering the original image quality [4], and some limit their studies to specific dataset(s) [3,4,7]. The goal of this work is to perform a systematic study on how effective a simple Convolutional Neural Network (CNN) and different transfer learning backbones can be in predicting the distribution of subjective quality scores given to images. For this purpose, we propose an Image Quality Distribution Network (IQDN) with five different configurations. Additional experiments have been performed, including experimentation with different loss functions and a study on whether using pre-trained backbones of a quality-aware network can lead to more robust results.

The contributions of this paper are as follows: (i) We introduce the IQDN, a blind IQA framework designed to predict full quality score distributions rather than single MOS values, implemented in five configurations (a custom CNN and four transfer learning backbones). (ii) We present a systematic evaluation of these architectures on three datasets (KonIQ-10k [9], CID2013 [10], and a new NAP540 [8] dataset), using a diverse set of distributional and point-wise error

[☆] This paper has been recommended for acceptance by Junsong Yuan.

* Corresponding author.

E-mail address: nikola.plavac@ntnu.no (N. Plavac).

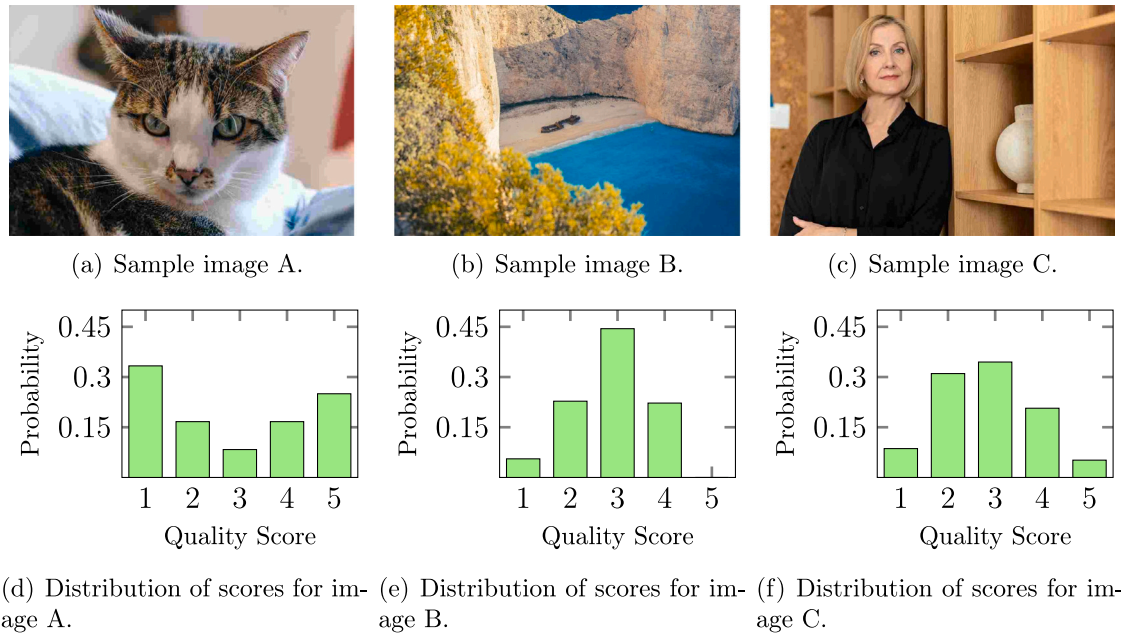


Fig. 1. Three image samples from the NAP540 [8] dataset, all with equal MOS (2.83). From the corresponding distribution histograms, it can be observed that the MOS does not capture the shape of the distribution nor variability in observers' ratings. This illustrates the need for predicting full subjective score distributions rather than single mean values.

metrics. (iii) We investigate whether employing a quality-aware backbone (Koncept512 [9]) enhances the cross-dataset generalization. (iv) We analyze the impact of different loss function choice, and propose a novel KL-Huber loss function for improved performance. (v) The performance of the proposed architectures is further validated on three additional datasets (CIDGD [11], VSR [12], and CIDOA [13]) on which the models were not trained.

The training and evaluation code, together with the best performing model, is publicly available on the following link: <https://github.com/plavacn/Predicting-Image-Quality-Score-Distribution-using-Deep-Neural-Networks>.

The remainder of the paper is structured as follows. A comprehensive overview of the previously proposed approaches is provided in Section 2. The methodology of this study is detailed in Section 3, followed by the experimental results in Section 4. Finally, a conclusion of the work and possible future directions are discussed in Section 5.

2. Related works

Putting emphasis on the importance of predicting the distribution of subjective image quality scores, several approaches have been proposed to go beyond the traditional prediction of MOS in IQA. NIMA [3] was among the first to propose such an architecture for predicting aesthetic quality scores. In their approach, they fine-tuned three CNN architectures (VGG16 [14], Inception-v2 [15], and MobileNet [16]) followed by a fully connected layer and a softmax layer that outputs values for ten normalized histogram bins. In their study, the Earth Mover's Distance (EMD) [17] was used as the loss function. The input images were resized to 256×256 pixels, and horizontal flipping was used as an augmentation technique. To evaluate the method, ideally, the quality score distributions obtained should be compared to the ground-truth distributions of subjective ratings. However, most subjective datasets used in the field do not contain ground-truth score distributions. This is why the performance assessment in NIMA [3] was based on a comparison of the MOS and the standard deviation of the ratings, assuming the ratings are Gaussian distributed. This assumption oversimplifies the often skewed or multi-modal nature of subjective quality score distributions.

DeepRN [4] introduced a resolution-independent model to predict the distribution of ratings using ResNet101 [18] as a backbone and spatial pyramid pooling. The network was trained on full-resolution images from the KonIQ-10k dataset [9] and evaluated using the ground-truth score distributions, represented as five-bin vectors of normalized histograms. Two loss functions were tested in the experiment, Euclidean distance and Huber loss, with the latter achieving better performance. To increase the size of the training set, various augmentation techniques were employed, some of which may have affected perceived image quality and influenced model training.

A closely related line of work was proposed by Tiotsoy et al. [19], who introduced CNN-based "artificial intelligence-based observers" to model individual observers in IQA. Their approach predicts a five-bin probability distribution over absolute category rating scores for each observer and reconstructs the MOS as the expectation over the predicted distribution. Although this work also acknowledges the importance of opinion score distributions, its primary goal is to simulate individual observers through separate subject-specific networks trained on single-rating labels, rather than directly learning population-level opinion score histograms.

Gao et al. [7] proposed to model the distribution of quality scores using an alpha stable distribution. Based on this model, they presented an objective metric that utilizes quality features such as structural and statistical information. In this approach, score distributions were parameterized using continuous functions, rather than fixed-bin histograms. A Support Vector Regressor (SVR) was then trained to predict the alpha stable model parameters. It was shown that using an alpha stable model can well reflect subjective quality score distribution. Properties such as skewness, heavy tail, and spikes present in the quality distributions can be accounted for using such a model. The model consists of four parameters that were regressed (estimated) using a SVR for each parameter. To obtain ground-truth quality distributions and evaluate the model, the authors re-collected subjective data for the LIVE image quality dataset [20] (and named it SJTU IQSD dataset). Quality ratings were given on a scale of 0 to 100 and binned in 10 bins with equal intervals on the quality scale. In another study [5], Gao et al. proposed a three-stage architecture to predict MOS from the distribution of quality scores. The proposed method first used a CNN architecture with ResNet50 [18] as the backbone to extract features from

images. The extracted features are then mapped to the image quality distribution using a graph convolutional network. In the third stage, the MOS value is extracted from the predicted image quality distribution. The proposed architecture was tested on two subjective datasets that contain the distribution of subjective scores (the SJTU IQSD [7] and KonIQ-10k [9] datasets). Additionally, the method was tested on the CSIQ [21], LIVE MD [22], LIVE Challenge [23], CID2013 [10], SPAQ [24], and VCLFER [25] datasets, where they assumed that the ratings distributions follow a Gaussian distribution. From the reported performance, the suggested architecture was superior to the state-of-the-art methods. Both studies [5,7] have some shortcomings. In the case of the first study [7], the model is based on handcrafted image features (statistical and structural features) and on a classical machine learning model (SVR), limiting its scalability. The second study [5] relies on a set of assumptions that may be somewhat strong or idealized. For instance, it assumes that the method can be appropriately evaluated by approximating the distributions as Gaussian using MOS and standard deviations, and it makes specific assumptions regarding the relationships between labels.

Other studies in this field incorporated fuzzy theory [6,26]. In [6], a pre-trained VGG16 network was used as a backbone to extract the features of the image. The features are then fuzzified using fuzzy theory. Finally, a fuzzy transfer network was proposed to predict the distribution of the quality scores. The loss function combined EMD and a novel quantile loss [6]. The method was trained and evaluated using the SJTU IQSD and KonIQ-10k datasets, and promising results were reported. Recently, this study has been extended [26], with the feature extraction network combining VGG16 and S-CNN [27]. Additionally, the method was tested on a wider variety of IQA datasets. Similar to [5], the ground-truth quality distributions are estimated by assuming a Gaussian distribution. Both [6,26] compared the proposed methods with the performance of transfer learning with the ResNet50, Inception-v2, MobileNet, and VGG backbones, and with altering Cross Entropy (CE) and EMD as loss functions. Although these methods demonstrated good results, they introduced significant model complexity.

Recently, Vision Transformers have emerged as a powerful alternative to CNNs for IQA tasks. The Image Quality Transformer [28] successfully adapted transformer architecture for full-reference IQA, using learnable quality and position embeddings to compare reference and distorted images. When it comes to blind IQA, Shi et al. [29] proposed a transformer-based NR-IQA model that combines CNN inductive bias with Transformer long-term dependency modeling through multi-stream blocks. This method achieved state-of-the-art performance by first training on large-scale synthetic datasets using supervised contrastive learning, and then fine-tuning for quality prediction. Another blind IQA architecture proposed was a local distortion-aware efficient Transformer adaptation architecture [30]. It utilized CNNs for extracting multi-scale distortion features and input them into Vision Transformer. A new direction in IQA has emerged with the integration of Multimodal Large Language Models (MLLMs). DepictQA [31] introduced a novel approach that provides a detailed, language-based evaluation of image quality. Unlike conventional IQA methods that output numerical scores, this work leverages MLLMs to interpret the image content and distortions descriptively and comparatively, closely aligning with human reasoning. Although CNN-based approaches could be seen as inferior to the Transformer- and MLLM-based models, there are several reasons why we selected CNN-based architectures for our study. Traditional CNN architectures require much less computational power compared to Transformer architectures. They also have well established training and optimization procedures and do not require very large training datasets. To the best of our knowledge, there are no existing datasets that have ground-truth score distributions and are larger than 10k images. In addition, training Transformer-based or multimodal architectures for distribution prediction poses additional challenges, as such models typically require substantially larger annotated datasets to avoid overfitting and to stabilize the learning of

multi-bin probability outputs under distribution-aware loss functions. The limited availability of large-scale datasets with reliable ground-truth opinion score histograms makes this particularly problematic at present. As data availability increases, future extensions could explore hybridization with Transformer or MLLM models, but at present CNNs offer the best balance of accuracy, efficiency, and practicality for distribution-aware IQA.

3. Methodology

This section describes the methodological framework employed to predict the distributions of image quality scores. The description of the datasets and a comparative analysis of their quality score distributions are given in Section 3.1. The deep learning architectures selected for this task are introduced in Section 3.2. The details of the implementation and an overview of the loss functions are given in Section 3.3. Finally, the evaluation protocol and the metrics used to quantify the prediction performance are outlined in Section 3.4.

3.1. Datasets

To ensure diversity of content, systematic training, and evaluation of different architectures, three IQA datasets were used: KonIQ-10k [9], CID2013 [10], and the NAP540 [8] dataset collected at our research laboratories. The KonIQ-10k dataset [9] is one of the largest IQA datasets, containing 10073 images. It contains mostly natural but also computer-generated images with a significant variety of content (e.g. nature, buildings, people, plants, animals, products). The observers rated the quality of the images on a scale from one to five. The Camera Image Database 2013 (CID2013) [10] consists of 480 images captured using 79 different cameras. The images depict eight scenes, each captured by 12–14 cameras. On average, 30 observers evaluated the quality of each image on a scale from zero to 100. To have a consistent quality rating scale throughout our work (i.e., five-bin histograms), the ratings in this work were scaled to a one-to-five scale by applying a simple binning of equal intervals. The NAP540 dataset [8], consists of 5400 images and 188760 individual observers' ratings. Of the 5400 images, 540 images are reference images (pristine quality images) covering three different content categories: animals (Fig. 1(a)), nature (Fig. 1(b)), and people (Fig. 1(c)). Each of the categories contained 180 images. Three types of distortions (blur, JPEG compression artifacts, and contrast) were applied to the images, each at three levels. Observers rated the quality of images on a scale from one to five. For each image, all the observers' scores were aggregated into a 5-bin histogram that was then normalized to represent the probability distribution of the quality scores.

