Quality and User Experience (2023) 8:8
https://doi.org/10.1007/s41233-023-00061-8

RESEARCH ARTICLE

Konx: cross-resolution image quality assessment

Oliver Wiedemann[1] · Vlad Hosu[1] · Shaolin Su[1,2] · Dietmar Saupe[1]

Received: 23 December 2022

Abstract

Scale-invariance is an open problem in many computer vision subfields. For example, object labels should remain constant across scales, yet model predictions diverge in many cases. This problem gets harder for tasks where the ground-truth labels change with the presentation scale. In image quality assessment (IQA), down-sampling attenuates impairments, e.g., blurs or compression artifacts, which can positively affect the impression evoked in subjective studies. To accurately predict perceptual image quality, cross-resolution IQA methods must therefore account for resolution-dependent discrepancies induced by model inadequacies as well as for the perceptual label shifts in the ground truth. We present the first study of its kind that disentangles and examines the two issues separately via *KonX*, a novel, carefully crafted cross-resolution IQA database. This paper contributes the following: 1. Through *KonX*, we provide empirical evidence of label shifts caused by changes in the presentation resolution. 2. We show that objective IQA methods have a scale bias, which reduces their predictive performance. 3. We propose a multi-scale and multi-column deep neural network architecture that improves performance over previous state-of-the-art IQA models for this task. We thus both raise and address a novel research problem in image quality assessment.

Keywords Image quality assessment · Cross-resolution quality prediction · IQA models · IQA databases

Introduction

The discipline of image quality assessment (IQA) aims to model how humans perceive the quality of digital images. Our interest lies mainly in technical aspects (distortions, sharpness, etc.), though artistic aspects (composition, motif, beauty, etc.) will likely affect human observers at least subconsciously. Recent no-reference (NR-)IQA algorithms predict quality scores for a given input without a pristine reference. They perform well when tested on the same domain they were trained on; however, model performance drops when cross-tested on different datasets [1–3]. We hypothesize that this decrease in performance is caused by two factors: a lack of *cross-resolution generalization* by the models and *domain shifts* across datasets. The latter is concerned with image contents and differences in the distributions of distortion types, combinations, and severities. We aim to isolate the first factor, which is also known as the *cross-resolution problem*, for image quality assessment. To this end, we created a first-of-its-kind dataset that provides a reliable benchmark for cross-resolution IQA. By resolution we mean *logical image size in pixels* (cf. Figure 1), which is to be distinguished from the *physical resolution as a pixel density*. On a display, the latter is manifested in terms of dots or pixels per inch (DPI/PPI). On the viewer's retina, a notion of angular resolution is better suited, as illustrated in Fig. 2.

Previous works in NR-IQA [1–5] assumed that the quality ratings of images gathered at one presentation resolution are valid at other resolutions as well. This is not the case. We subsequently show that perceived quality varies with the presentation resolution. When comparing images

Olive Wiedemann and Vlad Hosu have contributed equally to this work.

✉ Oliver Wiedemann
oliver.wiedemann@uni-konstanz.de

Vlad Hosu
vlad.hosu@uni-konstanz.de

Shaolin Su
shaolin_su@mail.nwpu.edu.cn

Dietmar Saupe
dietmar.saupe@uni-konstanz.de

[1] Department of Computer and Information Science, University of Konstanz, Konstanz, Germany