The differences in underlying score distributions across datasets can impose difficulties for cross-dataset evaluation. To examine the extent of these differences, we performed visual and statistical analysis of the score distribution present in the datasets. The KonIQ-10k dataset exhibits a unimodal, skewed distribution (Fig. 3(a)), with most images receiving quality scores of three, followed by four. It is important to point out that extreme values are underrepresented in the case of this dataset (Fig. 3(a)). In contrast, CID2013 demonstrates a comparatively uniform distribution across the five quality score bins, with none of the bins being severely underrepresented (Fig. 2(b)). Greater variation in per-image ratings across all quality levels can be observed (Fig. 3(b)), especially at score five, reflecting higher observer disagreement in this dataset. The NAP540 dataset lies between the two, with an average histogram approximating a Gaussian distribution, though skewed toward lower quality scores, with score two being most frequently selected by observers (Fig. 2(c)). Furthermore, variability of the scores is relatively uniform (Fig. 3(c)). The empirical mean (μ) and standard deviation (σ) of bin-wise probabilities across the datasets are summarized in Table 1. MANOVA analysis confirmed the significant differences ($p < 0.01$) between the three datasets. The differences in the underlying

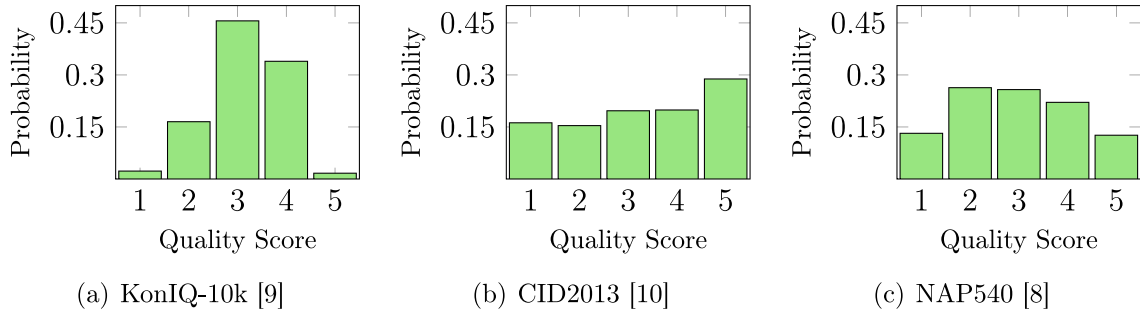


Fig. 2. Average normalized histograms of the quality scores for the datasets.

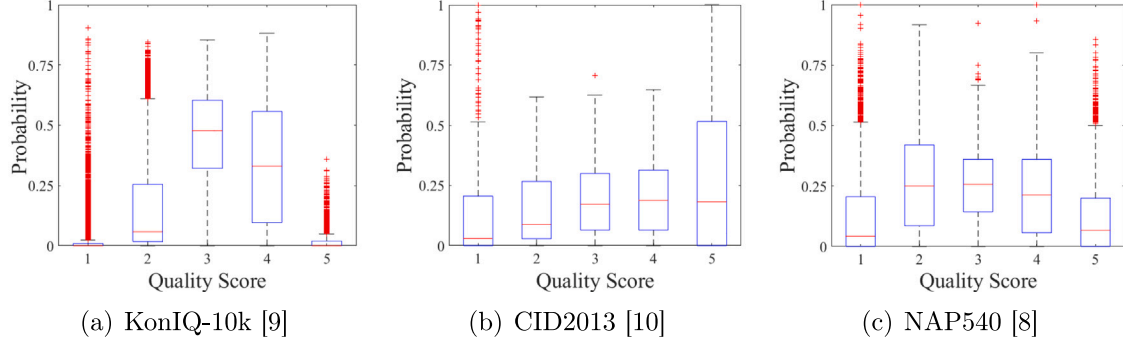


Fig. 3. Boxplots showing quality score distributions for the datasets.

Table 1

Empirical mean (μ) and empirical standard deviation (σ) for KonIQ-10k [9], CID2013 [10], and NAP540 [8] datasets.

Quality score	KonIQ-10k [9]		CID2013 [10]		NAP540 [8]	
	μ	σ	μ	σ	μ	σ
1	0.023	0.077	0.162	0.264	0.132	0.183
2	0.165	0.210	0.154	0.158	0.263	0.193
3	0.456	0.182	0.197	0.151	0.258	0.144
4	0.339	0.246	0.199	0.149	0.221	0.172
5	0.017	0.030	0.288	0.301	0.126	0.152
Average		0.149		0.205		0.169

distributions are not only visually apparent and statistically significant, but also practically important. For instance, if one dataset concentrates ratings around score three while another around score five, the trained models may be biased toward the dominant patterns of the training dataset and fail to generalize well across datasets.

In addition to the three main datasets used for training and performance evaluation, three smaller and more specialized datasets were used to test the models to gain additional insight into their performances. The Colourlab Image Database: Geometric Distortions (CIDGD) [11] was specifically designed to evaluate the robustness of image quality metrics against geometric transformations. It consists of 441 images: 49 reference images, each distorted by two levels of seam carving, four levels of lens distortion, and two levels of image rotation. The Variability in Subjective Ratings dataset (VSR) [12] includes 10 reference images degraded with 21 distortions at three levels each (640 images in total). It was designed to investigate which distortion causes the greatest disagreement among individuals. Finally, the Colourlab Image Database: Optical Aberrations (CIDA) [13] consists of 23 reference images distorted by a set of optical distortions, namely defocus, astigmatism, and spherical aberration. Each distortion was applied in four levels, making the total number of images in the dataset 299. These datasets, though smaller in scale, provide complementary perspectives that enrich the evaluation of the trained models.

3.2. Architectures

The task of predicting the distribution of subjective quality scores given to images requires robust and efficient models capable of generating multiclass regression output. A pilot study (Section 4.1) was conducted to select optimal network architectures from a group of ten candidate networks, based on their predictive performance and parameter efficiency. Based on the results of the pilot study, we propose IQDN in five configurations: a custom CNN architecture and four transfer learning-based architectures (ResNet50 [18], VGG16 [14], Xception [32], and DenseNet121 [33] backbones).

The basic IQDN, consisting of the custom CNN model (Fig. 4), was trained from scratch allowing us to experiment with different model parameters and gain general insight into the performance of the model. The input to the network is an RGB image with spatial dimensions of 384×512 pixels. This ensures consistency in input image size and allows efficient and unified computation in subsequent layers. The model comprises of four convolutional layers, max pooling, a Global Average Pooling (GAP) layer, dense layers, and dropout regularization. The convolutional layers serve as the primary feature extractors, progressively learning spatial hierarchies within the input images, from low-level edges to high-level patterns. Each layer uses a 3×3 kernel, ReLU activation, and increasing filter size from 32 to 256. Max-pooling operations follow each convolutional block, reducing the spatial dimensions of the feature maps by half and retaining the most significant values within each pooling window. By focusing on the most salient features, max-pooling ensures that the extracted feature maps are compact and meaningful. The GAP layer further reduces the spatial feature maps to a single vector, significantly reducing the number of parameters in the model. By summarizing the information from each feature map, the GAP layer provides a global representation of the learned features. A fully connected layer with 256 units and a dropout rate of 0.5 is applied to prevent overfitting by randomly deactivating half of the neurons in this layer during training. The final layer of the network consists of five neurons with a Softmax activation function. The Softmax function normalizes the outputs to represent a probability

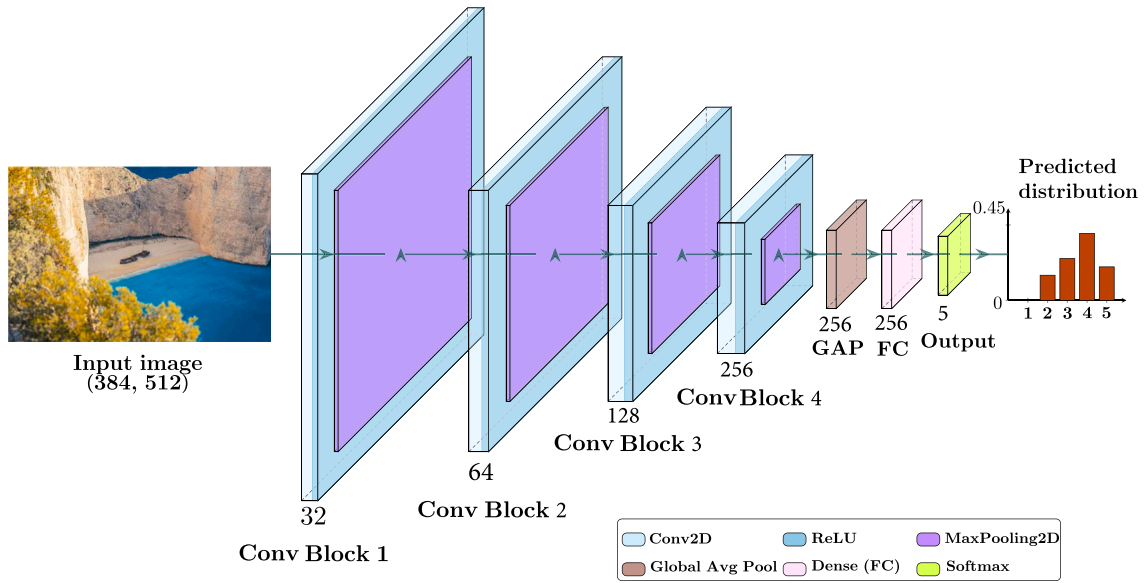


Fig. 4. Overview of the basic IQDN architecture. The input image passes through four convolutional blocks, each consisting of convolution and max-pooling operations that progressively extract higher-level features while reducing spatial resolution. A Global Average Pooling (GAP) layer reduces the final feature maps to a compact vector representation, which is followed by a Fully Connected (FC) layer with 256 units and dropout regularization. The output layer applies a Softmax activation to produce a predicted probability distribution of the image quality ratings.

distribution, with each value corresponding to the probability of one of the five quality scores.

Transfer learning allows models to use features learned from large-scale datasets such as ImageNet [34], reducing the need for extensive training on the new dataset. Four IQDN architectures were implemented using ResNet50 [18], VGG16 [14], Xception [32], and DenseNet121 [33] as backbones, all trained on millions of images from the ImageNet dataset. In each case, the pre-trained backbone served as a feature extractor, followed by a custom classification head tailored for the quality score prediction task. The four backbone networks are increasing in computational complexity, with DenseNet121 being the simplest with around 8.1M parameters, Xception with around 22.9M, ResNet50 with around 25.6M, and VGG16 being the most complex with around 138.4M parameters. In contrast, the basic IQDN network (Fig. 4) consists of around 455 K parameters. To our knowledge, the Xception and DenseNet121 backbones were not previously used for this kind of task, while the other two backbones were used in previous studies [3,5,6,26].

During the initial training phase of 15 epochs, the backbones were used with frozen pre-trained layers. This allowed the backbones to retain their general visual features. The regression head consisted of a GAP layer, a 128-neuron dense layer, a dropout layer with a rate of 0.5, and a softmax layer for the five output scores. In the second phase, fine-tuning was performed for an additional 30 epochs by unfreezing the last 10 layers of the backbone, allowing the model to adapt to the specific dataset. The fine-tuning strategy for the Xception and VGG16 backbone differed only in the second phase, where unfreezing was applied to the last 20 and last five layers, respectively.

3.3. Training and implementation details

All IQDN models were trained using the same dataset split and data augmentation pipeline to ensure fair comparisons. Identical early stopping and learning rate reduction were employed to prevent overfitting and optimize performance. For training, 70% of the images were randomly selected, leaving 10% for validation and 20% for testing. Although the datasets used do not have an official train/test split, these proportions match those used in other works (e.g., [3,4,6]), making it relatively possible to compare performance. We should point out that

this is not an optimal way of comparison, as the split of the dataset is random. However, this would be the only viable option to compare the proposed methods.