[2] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China

Published online: 18 August 2023

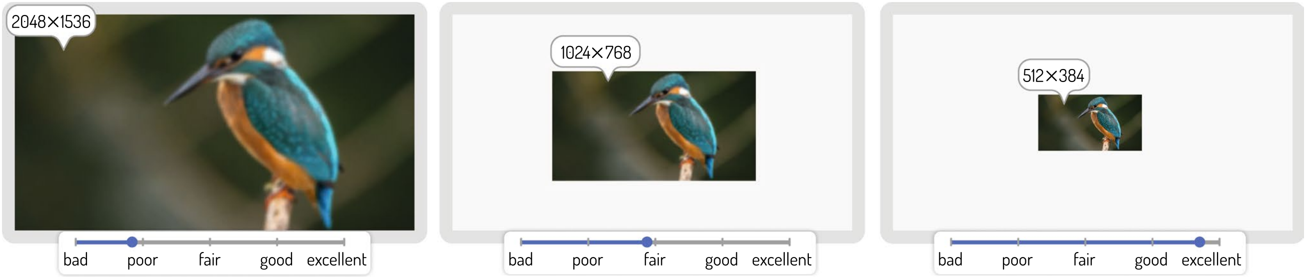
2048×1536
bad poor fair good excellent
1024×768
bad poor fair good excellent
512×384
bad poor fair good excellent

Fig. 1 Scaling affects human perception and IQA model predictions. Our interface displays images of different resolutions with varying spatial dimensions and keeps the pixel density constant across resolutions

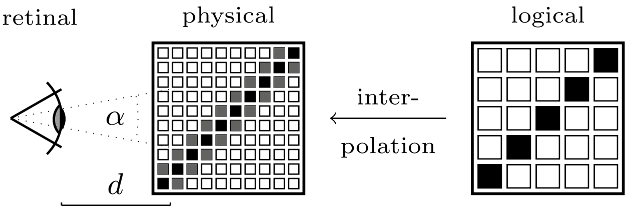

Fig. 2 The term resolution can be ambiguous. In this paper, we use it for the *logical* image size of $w \times h$ pixels. Presenting an image on a screen, possibly interpolated, yields a *physical* resolution, which defines the spatial dimensions and pixel density. What matters most for the human visual system is the perceivable *angular* resolution, which depends on the physical pixel density, the observer's distance to the screen (d), and the minimal discernible angle (α). The result is a representation of the image on the retina, which in turn evokes an impression in the visual cortex

across resolutions, we get a 0.93 Spearman rank-order correlation coefficient (SRCC) between their mean opinion scores (MOS) when the scaling factor is 4, compared to a 0.97 SRCC when scaling by a factor of only 2. Reliable IQA for modern high-resolution images is desirable, as it could pave the way for its wider application beyond academic research. Existing NR-IQA methods do not perform well in cross-resolution settings. This is in part because existing IQA databases are annotated at comparatively low resolutions and because of the prevalent approach of training and testing models on images that were resized to the same scale [1, 2, 5].

Some existing IQA datasets (e.g., [6]) contain images of various resolutions. However, there is none that was *annotated* at multiple resolutions, but the images were either scaled to a fixed presentation size or presented in their native resolution with different spatial sizes on screen. Rigorous cross-resolution comparisons on the same content were thus not possible. To address these limitations, we created *KonX*, a database in which the same image contents were annotated at multiple presentation scales. It serves as the first cross-resolution benchmark and allows to test quality predictors at multiple resolutions.

Contributions of this work

We introduce a novel problem, create a database that allows us to approach it for the first time, propose a deep neural network architecture that surpasses the state of the art, and add validation considerations that allow proper comparisons of cross-resolution model performances. The following subsections explain each of our contributions in greater detail:

A novel problem

The cross-resolution problem in NR-IQA arises by distinguishing between *cross-content* and purely *cross-resolution* predictions. The latter approach removes the confounding variable of image content from our experiments. This has not been studied before: previous IQA datasets only provided one annotation resolution per content. Particularly for crowdsourced studies, it is often unclear if and how well the actual presentation resolution was controlled for [1, 6–8]. A visualization of the cross-resolution problem in a different context is provided in Fig. 3.

A new dataset

KonX shows that the label shift is significant and that current NR-IQA models are unable to account for it.

We took the following measures to achieve precise annotations:

i) We invited expert freelancers as participants.
ii) We conducted a longitudinal study in which all items were rated twice, which provides information about participant reliability, self-consistency, and attention levels.

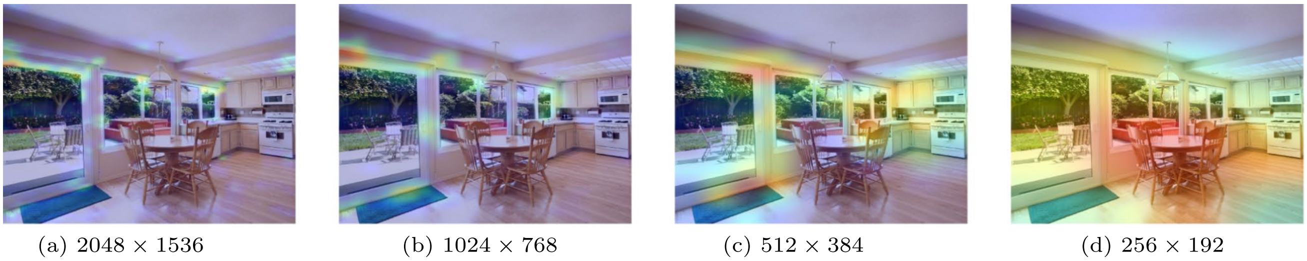
(a) 2048 × 1536 (b) 1024 × 768 (c) 512 × 384 (d) 256 × 192

Fig. 3 The cross-resolution problem: Grad-CAM [10] heatmaps depict aberrant regions-of-interest for the top predicted class of an `InceptionResNetV2` [11]. Analogous difficulties in CNN-based IQA methods are even more delicate, as perceptual quality varies with scale, unlike scale-invariant object class labels

iii) We controlled the presentation size. Our interface renders logical image pixels 1:1 to screen pixels, which was not ensured for any previous NR-IQA dataset.

KonX is publicly available[1].

A model architecture proposal

Multi-column deep neural networks (DNNs) conventionally process their input data in parallel streams. The features are merged at specific layers and then jointly utilized in the final prediction steps. For such models, weights are usually shared between the columns to limit the model capacity and prevent overfitting. We employ a transfer-learning backbone in a multi-column architecture *with individual weights for each column* that still does not overfit.

The central idea is to present different image resolutions to each column and create a bottleneck before combining per-column features. We additionally integrate information from multiple levels of the network, i.e., from all pseudo-repeated modules of the `EfficientNet` [9] backbones. These scale-variant features improve the cross-resolution performance for the quality assessment task.

Validation considerations

Absolute score prediction is crucial in cross-resolution IQA, as the ground-truth MOS changes with the image resolution. We demonstrate the limitations of singular metric choices by validating NR-IQA methods on absolute errors *and* rank correlation to ground truth. Our model outperforms recent competition in cross-database and cross-resolution comparisons regarding both metrics.

Related work

IQA models

Perceptual quality prediction evolved from statistical methods [12, 13] to an application area of deep learning. Most approaches crop or scale their input to a fixed, usually small resolution [3, 4, 14–19]. We aim to make IQA applicable at resolutions that are relevant in practice and focus on *no-reference* or *blind* IQA models, which take only the distorted image as an input and predict a quality score directly [2, 5, 20]. In comparison to *full-reference* IQA scenarios, where one has access to both the distorted image and a pristine original, the performance of NR-IQA methods in cross-resolution and cross-database tests is significantly reduced, especially on certain datasets [6, 7]. This is due to a more general problem in computer vision: scale variance [21], in this case, the *cross-resolution problem*.

We took inspiration from successful and recent works regarding model architectures, of which some already leaned towards improving robustness against input scale variance. Aggregating activations of multiple layers of pre-trained CNNs through a second network has shown success in image aesthetics assessment (IAA) [22, 23]. This inspired us to employ multi-level spatially pooled (MLSP) features in our proposed architecture as well. We noticed that CNNs [1] still perform well on *KonIQ-10k* even in comparison to transformer-based architectures [19, 24], in this case with SRCCs of 0.921 (`KonCept-512`) vs. 0.916 (`MUSIQ`[2]) and 0.915 (Golestaneh et al.). One hypothesis is that the use of both multi-scale inputs and multi-level features would be beneficial for cross-resolution prediction. It is unclear if transformers perform better in IQA than traditional CNNs, especially so for cross-resolution tasks.