The primary loss function used in this work is the Kullback–Leibler (KL) divergence. To the best of our knowledge, it has not been previously used for this kind of task. The KL divergence was chosen because the model output and ground-truth are represented as a normalized probability distribution over discrete quality score bins. Unlike point-wise error metrics such as Mean Absolute Error (MAE) or Root Mean Square Error (RMSE), which treat each bin independently, KL divergence measures the overall dissimilarity between two probability distributions, capturing differences in both magnitude and allocation of probability mass. One could point out that the KL divergence is an asymmetric metric. The asymmetry is advantageous in our setting because it will penalize cases where the predicted distribution assigns insufficient probability mass to high-likelihood quality bins. This aligns with the practical goal of accurately capturing the most probable quality scores, which dominate the perceptual interpretation of image quality. Conversely, overestimation of low-probability bins is penalized less, reflecting a tolerance for minor spread in predicted distributions when the primary peak is preserved. Additionally, an experiment was performed combining the KL divergence with Huber loss (Section 4.6). Huber loss behaves quadratically for small errors and linearly for large errors, balancing sensitivity to outliers with robustness to minor noise. It is symmetric, penalizing over and under predictions equally and focusing on the magnitude of bin-wise differences regardless of their relative probability. Combining both metrics in a unique loss function allows the model to capture the global shape of the distribution (KL divergence), while maintaining stable performance across all bins and avoiding overfitting to the dominant peak.

The following optimization parameters were chosen empirically, based on the result of a pilot study. The batch size was set to 32, the Adam optimizer was used, and the learning rate was set to 10^{-3} and 10^{-5} in the first and second phases of fine-tuning, respectively. Early stopping was used with patience of seven epochs, together with reduction of the learning rate on the plateau with a factor of 0.2 and patience of three epochs. All implementations were carried out in Keras with the TensorFlow backend.

3.4. Performance evaluation metrics

We employ a wide range of evaluation metrics that capture different aspects of distribution similarity and prediction error. These include KL divergence [35], MAE, RMSE, EMD [17], Root Cumulative Goodness of Fit Coefficient (RCGFC) [36], Chebyshev distance (CD), mean MOS prediction error (ΔMOS), Pearson Linear Correlation Coefficient (PLCC) and Spearman Ranked Correlation Coefficient (SRCC). As discussed in Section 3.3, KL divergence is chosen as the primary loss function during training, as it measures the dissimilarity between two probability distributions. MAE and RMSE are standard regression metrics that measure the average magnitude of errors between predicted and ground-truth scores without considering their direction (i.e. the shape of the distribution). RMSE penalizes larger errors more heavily than smaller ones, making it useful to identify cases where the model may struggle with outlier predictions. EMD computes the minimum effort required to transform one probability distribution into another, interpreted as the “cost” of moving the probability mass. It captures not only individual score differences, but also the structural similarity between distributions, making it particularly relevant for tasks where the overall shape and alignment of distributions matter. RCGFC is a metric designed to evaluate the similarity between two spectral vectors. CD measures the maximum absolute difference between the predicted and ground-truth scores in all dimensions. Unlike MAE and RMSE, which provide an average error magnitude, CD highlights the largest individual discrepancy, helping to identify the model’s worst-case performance. From the obtained quality score distribution predictions, the predicted MOS could be calculated by multiplying each quality score with its probability, and summing the results over all the bins. This allows for investigating how the proposed architectures perform in predicting the MOS compared to IQA metrics designed for that task. Such performance is usually measured by calculating the (ΔMOS), PLCC, and SRCC between the ground-truth and predicted MOS values. These metrics provide a comprehensive evaluation framework that covers both global alignment and point-wise error.

4. Results and discussion

This Section contains the details of the pilot study that was performed to select parameters and suitable architectures. After the pilot study is introduced, the performance of the selected architectures is presented, analyzed, and discussed. The performance of the basic IQDN model will be denoted with $IQDN$, while the transfer learning versions will contain the backbone name in the subscript (i.e., $IQDN_{ResNet}$, $IQDN_{VGG}$, $IQDN_{Xception}$, and $IQDN_{DenseNet}$ for the ResNet50, VGG16, Xception, and DenseNet121 backbones, respectively). Ideally, the proposed architectures should be benchmarked with similar approaches published in the field. However, a direct quantitative comparison with them can be quite challenging. This is due to the fact that many existing approaches rely on parametric assumptions on the shape of the score distributions, were evaluated on proprietary or re-collected datasets, and do not provide publicly available implementations, which prevents reproducible benchmarking. In cases where another method was evaluated on the same dataset, we report the performance that was reported in the original publication, while retaining the same train/test split proportions.

4.1. Pilot study: Architecture selection

Before conducting the main experiments, a pilot study was performed to identify suitable network architectures and training configurations to predict the distribution of quality scores. The study consisted of two stages: optimization of a custom CNN architecture trained from scratch and backbone selection among widely used deep learning architectures. Both stages were performed on the NAP540 dataset, as

it represents an intermediate case between the other two datasets both in terms of the size and the overall distribution shape.

In the case of the custom CNN architecture, the effect of network depth, batch size, and number of training epochs was explored. The results of using different numbers of convolutional blocks were evaluated. Increasing the depth beyond four convolutional blocks (e.g., five, six, seven blocks) did not yield performance gains, and therefore four blocks were selected as optimal. Although in some related works (e.g., [4]) a larger batch size is used, our experiments showed that smaller batches improved network performance. A batch size of 32 was chosen as the best trade-off between stability and accuracy. The number of epochs was initially set to 50, but validation the loss curves indicated overfitting. Early stopping was introduced and the nominal maximum was reduced to 30 epochs, with most runs converging within 20–25 epochs.

With the training parameters configuration fixed in the first stage, different backbone architectures were evaluated using transfer learning. The tested models included: VGG16, VGG19, ResNet50, ResNet101, EfficientNetB0, MobileNetV2, MobileNetV3Large, DenseNet121, DenseNet201, InceptionV3, and Xception. This set was selected to cover a broad range of network complexities in terms of the number of learnable parameters ($\#P$), allowing a fair comparison between lightweight and computationally intensive network designs. The fine-tuning was done in two stages, with the first 15 epochs keeping the whole backbone frozen. After that, the last 10 layers of each backbone were unfrozen for training for additional 30 epochs. For deeper architectures such as ResNet101, DenseNet201, InceptionV3, and Xception, the last 20 layers were unfrozen to allow greater adaptation to the task while maintaining computational feasibility.

Table 2 summarizes the performance of each backbone in terms of KL divergence, MAE, RMSE, EMD, CD, RCGFC, and $\#P$. It is important to note that $\#P$ refers to the number of learnable parameters in the second stage of fine-tuning. The results show that VGG16 achieved the best overall performance across all evaluation metrics, while being the backbone with the highest computational complexity among the tested models. It was selected for the further experiments because of its performance in the pilot study, as well as because it was previously used in prior IQA works (e.g., [6,26]). Xception backbone was the second best performing network, while having a middle-range number of trainable parameters (7.9M), providing a good balance between accuracy and computational cost. Compared to its more complex variant DenseNet201, DenseNet121 offered relatively comparable performance with a lowest count of trainable parameters (only 0.4M), making it the preferred lightweight option. Finally, ResNet50 was also selected for its moderate complexity (5M parameters) and its frequent use in prior works on IQA (e.g., [6]). These backbones were selected for further experiments (listed as Architectures in Section 3.2), ensuring diversity in model complexity while maintaining high predictive performance, and providing a strong experimental basis for the subsequent architecture-specific analyses in the main experiments.

4.2. Performance on the KonIQ-10k dataset

The performance of the IQDN architectures is displayed together with results reported in related studies (Table 3). From [4], several shared performance metrics are reported for the proposed DeepRN architecture using both the Euclidean (E) and Huber (H) loss functions. Gao et al. [5] presented performance comparisons between their GCN network, the NIMA network [3], and their earlier alpha-stable model [7]. In [6], a detailed performance comparison was performed between several backbone networks (ResNet50, IncepResNetV2, MobileNet, and VGG16) trained with Cross Entropy (CE) and EMD losses, in addition to the proposed fuzzy logic-based model.

Among the five proposed IQDN-based architectures, $IQDN_{Xception}$ achieved the best overall performance on the KonIQ-10k dataset, particularly excelling in KL divergence (0.120), EMD (0.043), and RCGFC

Table 2

Performance metrics for different backbone architectures on the NAP540 dataset, sorted by KL divergence (lower is better). The best performance per metric is shown in **bold**.

Backbone	KL	MAE	RMSE	EMD	CD	RCGFC	#P
VGG16	0.110	0.064	0.089	0.054	0.524	0.133	13.6M
VGG19	0.118	0.066	0.092	0.057	0.555	0.139	17.8M
Xception	0.119	0.064	0.090	0.055	0.521	0.134	7.9M
DenseNet201	0.121	0.065	0.092	0.056	0.593	0.137	1.1M
DenseNet121	0.127	0.068	0.095	0.060	0.549	0.144	0.4M
InceptionV3	0.128	0.069	0.096	0.060	0.564	0.147	2.4M
ResNet101	0.243	0.098	0.132	0.100	0.678	0.221	9.4M
ResNet50	0.260	0.107	0.136	0.112	0.677	0.234	5.0M
MobileNetV3Large	0.356	0.127	0.159	0.143	0.676	0.288	0.8M
EfficientNetB0	0.402	0.140	0.168	0.162	0.699	0.316	1.2M
MobileNetV2	0.651	0.141	0.183	0.153	0.741	0.338	1.0M

Table 3

Performance metrics for different architectures on the KonIQ-10k dataset (best performance per metric is shown in bold).

Architecture	KL	MAE	RMSE	EMD	CD	RCGFC
DeepRN(E) ^a	0.130	0.064	0.081	0.172	–	–
DeepRN(H) ^a	0.124	0.053	0.069	0.169	–	–
NIMA ^b	–	–	0.109	0.084	–	–
Gao ^b	–	–	0.072	0.060	–	–
GCN ^b	–	–	0.069	0.058	–	–
ResNet50(CE) ^c	–	–	0.185	0.130	–	–
ResNet50(EMD) ^c	–	–	0.093	0.079	–	–
IncepNetV2(CE) ^c	–	–	0.186	0.129	–	–
IncepNetV2(EMD) ^c	–	–	0.108	0.094	–	–
MobileNet(CE) ^c	–	–	0.249	0.175	–	–
MobileNet(EMD) ^c	–	–	0.107	0.084	–	–
VGG16(CE) ^c	–	–	0.201	0.141	–	–
VGG16(EMD) ^c	–	–	0.093	0.080	–	–
Fuzzy ^c	–	–	0.072	0.060	–	–
<i>IQDN</i>	0.157	0.077	0.120	0.157	0.617	0.177
<i>IQDN_{ResNet}</i>	0.200	0.088	0.135	0.066	0.755	0.205
<i>IQDN_{VGG}</i>	0.137	0.069	0.112	0.049	0.717	0.164
<i>IQDN_{Xcept}</i>	0.120	0.062	0.103	0.043	0.649	0.146
<i>IQDN_{DenNet}</i>	0.139	0.068	0.110	0.048	0.711	0.159

^a The reported performance in [4].

^b The reported performance in [5].

^c The reported performance in [6].

(0.146), indicating that this architecture effectively captures the full quality score distribution while minimizing divergence from ground-truth ratings. *IQDN_{DenNet}* followed closely, producing the second-best scores for EMD (0.048) and RCGFC (0.159), while also producing strong MAE and RMSE results, positioning it as a lightweight but effective alternative. The second-best KL divergence score (0.137) was obtained for *IQDN_{VGG}*. The basic *IQDN* architecture performed well overall, notably achieving the lowest CD value (0.617), which reflects a minimal maximum deviation between the predicted and true distributions. This indicates that although most predictions diverged from ground-truth, the single worst-case discrepancy was smaller than those for the other four *IQDN* architectures. The *IQDN_{ResNet}* model, although an often used backbone in IQA, underperformed compared to the other variants in KL divergence and CD. However, it did produce a relatively strong EMD score (0.066), suggesting a good alignment in the shape of the distributions despite higher mean error magnitudes. Compared to related works, the proposed architectures demonstrated strong results in terms of distributional alignment. In particular, while the RMSE values of most proposed architectures remain higher than those reported in [4–6], the EMD scores of *IQDN_{Xcept}*, *IQDN_{DenNet}*, and *IQDN_{VGG}* show better performance than previously reported models, including those that use the EMD loss function. Since EMD captures how well the full predicted distribution aligns with human-perceived quality distribution, this supports the claim that our models capture the subjective uncertainty more faithfully than state-of-the-art

Table 4

Performance metrics for different architectures on the CID2013 dataset (best performance per metric is shown in bold).

Architecture	KL	MAE	RMSE	EMD	CD	RCGFC
<i>IQDN</i>	0.592	0.171	0.224	0.592	0.798	0.358
<i>IQDN_{ResNet}</i>	0.517	0.162	0.207	0.200	0.733	0.284
<i>IQDN_{VGG}</i>	0.374	0.127	0.176	0.143	0.881	0.243
<i>IQDN_{Xcept}</i>	0.393	0.126	0.178	0.141	0.841	0.222
<i>IQDN_{DenNet}</i>	0.334	0.118	0.162	0.125	0.697	0.206

models. These results highlight that the proposed models, especially *IQDN_{Xcept}*, better account for the variability of the perceptual score, even if the mean score prediction (reflected in the RMSE) is less precise than in some baseline models. This trade-off is justified in applications where understanding perceptual ambiguity and rating spread is critical.

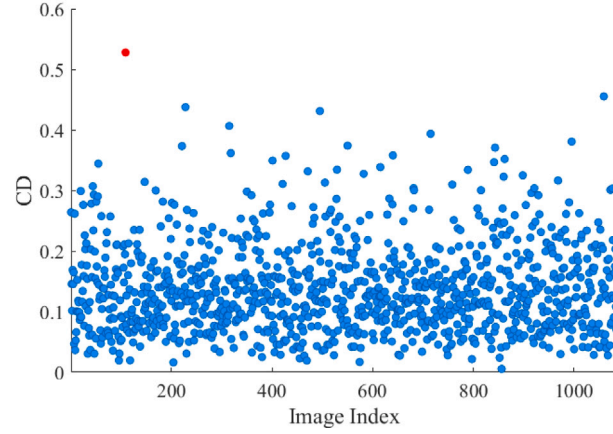
4.3. Performance on the CID2013 dataset

None of the previous works in the literature evaluated their models on the ground-truth distribution data of the CID2013 dataset, making this comparison novel. As shown in Table 4, the *IQDN_{DenNet}* architecture achieved the best overall performance, achieving the best scores on all metrics. This shows that the lightweight DenseNet-based model was more effective than the other variants in aligning the predicted distributions with the ground-truth, despite the rather small size of the dataset. *IQDN_{Xcept}* and *IQDN_{VGG}* also performed competitively, with *IQDN_{Xcept}* performing slightly better *IQDN_{VGG}* in terms of MAE (0.126 vs. 0.127), EMD (0.141 vs. 0.143), CD (0.841 vs. 0.881), and RCGFC (0.222 vs. 0.243) while *IQDN_{VGG}* achieved lower KL divergence (0.374 vs. 0.393) and RMSE (0.176 vs. 0.178), indicating comparable performance between the two. The *IQDN_{ResNet}* and the basic *IQDN* model lagged behind, suggesting poor alignment with the human rating distributions. The results on CID2013 were notably weaker than those obtained on the other two datasets. This could be attributed to multiple factors. First, the dataset size is considerably smaller, which can limit the model's ability to generalize. Second, the types and combinations of distortions in CID2013 differ from those in more recent datasets, possibly introducing a mismatch between the training and test distributions. Finally, CID2013 exhibits distinct patterns in subjective score distributions (as seen in Section 3.1), which may further complicate learning. Despite these challenges, the proposed *IQDN_{DenNet}* still demonstrated encouraging results, showing that relatively effective learning of perceptual distributions is achievable even under dataset constraints.

4.4. Performance on the NAP540 dataset

As shown in Table 5, *IQDN_{Xcept}* achieved the best overall performance across the majority of the metrics, including MAE (0.064), RMSE (0.091), EMD (0.055), and RCGFC (0.136). It also performed competitively in terms of KL divergence (0.121), only slightly worse than *IQDN_{VGG}*, which recorded the lowest KL value (0.117). *IQDN_{VGG}* performed particularly well, achieving the top two performance for all metrics and performing better than the other variants by a considerable margin in KL divergence and CD. The simple *IQDN* model demonstrated decent performance, especially when considering its simplicity. *IQDN_{DenNet}* slightly underperformed compared to its strong performance on CID2013. On the other hand, *IQDN_{ResNet}* achieved significantly higher error values across all metrics, especially KL divergence (0.302), EMD (0.124), and RCGFC (0.256), indicating that it struggled to align predicted distributions effectively on this dataset.

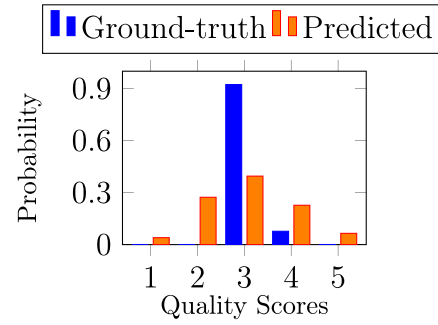
It should be noted that the performance of the networks in terms of CD was very dataset-dependent. As CD captures the maximum absolute deviation between the predicted and ground-truth probability bins, it can identify which architectures avoid large singular deviations, even



(a) The CD for the example image shown with red dot.



(b) The image with the highest CD value.



(c) The predicted and the ground-truth distribution for the example image.

Fig. 5. The obtained CD values for $IQDN_{VGG}$ on the NAP540 dataset.**Table 5**

Performance metrics for different architectures on the NAP540 dataset (best performance per metric is shown in bold).

Architecture	KL	MAE	RMSE	EMD	CD	RCGFC
$IQDN$	0.157	0.078	0.110	0.157	0.696	0.163
$IQDN_{ResNet}$	0.302	0.115	0.148	0.124	0.714	0.256
$IQDN_{VGG}$	0.117	0.066	0.095	0.057	0.528	0.138
$IQDN_{Xcept}$	0.121	0.064	0.091	0.055	0.549	0.136
$IQDN_{DenNet}$	0.189	0.082	0.113	0.080	0.560	0.185

if their overall distribution match may be weaker. To further explore which portion of the test images had high CD values, we examined the per-image CD values for different architectures. The example for $IQDN_{VGG}$ architecture on the NAP540 dataset is shown in Fig. 5. In this case, the mean CD value was 0.138 with a standard deviation of 0.073. The images for which the CD value is outside the three standard deviation range can be considered as outliers (here the threshold would correspond to 0.356). According to this criteria, 13 outliers have been detected. One of the outliers, the image with the highest CD value is marked in red (Fig. 5(a)), corresponding to a JPEG-distorted image (Fig. 5(b)) for which the prediction strongly deviated from the ground-truth (Fig. 5(c)).

Together, these results demonstrate the robustness and consistency of the $IQDN_{Xcept}$ and $IQDN_{VGG}$ architectures across datasets. $IQDN_{DenNet}$ also demonstrated solid performance, especially excelling in the case of the CID2013 dataset, offering a favorable trade-off between accuracy and computational efficiency due to its lightweight design.

4.5. Fine-tuning the quality-aware backbone

To assess the impact of using a quality-aware feature extractor, the pre-trained Koncept512 network from the KonIQ-10k paper [9] was used as a backbone. Consistent with the workflow described in Section 3.2, the custom regression head was appended to this architecture and the two-stage fine-tuning was applied. The results for all three datasets are summarized in Table 6. It is important to emphasize that since the Koncept512 model was originally trained on the KonIQ-10k dataset, its performance on this dataset may be biased or overoptimistic due to potential overfitting to the data distribution. Therefore, its performance on KonIQ-10k should be interpreted with caution and not used for direct comparisons. Nevertheless, the results on the other two datasets (CID2013 and NAP540) offer a more meaningful assessment of the generalization capability of this quality-aware backbone.

In the case of the CID2013 dataset, the Koncept512 backbone performs better than the basic $IQDN$ and $IQDN_{ResNet}$ architectures in all metrics. Compared to the best performing variant $IQDN_{DenNet}$, much lower KL divergence (0.289) and CD (0.518) values were obtained, confirming the benefit of using quality-aware feature representations. In terms of MAE, RMSE, EMD, and RCGFC, the quality-aware backbone achieved performance similar to, or slightly worse than, the best performing variant from Table 4. Similarly, for the NAP540 dataset, the fine-tuned Koncept512 backbone achieves highly competitive results, rivaling the top-performing configurations $IQDN_{Xcept}$ and $IQDN_{VGG}$ reported in Table 3. The obtained CD (0.489) is once again lower than those of the best performing architectures in Table 3, confirming that the use of quality-aware features can reduce the extreme prediction deviations. It demonstrates its potential value in settings where minimizing the maximum per-bin error is important. These results suggest

Table 6

Performance metrics obtained using the KonCept512 backbone on all three datasets.

Dataset	KL	MAE	RMSE	EMD	CD	RCGFC
KonIQ-10k	0.139	0.067	0.107	0.053	0.788	0.154
CID2013	0.289	0.122	0.174	0.126	0.518	0.209
NAP540	0.122	0.065	0.090	0.055	0.489	0.139

that using a pre-trained quality-aware backbone can enhance prediction performance, particularly in cross-dataset scenarios.

4.6. Combined loss experiment

To assess the effect of augmenting the training loss with a secondary regularization term, we conducted experiments combining the KL divergence with the Huber loss. The total loss function used during training was defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{KL}}(P \parallel Q) + \lambda \cdot \mathcal{L}_{\text{Huber}}(P, Q) \quad (1)$$

where P denotes the ground-truth quality score distribution, Q is the predicted distribution, $\mathcal{L}_{\text{KL}}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$, and $\mathcal{L}_{\text{Huber}}(P, Q)$ is computed as:

$$\mathcal{L}_{\text{Huber}}(P, Q) = \frac{1}{K} \sum_{i=1}^K \begin{cases} \frac{1}{2} (P(i) - Q(i))^2, & \text{if } |P(i) - Q(i)| \leq \delta \\ \delta \cdot (|P(i) - Q(i)| - \frac{1}{2} \delta), & \text{otherwise} \end{cases} \quad (2)$$

with $K = 5$ denoting the number of bins, and $\delta = 0.1$ as the transition point between quadratic and linear error regimes, following the recommendation in [4].

The weighting parameter λ controls the relative contribution of the distribution-level KL divergence and the bin-wise Huber term. We evaluated several values of the weighting parameter $\lambda \in \{0, 0.5, 1, 10, 50, 100\}$ on the validation set to examine its influence on prediction performance. Surprisingly, the choice of λ had a negligible effect on the results. Models trained with any non-zero λ consistently performed better than the KL-only baseline, but the difference in the results between different λ values was marginal. To investigate this, we analyzed the scale of the two individual loss components. A substantial scale mismatch has been observed, with the KL divergence component being up to 40× larger than the Huber component. For a typical training run ($IQDN_{\text{DenNet}}$ trained on KonIQ-10k), the average magnitude of $\mathcal{L}_{\text{KL}}$ was around 0.125, while the average magnitude of $\mathcal{L}_{\text{Huber}}$ was 0.003. Consequently, even for large values of λ , the Huber term contributes modestly to the overall gradient. This might explain why performance remains stable across a wide range of λ values, as the KL term continues to dominate optimization behavior. These results suggest that the addition of the Huber term functions more as a regularizer than a competing objective. Its presence stabilizes training and improves prediction performance (Table 7), even when its contribution is small in magnitude. Across the tested λ values, the standard deviation of the average EMD values obtained on the test set was 0.0005 for a typical training run ($IQDN_{\text{DenNet}}$ trained on KonIQ-10k). The reported standard deviation reflects the variation of performance across the tested λ values, rather than variability due to random initialization. This indicates that further performance gains do not require precise tuning of λ within the tested range. Rather, the benefit arises from simply including the Huber loss in the training objective.

When comparing the results of the KL-only baseline (Tables 3, 4, and 5) with the KL + Huber variant (Table 7, $\lambda = 1$), several important trends emerge across datasets and architectures. On the KonIQ-10k dataset, most architectures benefited from the combined loss, especially $IQDN_{\text{Xcept}}$, with substantial reductions in all metrics, including a drop

in KL divergence (from 0.120 to 0.087) and EMD (from 0.043 to 0.036), both representing the best scores observed in all experiments.

The performance of $IQDN_{\text{ResNet}}$, previously the lowest performing variant, improved greatly (KL divergence from 0.200 to 0.136, EMD from 0.066 to 0.047), closing much of the performance gap to stronger backbones. The only case where the combined loss did not lead to an increase in performance was $IQDN_{\text{VGG}}$. In this case a small but consistent increase in almost all error metrics was observed, suggesting that for architectures already well-optimized under KL divergence only, the Huber term may slightly perturb convergence. On the CID2013 dataset, the effect was more architecture-dependent. The strongest improvements came from $IQDN_{\text{Xcept}}$ (KL divergence from 0.393 to 0.253) and $IQDN$ (KL divergence from 0.592 to 0.364, EMD from 0.592 to 0.175). For $IQDN_{\text{DenNet}}$ and $IQDN_{\text{ResNet}}$, the KL divergence improved, but the other metrics were slightly reduced, suggesting that while the combined loss aligned global distribution shapes better, it did not necessarily reduce bin-wise absolute errors for this lightweight model. As in the case of KonIQ-10k, introducing the combined loss disrupted the fine-tuning dynamics for $IQDN_{\text{VGG}}$. The constrained size of this dataset might have played a crucial role in this rather unstable behavior across different architectures. For the NAP540 dataset, all architectures except $IQDN_{\text{VGG}}$ improved consistently in nearly all metrics. For the $IQDN_{\text{ResNet}}$ variant, a large drop in KL divergence (from 0.302 to 0.101) and EMD (from 0.124 to 0.051) moved this variant from the weakest to the second strongest in terms of KL divergence. $IQDN_{\text{Xcept}}$ retained its leading performance, with a further reduction in the KL divergence (from 0.121 to 0.095) and the EMD (from 0.055 to 0.047). The performance of $IQDN_{\text{VGG}}$ was again consistently worse compared to the KL-only loss variant. $IQDN$ and $IQDN_{\text{DenNet}}$ also experienced slight improvements.