[1] http://database.mmsp-kn.de.

[2] We could not rule out a data-overlap between the `MUSIQ` training set and *KonX* for the weights published by the authors, which renders a comparison on *KonX* questionable.

Some works on full-reference IQA [25, 26] integrate information from downscaled versions of their input internally. However, they are only evaluated on predictions for a single fixed resolution, so they do not approach the problem of resolution-dependent scores. NR-IQA models additionally have to intrinsically encode both the knowledge about visual distortions and their connection to the image resolution. Only a few attempts on multi-scale approaches in NR-IQA [24, 27] have been made. We considered adding explicit information about the scale similar to [24], but [28] has shown that CNNs can infer the input dimensions from the zero-padding that is often added to images before convolutional layers are applied. Another factor to consider is the prediction target. Three main types are found in the IQA literature: a single rating per image [1], the distribution of ratings from multiple annotators [3, 24], and scale-free rankings rather than absolute ratings [5, 29]. We aim to predict a single rating per image as accurately as possible across resolutions.

The MSE loss is a reasonable choice due to its characteristics when training for absolute scores. In our experiments, it did not perform worse than alternative choices, even when the evaluation metric is Spearman's rank correlation coefficient between predictions and ground-truth ratings, as commonly used in IQA. This applies to all three types of losses previously mentioned, including the scale-free rating loss introduced by Li et al. [5]. The latter work's improved performance seems to be primarily due to the choice of training resolution rather than the loss itself. Though it appears to converge faster in the early epochs, there is no clear advantage over the MSE.

Scale generalization

We incorporated works on scale generalization and transfer-learned CNNs in order to build a model that accurately predicts quality scores across resolutions. The base architecture, usually a pre-trained (e.g., on ImageNet) feature extractor, is a key choice. We expect that newer architectures will generally transfer their improved performance to downstream tasks, but multiple aspects play a role. ImageNet CNNs are usually trained at small resolutions, many at 224 × 224 pixels, up to 800 × 800px for `EfficientNet-L2` [30]. Pre-training on such small resolutions might limit the performance in large-resolution IQA. `InceptionResNet-v2` was applied successfully in IAA [23] on AVA [31], an aesthetics database that contains images of various resolutions (up to 800 × 800px). It outperformed other proposals in the past years since its introduction, raising the question: *what makes this architecture more suitable for cross-resolution tasks?*

Recent quality and aesthetics models [5, 23, 32] combine activations from multiple layers of pre-trained backbone models. Later-stage layers usually represent more abstract, scale-invariant concepts [28], whereas earlier layers tend towards scale-dependent features. IQA depends on both, e.g., object classes and pixel-level distortion patterns. This explains the benefit of utilizing features from multiple network layers for IQA.

CNNs trained on a single resolution [28, 30] exhibit scale-wise overfitting, which can be mitigated by multi-resolution ensembles [33]. Multi-column architectures have shown success in crowd-counting [34–37], a task that involves varying object scales within single images. Again, this integrates information from multiple scales. The authors of [34] feed rescaled images to a shared-weight CNN column. In contrast to most crowd-counting works, which use directly trained custom architectures for the task, we employed pre-trained networks as columns and show that they can jointly handle different scales.

Image quality databases

IQA databases are generally divided into two classes: those with *artificially* distorted images and those with *authentically* distorted images. The former are derived from pristine originals by applying distortions of various types and magnitudes, either in isolation or in combined forms [8, 38–40]. This class has been criticized for lacking diversity due to the comparatively small sets of source images and the limited variety of distortions. Models trained thereon usually have poor generalization to new impairments [41].