To qualitatively illustrate the trade-off between distribution alignment and point-wise accuracy, Fig. 6 shows representative examples with their ground-truth and predicted score distributions. Although the predicted distribution in the first case (Fig. 6(b)) appears visually similar to the ground truth, a small but systematic shift of probability mass toward higher quality bins leads to a noticeable MOS error, resulting in a high RMSE despite relatively low EMD. This example demonstrates that good distribution alignment alone does not guarantee accurate mean opinion score prediction. In the second case (Fig. 6(d)), the predicted MOS is relatively close to the ground-truth, yielding a moderate RMSE, while the predicted distribution fails to capture the spread of observer ratings, resulting in a high EMD value. This highlights how acceptable point-wise accuracy can mask substantial mismatches in perceptual uncertainty. Finally, the third case (Fig. 6(f)) shows low values for both RMSE and EMD, where the predicted distribution closely matches the ground-truth in both shape and mean. Together, these examples emphasize that EMD and RMSE capture complementary aspects of prediction quality, and that relying on a single metric may obscure important failure modes in distribution-aware IQA.

In order to investigate how the proposed architectures perform in predicting the MOS compared to other IQA metrics designed for that task, the predicted MOS can be calculated from the obtained quality score distribution predictions. From the predicted MOS, the mean difference between the ground-truth and predicted MOS values ($|\Delta \text{MOS}|$) and correlation coefficients (PLCC and SRCC) can be calculated. The predicted MOS values for the best performing architecture from Table 7 are shown on Fig. 7. Very good alignment between predicted and ground-truth MOS values can be observed, further confirmed by PLCC of 0.917 and SRCC of 0.896. The mean MOS prediction error ($|\Delta \text{MOS}|$) was 0.169. The correlation coefficients and the mean MOS prediction errors for other architectures can be found in Appendix (Table C.16).

While the results from Fig. 7 confirm that the predicted MOS values strongly correlate with the ground-truth, this alone does not reveal how well the full quality score distributions are captured. To address this, we complement the quantitative analysis with a qualitative evaluation shown in Figs. 8 and 9. The best and worst test set predictions based on

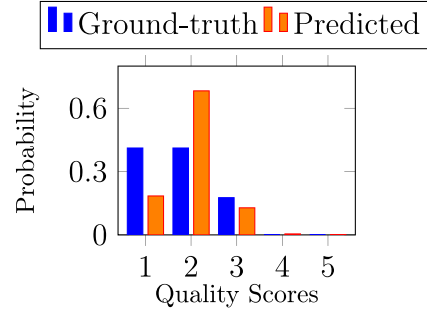
Table 7

Performance metrics for the IQDN architectures across datasets using the combined loss (KL + Huber, Eq. (1)), for the case $\lambda = 1$. Values in green indicate improvement over KL-only baseline, while red indicates degradation.

Architecture	Dataset	KL	MAE	RMSE	EMD	CD	RCGFC
<i>IQDN</i>	KonIQ-10k	0.142	0.071	0.114	0.051	0.666	0.167
<i>IQDN_{ResNet}</i>	KonIQ-10k	0.136	0.067	0.110	0.047	0.634	0.159
<i>IQDN_{VGG}</i>	KonIQ-10k	0.144	0.072	0.116	0.052	0.661	0.169
<i>IQDN_{Xcept}</i>	KonIQ-10k	0.087	0.053	0.088	0.036	0.622	0.125
<i>IQDN_{DenNet}</i>	KonIQ-10k	0.116	0.063	0.103	0.044	0.608	0.149
<i>IQDN</i>	CID2013	0.364	0.145	0.196	0.175	0.808	0.253
<i>IQDN_{ResNet}</i>	CID2013	0.440	0.170	0.221	0.216	0.817	0.298
<i>IQDN_{VGG}</i>	CID2013	0.429	0.171	0.218	0.215	0.780	0.291
<i>IQDN_{Xcept}</i>	CID2013	0.253	0.124	0.166	0.137	0.778	0.212
<i>IQDN_{DenNet}</i>	CID2013	0.248	0.119	0.164	0.128	0.695	0.209
<i>IQDN</i>	NAP540	0.133	0.071	0.101	0.064	0.576	0.154
<i>IQDN_{ResNet}</i>	NAP540	0.101	0.061	0.086	0.051	0.508	0.126
<i>IQDN_{VGG}</i>	NAP540	0.140	0.075	0.103	0.067	0.610	0.160
<i>IQDN_{Xcept}</i>	NAP540	0.095	0.058	0.082	0.047	0.450	0.118
<i>IQDN_{DenNet}</i>	NAP540	0.010	0.060	0.085	0.050	0.489	0.124



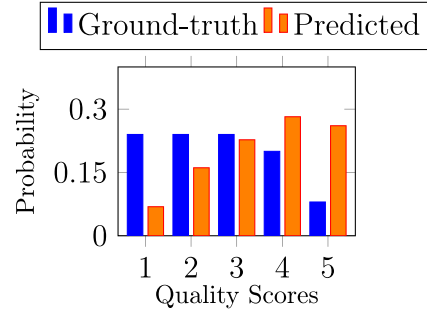
(a) Image for which low EMD but high RMSE were obtained.



(b) The predicted distribution for image 6(a), RMSE = 0.1596, EMD = 0.0551.



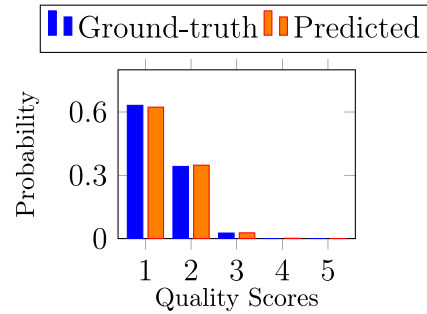
(c) Image for which moderate RMSE but high EMD were obtained.



(d) The predicted distribution for image 6(c), RMSE = 0.1225, EMD = 0.1729.



(e) Image for which low RMSE and low EMD were obtained.



(f) The predicted distribution for image 6(e), RMSE = 0.0047, EMD = 0.0026.

Fig. 6. Qualitative examples illustrating the complementary nature of RMSE and EMD.

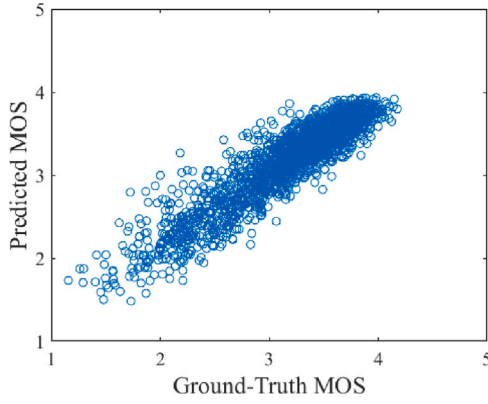


Fig. 7. The predicted and the ground-truth MOS values for $IQDN_{Xcept}$ on the KonIQ-10k dataset.

Table 8

Pairwise Wilcoxon signed-rank test results on KonIQ-10k (metric: EMD). Lower values indicate better performance. Significant adjusted p -values ($p_{adj} < 0.05$) are in **bold**.

A	B	Mean Δ	Median Δ	Pr(B<A)	p_{adj}
$IQDN$	$IQDN_{DenseNet}$	0.007	0.006	0.595	4.03e-21
$IQDN$	$IQDN_{ResNet}$	0.004	0.004	0.571	9.95e-09
$IQDN$	$IQDN_{VGG}$	-0.001	-0.001	0.476	0.051
$IQDN$	$IQDN_{Xcept}$	0.015	0.012	0.685	2.22e-87
$IQDN_{DenseNet}$	$IQDN_{ResNet}$	-0.003	-0.001	0.463	1.95e-05
$IQDN_{DenseNet}$	$IQDN_{VGG}$	-0.008	-0.007	0.400	3.70e-24
$IQDN_{DenseNet}$	$IQDN_{Xcept}$	0.008	0.006	0.616	6.65e-38
$IQDN_{ResNet}$	$IQDN_{VGG}$	-0.005	-0.004	0.420	1.37e-11
$IQDN_{ResNet}$	$IQDN_{Xcept}$	0.011	0.008	0.639	2.19e-55
$IQDN_{VGG}$	$IQDN_{Xcept}$	0.016	0.012	0.680	7.76e-89

EMD and $|\Delta MOS|$ for the best performing architecture from Table 7 are given in Fig. 8. Additionally, Fig. 9 highlights the three best and three worst test images ranked by EMD, with KL and $|\Delta MOS|$ also reported. These examples illustrate where the model excels and where it fails, and reinforce the trends observed in quantitative metrics. The three best and worst test images for CID2013 and NAP540 are given in Appendix (Figs. D.11 and D.12, respectively).

4.7. Statistical significance analysis

To further assess the robustness of the observed differences between architectures, we conducted a statistical significance analysis using the Wilcoxon signed-rank test on paired per-image errors. The test evaluates the null hypothesis that the median difference between two models is zero, thus directly quantifying whether performance differences are systematic rather than due to chance. The focus was on two complementary metrics, EMD and $|\Delta MOS|$. The first captures how well the predicted distributions match the ground-truth ones in shape, while the latter quantifies the absolute difference between the ground-truth and predicted MOS values. The p -values were corrected for multiple comparisons using Bonferroni-Holm adjustment. For each dataset, the median difference (aligned with the Wilcoxon null hypothesis), the mean difference, the proportion of the cases in which model B achieved lower error than model A (Pr(B<A)), and the Holm-adjusted p -values (p_{adj}) are reported.

For the KonIQ-10k dataset (Tables 8 and 9), all pairwise comparisons were statistically significant ($p_{adj} < 0.05$), except for $IQDN$ vs. $IQDN_{VGG}$ comparison for both metrics. $IQDN_{Xcept}$ consistently achieved lower errors than all other architectures, with highly significant differences in every comparison.

In the case of CID2013 (Tables B.12 and B.13), $IQDN_{DenseNet}$ and $IQDN_{Xcept}$ were the best performing architectures, each performing

Table 9

Pairwise Wilcoxon signed-rank test results on KonIQ-10k (metric: $|\Delta MOS|$). Lower values indicate better performance. Significant adjusted p -values ($p_{adj} < 0.05$) are in **bold**.

A	B	Mean Δ	Median Δ	Pr(B<A)	p_{adj}
$IQDN$	$IQDN_{DenseNet}$	0.029	0.019	0.554	1.92e-10
$IQDN$	$IQDN_{ResNet}$	0.013	0.010	0.530	0.004
$IQDN$	$IQDN_{VGG}$	-0.004	-0.005	0.478	0.027
$IQDN$	$IQDN_{Xcept}$	0.068	0.051	0.644	2.56e-57
$IQDN_{DenseNet}$	$IQDN_{ResNet}$	-0.016	-0.008	0.469	6.98e-05
$IQDN_{DenseNet}$	$IQDN_{VGG}$	-0.033	-0.029	0.432	5.78e-14
$IQDN_{DenseNet}$	$IQDN_{Xcept}$	0.038	0.031	0.601	2.61e-29
$IQDN_{ResNet}$	$IQDN_{VGG}$	-0.017	-0.014	0.460	2.96e-05
$IQDN_{ResNet}$	$IQDN_{Xcept}$	0.054	0.043	0.627	1.51e-45
$IQDN_{VGG}$	$IQDN_{Xcept}$	0.071	0.056	0.642	5.34e-60

significantly better than other architectures in pair-wise comparisons. However, these two architectures compared to each other did not show significant differences, nor did $IQDN_{ResNet}$ and $IQDN_{VGG}$ compared to each other.

For the NAP540 dataset (Tables B.14 and B.15), the magnitude of the differences was greater than that of the KonIQ-10k dataset. Again, $IQDN_{Xcept}$ performed significantly better than all other architectures in pairwise tests. Comparisons between $IQDN$ and $IQDN_{VGG}$, as well as between $IQDN_{DenseNet}$ and $IQDN_{ResNet}$ were not significant, suggesting comparable performance between those pairs.