On the other hand, authentically distorted IQA databases are commonly sampled directly from online photography communities or adequate social media platforms. Their images are affected by mixtures of naturally occurring distortions. The state-of-the-art for general authentically distorted IQA databases is currently *KonIQ-10k* [20] with 10,073 images. *SPAQ* [42] is the largest domain-specific *authentic* dataset with 11,125 images taken with smartphone cameras.

Another subclass of databases focuses on local image quality, a concept introduced by *KonPatch-30k* [15] and extended through *Paq-2-Piq* [6]. They allow comparing the quality of patches to the quality of the entire image, which generalizes the concept of a global MOS to local image quality.

However, using only these existing IQA datasets, one *cannot* reliably study the cross-resolution problem. Though there are datasets that annotate different images (or crops thereof) at different resolutions, such as *SPAQ* [42] and *Paq-2-Piq* [6], *no dataset so far annotated the same image contents at multiple presentation resolutions*. This means neither the subjective perceptual shifts across resolutions nor the reason IQA models perform poorly in cross-resolution (and cross-dataset) tests was studied thoroughly.

Our proposed dataset, *KonX*, allows us to properly validate the cross-resolution performance of IQA models for the first time by comparing predictions versus three resolution-specific mean opinion scores. We conducted a crowdsourcing-based user study to obtain subjective ratings specifically for cross-resolution tests, and we anticipate that our work will thus pave the way for new directions in image quality research.

Subjective factors in QoE

Previous studies in which existing IQA databases were annotated did not consider well-known aspects of *quality of experience* (QoE) [43]. Reiter et al. [44] introduced three classes of influence factors (IFs) in this regard: *Human* IFs affect the lower-level (visual acuity, age, mood, etc.) and higher-level (cognitive processes, personality traits, expectations, etc.) perception of quality. *System* IFs are related to content, network, and device aspects (screen resolution, display size, etc.), while *context* IFs are affected by the environment (temporal, social, technical peculiarities, etc.). Many Reiter IFs are difficult to study, especially in crowdsourcing, where control mechanisms are lacking and self-reports can be unreliable. Several studies [45–51] report on the influence of the display device (System IF) on the perceived quality, especially regarding device characteristics.

The *visual resolution* [52] of an image presentation imposes a limit on the pixels that are discernible by the human visual system. It depends on the display size, its physical resolution, the mapping from virtual- to physical pixels, the viewing distance, and finally, the viewer's physiological capabilities, as shown in Fig. 2. Opposing effects can occur when altering the visual resolution:

- Presenting a pristine image at a higher visual resolution can increase its perceptual quality, as additional details become visible [53].
- A reduced visual resolution of a degraded image can mask impairments, which in turn can *also* increase perceptual quality.

Both effects play a role in quality assessment but have not been considered in previous works, let alone handled consistently. Moorthy et al. [45] presented videos centered on mobile screens, while Gong et al. [46] resized images to ensure a constant physical size. On the other hand, Zou et al. [48] and Kara et al. [50] opted for full-screen, rescaled as needed. The source images were not always the same size as the screen resolution.

Rehman et al. [47] did not state the presentation size, but it can be assumed to be full-screen. None of the authors mention possible discrepancies between the virtual and physical resolutions. This is relevant nowadays, especially when presenting images in browser-based user interfaces due to the reliance on rendering at smaller virtual resolutions than the physical ones. Apple Retina displays, for example, have ratios between the physical and virtual resolution up to 3:1. We consider these aspects in our study and control for them as much as possible.

The viewing distance (*Human/Context IF*) between participants and the screen was considered before. Studies involving 4K TVs [50] deemed it essential to be controlled, less so those on mobile and desktop devices [45, 48]. The latter emphasizes the freedom to choose one's preferred viewing distance to best express natural behavior instead of enforcing strict, possibly awkward or even uncomfortable setups such as chin rests.

Following this line of reasoning, we did not expect participants in our study to maintain a fixed viewing distance. It is not only difficult to enforce this in crowdsourcing, but feeling uncomfortable might reduce the participants' ability to focus on the assessment task and negatively affect their judgments.