The general results obtained confirm that $IQDN_{Xcept}$ is the best performing model in all datasets, with significantly and consistently lower prediction errors in nearly all pairwise comparisons. $IQDN_{DenseNet}$ emerged as a competitive alternative on the CID2013 dataset. It is important to note that, while the statistical test demonstrated robust and systematic differences, the absolute effect sizes were small, with the median difference in predicted MOS error typically between 0.03 and 0.10 units on a 1–5 scale. This suggests that although the differences are statistically significant due to large number of test images, their practical perceptual impact is limited.

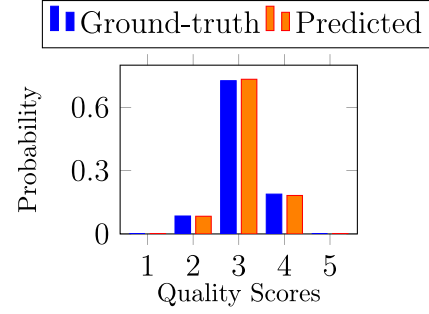
4.8. Cross-dataset testing

Cross-dataset testing was attempted in two scenarios. First, the NAP540 dataset was used for training and the CID2013 and KonIQ-10k datasets for testing. In the second scenario, the KonIQ-10k dataset was used for training and the other two datasets were used for testing. A closer inspection of Table 10 reveals that performance degradation was not uniform between datasets. Models trained on NAP540 transfer relatively well to KonIQ-10k (KL $\approx$ 0.36–0.44), but fail on CID2013 (KL $\approx$ 0.95–1.06). Training on KonIQ-10k yields moderate transfer to NAP540 (KL $\approx$ 0.73–0.87) but poor transfer to CID2013 (KL $\approx$ 1.58–1.77). This pattern indicates that CID2013 is particularly challenging for cross-dataset generalization. Possible reasons for this behavior can be the limited size of the CID2013 dataset compared to the other two datasets, the simple rescaling from 0–100 to 0–5 scale, as well as the significant differences in the underlying score distributions compared to the other two datasets (Fig. 2). These factors could explain why CID2013 consistently emerges as an outlier in cross-dataset evaluation. Among the architectures, $IQDN_{DenseNet}$ consistently showed the best robustness, especially when trained on NAP540 and tested on the KonIQ dataset.

Cross-dataset evaluation was expanded on the three additional datasets (CIDGD, VSR, CIDOA), which none of the models had encountered during training (Table 11). In general, models trained on NAP540 transfer better to these unseen datasets than those trained on KonIQ-10k, particularly for CIDGD where the KL divergence drops to 0.40–0.63 across architectures, compared to 0.94–1.35 when trained on KonIQ-10k. The CIDGD results suggest that its distortion types and score distributions might be more closely aligned with those in



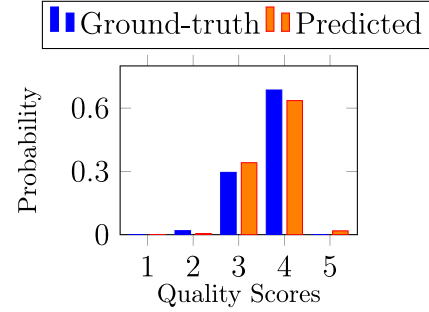
(a) Image with the lowest EMD obtained.



(b) The predicted distribution for image 8(a).



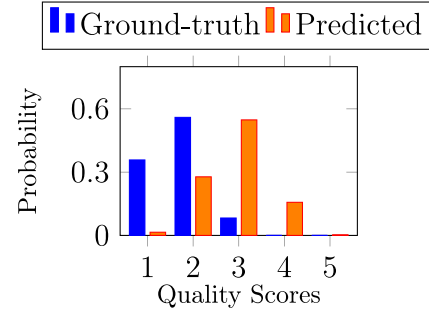
(c) Image with the best MOS prediction (lowest $|\Delta\text{MOS}|$).



(d) The predicted distribution for image 8(c).



(e) Image with the highest EMD and $|\Delta\text{MOS}|$.



(f) The predicted distribution for image 8(e).

Fig. 8. The best and the worst obtained results for $IQDN_{Xcept}$ on the KonIQ-10k dataset.

NAP540, enabling more reliable predictions. One reason for this could be the fact that NAP540 and CIDGD dataset share more systematic distortion variations and content characteristics compared to the more diverse KonIQ-10k dataset. On the other hand, performance on VSR and CIDOA remains weaker across all models ($KL \approx 0.60\text{--}1.42$), indicating a larger distribution shift in terms of both content diversity and distortion representation. Among the architectures, $IQDN_{VGG}$ and the simpler $IQDN$ achieve the most competitive results on CIDGD, with the $IQDN$ obtaining the best overall scores in terms of CD, RCGFC, MAE, and RMSE. This suggests that model depth does not necessarily guarantee robustness in cross-dataset settings. Taken together, these findings reinforce that unseen datasets pose substantial challenges, with generalization success depending not only on architecture but, critically, on the degree of similarity in present image distortions and label distributions between training and testing sets.

5. Conclusion and future work

The study addressed the limitations of conventional blind IQA methods that predict only single MOS values by proposing the Image Quality Distribution Network (IQDN), which is capable of predicting the complete distribution of subjective quality scores. IQDN was implemented in five configurations, including a custom CNN trained from scratch and four transfer learning variants with ResNet50, VGG16, Xception, and DenseNet121 backbones. Unlike previous studies that often evaluated the configuration of a proposed model on a single dataset containing ground-truth quality distribution scores, our study provides a broader perspective on the generalization strengths and weaknesses of the proposed architectures. The models were trained and evaluated on three datasets (KonIQ-10k, CID2013, and NAP540 datasets), using a wide range of point-wise error and distribution similarity metrics. In

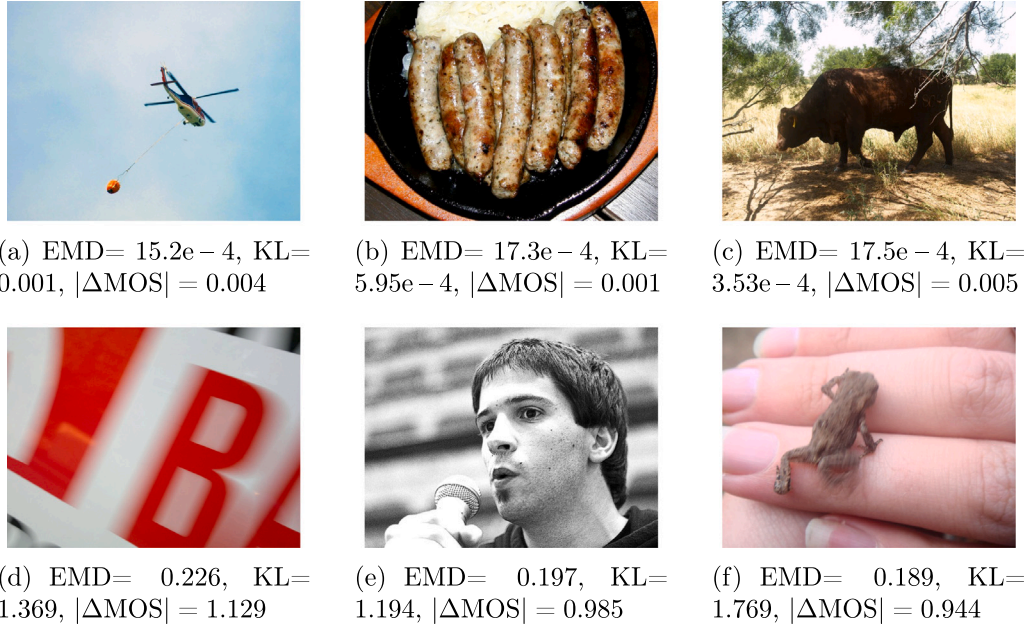


Fig. 9. Three images with the best (top row) and the worst (bottom row) EMD values for $\text{IQDN}_{\text{Xcept}}$ on the KonIQ-10k dataset.

Table 10

Cross-dataset performance for each architecture using the custom loss ($\lambda = 1$). Metrics reported per (Train $\rightarrow$ Test) combination. The best performance is shown in **bold**.

Architecture	Train $\rightarrow$ Test	KL	MAE	RMSE	EMD	CD	RCGFC
IQDN	KonIQ-10k $\rightarrow$ CID2013	1.573	0.215	0.278	0.187	0.978	0.511
	KonIQ-10k $\rightarrow$ NAP540	0.728	0.140	0.178	0.120	0.987	0.320
	NAP540 $\rightarrow$ CID2013	0.966	0.169	0.204	0.189	0.971	0.415
	NAP540 $\rightarrow$ KonIQ-10k	0.438	0.146	0.183	0.116	0.784	0.299
$\text{IQDN}_{\text{ResNet}}$	KonIQ-10k $\rightarrow$ CID2013	1.727	0.213	0.279	0.173	0.985	0.498
	KonIQ-10k $\rightarrow$ NAP540	0.879	0.143	0.184	0.117	0.988	0.320
	NAP540 $\rightarrow$ CID2013	1.022	0.168	0.204	0.186	0.968	0.414
	NAP540 $\rightarrow$ KonIQ-10k	0.378	0.135	0.172	0.104	0.790	0.279
IQDN_{VGG}	KonIQ-10k $\rightarrow$ CID2013	1.573	0.215	0.278	0.187	0.978	0.511
	KonIQ-10k $\rightarrow$ NAP540	0.728	0.140	0.178	0.120	0.981	0.320
	NAP540 $\rightarrow$ CID2013	0.641	0.158	0.193	0.172	0.994	0.385
	NAP540 $\rightarrow$ KonIQ-10k	0.439	0.148	0.184	0.118	0.806	0.299
$\text{IQDN}_{\text{Xcept}}$	KonIQ-10k $\rightarrow$ CID2013	1.773	0.215	0.286	0.166	0.964	0.499
	KonIQ-10k $\rightarrow$ NAP540	0.812	0.144	0.186	0.112	0.944	0.318
	NAP540 $\rightarrow$ CID2013	1.064	0.175	0.212	0.190	0.990	0.431
	NAP540 $\rightarrow$ KonIQ-10k	0.414	0.141	0.181	0.112	0.806	0.310
$\text{IQDN}_{\text{DenseNet}}$	KonIQ-10k $\rightarrow$ CID2013	1.587	0.210	0.274	0.172	0.972	0.493
	KonIQ-10k $\rightarrow$ NAP540	0.843	0.139	0.178	0.113	0.969	0.309
	NAP540 $\rightarrow$ CID2013	0.957	0.165	0.200	0.178	0.933	0.405
	NAP540 $\rightarrow$ KonIQ-10k	0.363	0.131	0.168	0.101	0.951	0.275

addition, three smaller scale datasets were used to perform the cross-dataset evaluation (CIDGD, VSR, and CIDOA). The networks were initially trained using KL divergence as a loss function. It was shown that using a combination of KL divergence and Huber loss improves the performance of the networks. The results demonstrated that IQDN with Xception backbone achieved the strongest overall performance, particularly in Earth Mover's Distance, indicating a high fidelity in capturing the shape and spread of subjective score distributions. The variant with DenseNet121 backbone offered competitive performance at a fraction of computational cost, making it attractive for scenarios where resources are limited. Although cross-dataset generalization remains challenging, incorporating a quality-aware backbone showed potential for improving transfer performance, suggesting the value of task-specific pre-training.

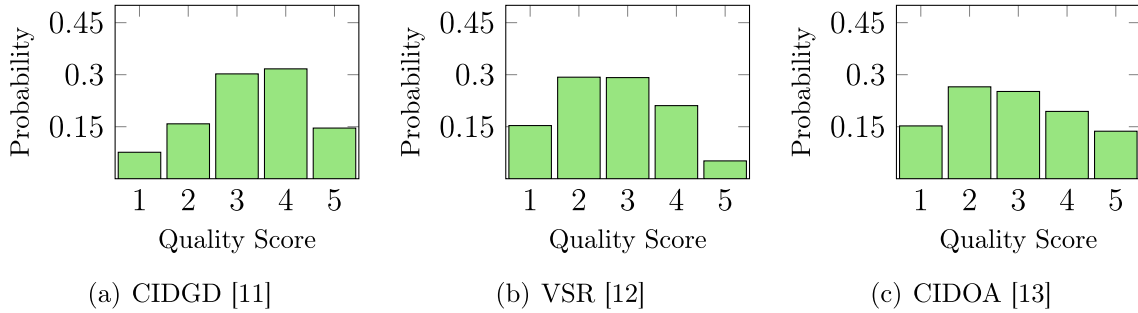
The study opened several new directions for future work. In order to improve cross-dataset generalization, further investigating domain adaptation and distribution alignment techniques is needed in order

to handle prior distribution differences between datasets better. Using a quality-aware backbone showed potential for improving the cross-dataset robustness. Future work will explore this direction more extensively. One promising direction is to adopt a student-teacher framework, where a quality-aware teacher (such as Koncept512) guides the training of a lighter or more specialized student network. Another option is to utilize other types of quality-aware networks that were not originally trained for this specific task, for example, models trained for image distortion classification. Integrating such backbones with transformer-based modules in hybrid architectures could potentially lead to better capture of both local detail and long-range dependencies in quality perception. As transformer architectures require much higher amounts of ground-truth data for training, larger-scale datasets containing full score distributions are needed. Loss function experiments showed a promising direction, opening a possibility to further develop scale-balanced and multi-term objective functions that jointly optimize distribution alignment and robustness to outliers.

Table 11

Cross-dataset performance on VSR, CIDGD, and CIDOA for each architecture using the custom loss ($\lambda = 1$). Metrics reported per (Train $\rightarrow$ Test) combination. The best performance is shown in **bold**.

Architecture	Train $\rightarrow$ Test	KL	MAE	RMSE	EMD	CD	RCGFC
<i>IQDN</i>	KonIQ-10k $\rightarrow$ VSR	1.261	0.172	0.217	0.161	0.954	0.406
	KonIQ-10k $\rightarrow$ CIDGD	1.157	0.155	0.198	0.140	0.719	0.344
	KonIQ-10k $\rightarrow$ CIDOA	1.058	0.175	0.229	0.150	0.917	0.427
	NAP540 $\rightarrow$ VSR	0.659	0.153	0.180	0.170	0.941	0.360
	NAP540 $\rightarrow$ CIDGD	0.447	0.114	0.135	0.122	0.663	0.257
	NAP540 $\rightarrow$ CIDOA	0.493	0.145	0.183	0.132	0.747	0.346
<i>IQDN_{ResNet}</i>	KonIQ-10k $\rightarrow$ VSR	1.367	0.168	0.212	0.145	0.954	0.387
	KonIQ-10k $\rightarrow$ CIDGD	1.212	0.152	0.191	0.134	0.667	0.338
	KonIQ-10k $\rightarrow$ CIDOA	1.217	0.175	0.226	0.151	0.920	0.424
	NAP540 $\rightarrow$ VSR	0.900	0.161	0.190	0.183	0.951	0.387
	NAP540 $\rightarrow$ CIDGD	0.513	0.115	0.137	0.124	0.666	0.265
	NAP540 $\rightarrow$ CIDOA	0.445	0.137	0.175	0.121	0.803	0.326
<i>IQDN_{VGG}</i>	KonIQ-10k $\rightarrow$ VSR	1.197	0.170	0.212	0.157	0.954	0.401
	KonIQ-10k $\rightarrow$ CIDGD	0.942	0.147	0.183	0.133	0.666	0.328
	KonIQ-10k $\rightarrow$ CIDOA	0.995	0.173	0.224	0.152	0.916	0.420
	NAP540 $\rightarrow$ VSR	0.614	0.147	0.173	0.159	0.919	0.341
	NAP540 $\rightarrow$ CIDGD	0.402	0.118	0.139	0.127	0.662	0.267
	NAP540 $\rightarrow$ CIDOA	0.580	0.157	0.197	0.151	0.757	0.379
<i>IQDN_{Xcept}</i>	KonIQ-10k $\rightarrow$ VSR	1.424	0.174	0.221	0.150	0.955	0.403
	KonIQ-10k $\rightarrow$ CIDGD	1.194	0.150	0.190	0.128	0.667	0.329
	KonIQ-10k $\rightarrow$ CIDOA	0.987	0.162	0.215	0.126	0.916	0.387
	NAP540 $\rightarrow$ VSR	0.950	0.165	0.194	0.187	0.953	0.398
	NAP540 $\rightarrow$ CIDGD	0.491	0.117	0.138	0.127	0.664	0.272
	NAP540 $\rightarrow$ CIDOA	0.454	0.141	0.180	0.124	0.852	0.338
<i>IQDN_{DenNet}</i>	KonIQ-10k $\rightarrow$ VSR	1.393	0.168	0.212	0.149	0.954	0.391
	KonIQ-10k $\rightarrow$ CIDGD	1.359	0.152	0.192	0.135	0.682	0.337
	KonIQ-10k $\rightarrow$ CIDOA	1.091	0.167	0.217	0.141	0.895	0.404
	NAP540 $\rightarrow$ VSR	0.826	0.156	0.185	0.173	0.952	0.375
	NAP540 $\rightarrow$ CIDGD	0.631	0.121	0.144	0.133	0.667	0.285
	NAP540 $\rightarrow$ CIDOA	0.428	0.132	0.170	0.114	0.796	0.316

**Fig. A.10.** Average normalized histograms of the quality scores for the three additional datasets.

CRediT authorship contribution statement

Nikola Plavac: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Seyed Ali Amirshahi:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Marius Pedersen:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Sophie Triantaphillidou:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Funding sources

Marius Pedersen and Sophie Triantaphillidou were supported by the project “Quality and Content: understanding the influence of content on subjective and objective image quality assessment” (grant 324663) from the Research Council of Norway. Seyed Ali Amirshahi was supported by the project “VQ4MedicS: Video Quality Assessment and Enhancement for Pre-Hospital Medical Services” (grant number 329034, approved on 1 September 2021) from the Research Council of Norway.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Seyed Ali Amirshahi, Marius Pedersen, Sophie Triantaphillidou reports financial support was provided by Research Council of Norway. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary analysis for CIDGD, VSR, and CIDOA datasets.

To investigate the probability distributions of quality scores in three additional datasets (CIDGD, VSR, and CIDOA), the average normalized histograms of the scores are shown in Fig. A.10. Compared to those obtained for the main three datasets (Fig. 2), the average normalized histograms resemble those of the NAP540 (Fig. 2(c)), especially in the case of the VSR and CIDOA datasets.

Table B.12

Pairwise Wilcoxon signed-rank test results on CID2013 (metric: EMD). Lower values indicate better performance. Significant adjusted p -values ($p_{\text{adj}} < 0.05$) are in **bold**.

A	B	Mean Δ	Median Δ	Pr(B<A)	p_{adj}
<i>IQDN</i>	<i>IQDN</i> _{DenNet}	0.047	0.044	0.667	0.001
<i>IQDN</i>	<i>IQDN</i> _{ResNet}	-0.042	-0.024	0.354	0.004
<i>IQDN</i>	<i>IQDN</i> _{VGG}	-0.040	-0.025	0.292	6e-05
<i>IQDN</i>	<i>IQDN</i> _{Xcept}	0.038	0.025	0.688	2.57e-04
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{ResNet}	-0.089	-0.080	0.208	3.01e-08
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{VGG}	-0.087	-0.079	0.229	2.11e-08
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{Xcept}	-0.009	-0.016	0.396	0.431
<i>IQDN</i> _{ResNet}	<i>IQDN</i> _{VGG}	0.001	-0.007	0.417	0.991
<i>IQDN</i> _{ResNet}	<i>IQDN</i> _{Xcept}	0.080	0.062	0.792	6.36e-10
<i>IQDN</i> _{VGG}	<i>IQDN</i> _{Xcept}	0.078	0.058	0.833	5.30e-12

Table B.13

Pairwise Wilcoxon signed-rank test results on CID2013 (metric: $|\Delta\text{MOS}|$). Lower values indicate better performance. Significant adjusted p -values ($p_{\text{adj}} < 0.05$) are in **bold**.

A	B	Mean Δ	Median Δ	Pr(B<A)	p_{adj}
<i>IQDN</i>	<i>IQDN</i> _{DenNet}	0.214	0.224	0.656	0.011
<i>IQDN</i>	<i>IQDN</i> _{ResNet}	-0.197	-0.211	0.417	0.023
<i>IQDN</i>	<i>IQDN</i> _{VGG}	-0.167	-0.076	0.385	0.023
<i>IQDN</i>	<i>IQDN</i> _{Xcept}	0.189	0.176	0.646	0.002
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{ResNet}	-0.411	-0.449	0.260	3.15e-06
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{VGG}	-0.381	-0.364	0.302	1.25e-05
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{Xcept}	-0.025	-0.065	0.427	0.497
<i>IQDN</i> _{ResNet}	<i>IQDN</i> _{VGG}	0.030	-0.026	0.490	0.272
<i>IQDN</i> _{ResNet}	<i>IQDN</i> _{Xcept}	0.386	0.328	0.729	3.38e-08
<i>IQDN</i> _{VGG}	<i>IQDN</i> _{Xcept}	0.357	0.240	0.750	1.40e-08

Table B.14

Pairwise Wilcoxon signed-rank test results on NAP540 (metric: EMD). Lower values indicate better performance. Significant adjusted p -values ($p_{\text{adj}} < 0.05$) are in **bold**.

A	B	Mean Δ	Median Δ	Pr(B<A)	p_{adj}
<i>IQDN</i>	<i>IQDN</i> _{DenNet}	0.015	0.007	0.611	3.24e-23
<i>IQDN</i>	<i>IQDN</i> _{ResNet}	0.013	0.008	0.628	1.60e-25
<i>IQDN</i>	<i>IQDN</i> _{VGG}	-0.003	-0.002	0.467	0.005
<i>IQDN</i>	<i>IQDN</i> _{Xcept}	0.017	0.009	0.641	2.12e-32
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{ResNet}	-0.001	-0.001	0.482	0.062
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{VGG}	-0.017	-0.013	0.344	1.14e-37
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{Xcept}	0.002	0.002	0.540	0.005
<i>IQDN</i> _{ResNet}	<i>IQDN</i> _{VGG}	-0.016	-0.012	0.318	6.78e-39
<i>IQDN</i> _{ResNet}	<i>IQDN</i> _{Xcept}	0.004	0.003	0.560	1.78e-05
<i>IQDN</i> _{VGG}	<i>IQDN</i> _{Xcept}	0.020	0.014	0.685	2.53e-47

Appendix B. Supplementary tables for the statistical significance analysis

The results for pairwise Wilcoxon signed-rank test on CID2013 dataset with EMD as the target metric are given in Table B.12. The corresponding results for the mean MOS prediction error ($|\Delta\text{MOS}|$) as the target metric are given in Table B.13.

The results for pairwise Wilcoxon signed-rank test on NAP540 dataset with EMD as the target metric are given in Table B.14. The corresponding results for the mean MOS prediction error ($|\Delta\text{MOS}|$) as the target metric are given in Table B.15.

Table B.15

Pairwise Wilcoxon signed-rank test results on NAP540 (metric: $|\Delta\text{MOS}|$). Lower values indicate better performance. Significant adjusted p -values ($p_{\text{adj}} < 0.05$) are in **bold**.

A	B	Mean Δ	Median Δ	Pr(B<A)	p_{adj}
<i>IQDN</i>	<i>IQDN</i> _{DenNet}	0.075	0.046	0.612	3.58e-19
<i>IQDN</i>	<i>IQDN</i> _{ResNet}	0.070	0.055	0.618	6.78e-21
<i>IQDN</i>	<i>IQDN</i> _{VGG}	-0.009	-0.006	0.492	0.210
<i>IQDN</i>	<i>IQDN</i> _{Xcept}	0.091	0.058	0.629	3.74e-28
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{ResNet}	-0.006	0.000	0.500	0.335
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{VGG}	-0.085	-0.062	0.392	8.65e-25
<i>IQDN</i> _{DenNet}	<i>IQDN</i> _{Xcept}	0.016	0.013	0.559	0.000
<i>IQDN</i> _{ResNet}	<i>IQDN</i> _{VGG}	-0.079	-0.065	0.358	1.44e-26
<i>IQDN</i> _{ResNet}	<i>IQDN</i> _{Xcept}	0.021	0.016	0.559	1.03e-05
<i>IQDN</i> _{VGG}	<i>IQDN</i> _{Xcept}	0.100	0.073	0.643	8.93e-36

Table C.16

Performance metrics ($|\Delta\text{MOS}|$, PLCC, SRCC) for different backbone architectures across KonIQ-10k, CID2013, and NAP540 datasets. The best performance per metric is shown in **bold**.

Dataset	Architecture	$ \Delta\text{MOS} $	PLCC	SRCC
KonIQ-10k	<i>IQDN</i>	0.236	0.826	0.799
	<i>IQDN</i> _{ResNet}	0.223	0.852	0.825
	<i>IQDN</i> _{VGG}	0.240	0.822	0.787
	<i>IQDN</i> _{Xcept}	0.169	0.917	0.896
	<i>IQDN</i> _{DenNet}	0.207	0.873	0.851
CID2013	<i>IQDN</i>	0.808	0.554	0.516
	<i>IQDN</i> _{ResNet}	1.005	-0.288	-0.283
	<i>IQDN</i> _{VGG}	0.975	0.389	0.397
	<i>IQDN</i> _{Xcept}	0.619	0.812	0.801
	<i>IQDN</i> _{DenNet}	0.594	0.785	0.761
NAP540	<i>IQDN</i>	0.284	0.904	0.902
	<i>IQDN</i> _{ResNet}	0.214	0.949	0.949
	<i>IQDN</i> _{VGG}	0.293	0.902	0.892
	<i>IQDN</i> _{Xcept}	0.193	0.959	0.959
	<i>IQDN</i> _{DenNet}	0.208	0.952	0.952

Appendix C. Supplementary table for predicted MOS error and correlation coefficients

The mean predicted MOS error ($|\Delta\text{MOS}|$), Pearson Linear Correlation Coefficient (PLCC), and Spearman Rank Correlation Coefficient (SRCC) were calculated for each proposed architecture. The values obtained across the KonIQ-10k, CID2013, and NAP540 datasets are given in Table C.16.

Appendix D. Supplementary figures displaying the best and the worst performing examples for CID2013 and NAP540 datasets

For additional qualitative analysis, the Top-3 and Bottom-3 test images by Earth Mover's Distance (EMD) for *IQDN*_{Xcept} trained with the combined KL+Huber loss (Section 4.6) are shown for the CID2013 and NAP540 datasets in Figs. D.11 and D.12, respectively.

Data availability

The repository link with the codes and dataset organization is shared in the article.

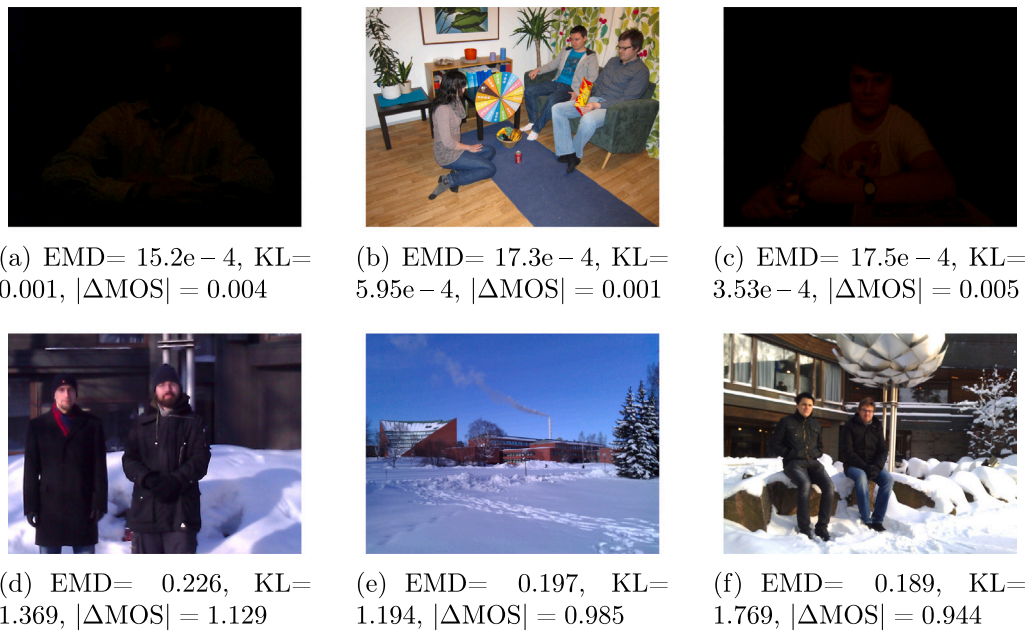


Fig. D.11. Three images with the best (top row) and the worst (bottom row) EMD values for $\text{IQDN}_{\text{Xcept}}$ on the CID2013 dataset.

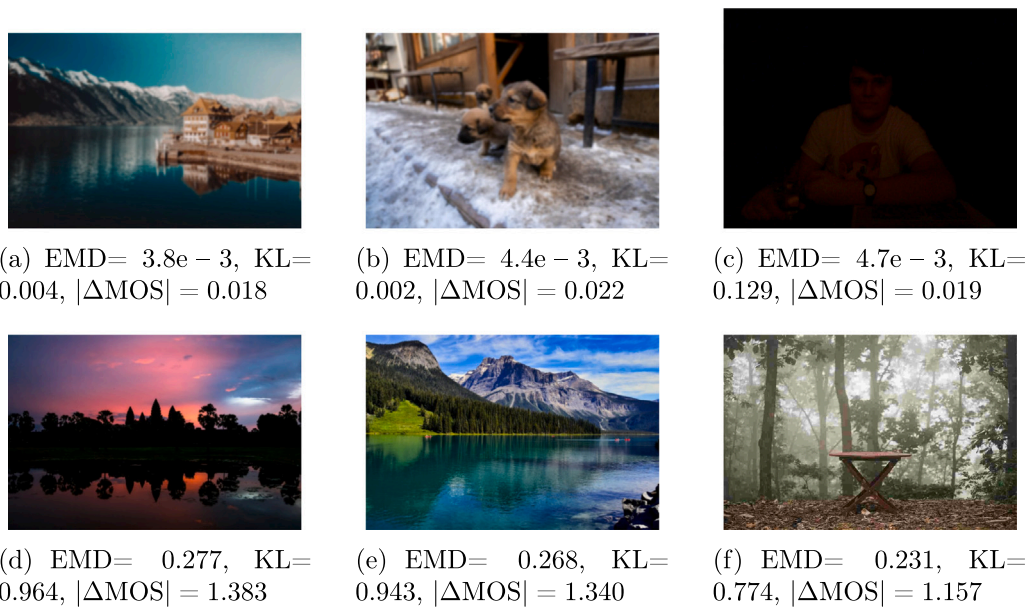


Fig. D.12. Three images with the best (top row) and the worst (bottom row) EMD values for $\text{IQDN}_{\text{Xcept}}$ on the NAP540 dataset.

References

- [1] R.K. Mantiuk, A. Tomaszewska, R. Mantiuk, Comparison of four subjective methods for image quality assessment, Comput. Graph. Forum (1981) 1–13.
- [2] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
- [3] H. Talebi, P. Milanfar, NIMA: Neural image assessment, IEEE Trans. Image Process. 27 (8) (2018) 3998–4011.
- [4] D. Varga, D. Saupe, T. Sziranyi, Deepn: A content preserving deep architecture for blind image quality assessment, in: 2018 IEEE International Conference on Multimedia and Expo, ICME, IEEE Computer Society, Los Alamitos, CA, USA, 2018, pp. 1–6.
- [5] Y. Gao, X. Min, Y. Cao, X. Liu, G. Zhai, No-reference image quality assessment: Obtain MOS from image quality score distribution, IEEE Trans. Circuits Syst. Video Technol. 35 (2) (2025) 1840–1854.
- [6] Y. Gao, X. Min, Y. Zhu, J. Li, X. Zhang, G. Zhai, Image quality assessment: From mean opinion score to opinion score distribution, in: Proceedings of the 30th ACM International Conference on Multimedia, MM'22, 2022, pp. 997–1005.
- [7] Y. Gao, X. Min, W. Zhu, X. Zhang, G. Zhai, Image quality score distribution prediction via alpha stable model, IEEE Trans. Circuits Syst. Video Technol. 33 (6) (2023) 2656–2669.
- [8] NTNU Colourlab, NAP540 image quality assessment dataset, 2025, (pending article publication) https://osf.io/vak62/?view_only=9e15ba8200c74bb9a70cf509ad7b251e. (Accessed 26 August 2025).
- [9] V. Hosu, H. Lin, T. Sziranyi, D. Saupe, KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment, IEEE Trans. Image Process. 29 (2020) 4041–4056.
- [10] T. Virtanen, M. Nuutinen, M. Vaaherankoska, P. Oittinen, J. Häkkinen, CID2013: A database for evaluating no-reference image quality assessment algorithms, IEEE Trans. Image Process. 24 (1) (2015) 390–402.
- [11] M. Pedersen, S.A. Amirshahi, Colourlab image database: Geometric distortions, in: Color and Imaging Conference, CIC, 2021, pp. 258–263.
- [12] O. Cherepkova, S.A. Amirshahi, M. Pedersen, Analyzing the variability of subjective image quality ratings for different distortions, in: 2022 Eleventh

- International Conference on Image Processing Theory, Tools and Applications, IPTA, 2022, pp. 1–6.
- [13] R. Hlayhel, M. Mobini, B.E. Agossou, M. Pedersen, S.A. Amirshahi, Colourlab image database: Optical aberrations, in: London Imaging Meeting, LIM, 2024, pp. 22–26.
- [14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representation, 2015, Available on <https://arxiv.org/abs/1409.1556>. (Accessed 29 September 2025).
- [15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 2818–2826.
- [16] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint [arXiv:1704.04861](https://arxiv.org/abs/1704.04861). (Accessed 29 September 2025).
- [17] Y. Rubner, C. Tomasi, L.J. Guibas, The earth mover's distance as a metric for image retrieval, *Int. J. Comput. Vis.* 40 (2) (2000) 99–121.
- [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, CVPR, 2016, pp. 770–778.
- [19] L. Fotio Tiotop, A. Servetti, M. Barkowsky, P. Pocta, T. Mizdos, G. Van Wallendaal, E. Masala, Predicting individual quality ratings of compressed images through deep CNNs-based artificial observers, *Signal Process., Image Commun.* 112 (2023) 116917.
- [20] H.R. Sheikh, Z. Wang, L. Cormack, A.C. Bovik, LIVE: Image Quality Assessment Database Release 2, 2005, <http://live.ece.utexas.edu/research/quality>.
- [21] E.C. Larson, D.M. Chandler, Most apparent distortion: Full-reference image quality assessment and the role of strategy, *J. Electron. Imaging* 19 (1) (2010) 1–21.
- [22] D. Jayaraman, A. Mittal, A.K. Moorthy, A.C. Bovik, Objective quality assessment of multiply distorted images, in: Proceedings of the 46th Asilomar Conference on Signals, Systems and Computers, ASIOMAR, 2012, pp. 1693–1697.
- [23] D. Ghadiyaram, A.C. Bovik, Massive online crowdsourced study of subjective and objective picture quality, *IEEE Trans. Image Process.* 25 (1) (2015) 372–387.
- [24] Y. Fang, H. Zhu, Y. Zeng, K. Ma, Z. Wang, Perceptual quality assessment of smartphone photography, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2020, pp. 3677–3686.
- [25] A. Zaric, N. Tatalovic, N. Brajkovic, H. Hlevnjak, M. Loncaric, E. Dumic, VCL@FER image quality assessment database, in: Proceedings of International Symposium on Electronics in Marine, ELMAR, 2011, pp. 105–110.
- [26] Y. Gao, X. Min, Y. Zhu, X. Zhang, G. Zhai, Blind image quality assessment: A fuzzy neural network for opinion score distribution prediction, *IEEE Trans. Circuits Syst. Video Technol.* 34 (3) (2024) 1641–1653.
- [27] W. Zhang, K. Ma, J. Yan, D. Deng, Z. Wang, Blind image quality assessment using a deep bilinear convolutional neural network, *IEEE Trans. Circuits Syst. Video Technol.* 30 (1) (2020) 36–47.
- [28] M. Cheon, S. Yoon, B. Kang, J. Lee, Perceptual image quality assessment with transformers, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2021, pp. 433–442.
- [29] J. Shi, P. Gao, J. Qin, Transformer-based no-reference image quality assessment via supervised contrastive learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (no. 5) 2024, pp. 4829–4837.
- [30] K. Xu, L. Liao, J. Xiao, C. Chen, H. Wu, Q. Yan, W. Lin, Boosting image quality assessment through efficient transformer adaptation with local feature enhancement, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 2662–2672.
- [31] Z. You, Z. Li, J. Gu, Z. Yin, T. Xue, C. Dong, Depicting beyond scores: Advancing image quality assessment through multi-modal language models, in: 2024 European Conference on Computer Vision, ECCV, 2024, pp. 256–276.
- [32] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 1251–1258.
- [33] G. Huang, Z. Liu, L. van der Maaten, K. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 4700–4708.
- [34] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009, pp. 248–255.
- [35] S. Kullback, R.A. Leibler, On information and sufficiency, *Ann. Math. Stat.* 22 (1) (1951) 79–86.
- [36] J.A.S. Viggiano, Metrics for evaluating spectral matches: A quantitative comparison, in: CGIV2004: The Second European Conference on Computer Graphics, Imaging and Vision, 2004, pp. 286–291.