The KonX database

Our novel cross-resolution IQA database *KonX* was annotated with subjective quality scores at three presentation resolutions. It is primarily intended as a benchmark for IQA models. With its emphasis on annotation reliability, it allows to investigate the relationship between perceived quality and scale for the first time (Fig. 4).

Overview

KonX consists of 210 images from Flickr,[3] which were already included in *KonIQ-10k* [1], and another 210 images from Pixabay[4] to supplement the high-quality range. The images were sampled using a stratified approach based on discretized metadata and image properties. The aim was to diversify both their perceptual quality levels and contents. We center-cropped all Pixabay candidates and smart-cropped [20] the original images used for *KonIQ-10k* to an aspect ratio of 4:3. These were then downsampled using the Lanczos-interpolation to 2048 × 1536px, 1024 × 768px and 512 × 384px.

Nineteen freelancers[5] with a professional background in photography or graphics design rated each image twice at each resolution. The study participants were thoroughly screened for their ability to detect image defects. We

[3] https://flickr.com.

[4] https://pixabay.com.

[5] http://freelancer.com

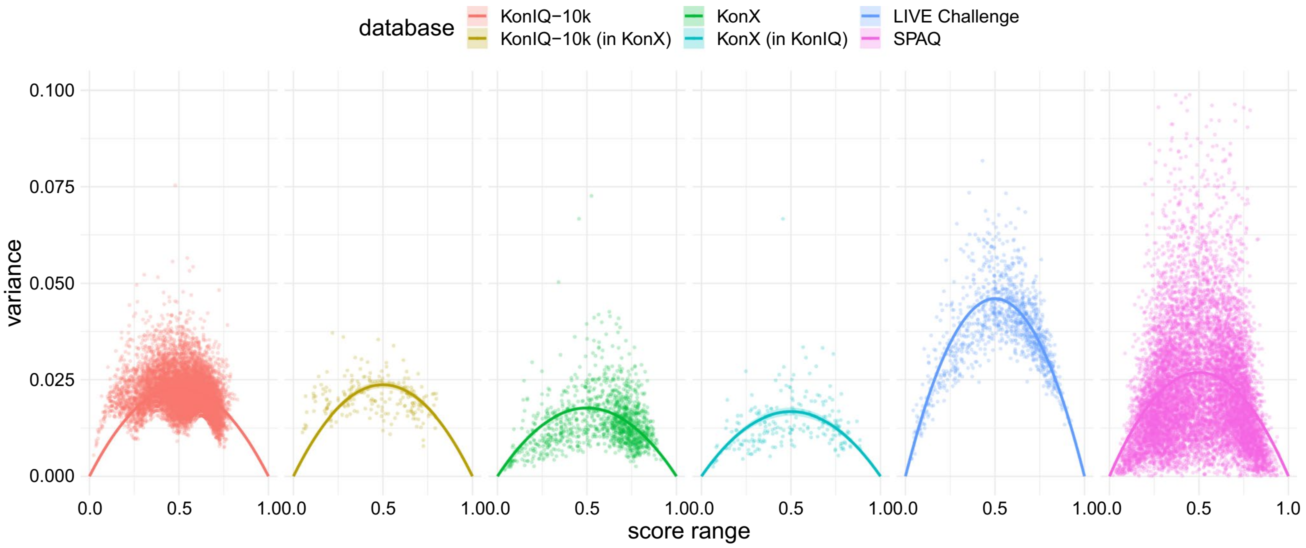

Fig. 4 Variance vs. MOS of authentically distorted, crowdsourced datasets. The a-values according to the SOS-hypothesis [54] for *KonX*, *KonX* scores at 1024 × 768 for the subset of images sampled from *KonIQ-10k*, *KonIQ-10k*, *KonIQ-10k* scores for the subset of images sampled for *KonX*, *Live Challenge*, and *SPAQ* are 0.071, 0.067, 0.091, 0.095, 0.184, and 0.107 respectively. The shaded region around the main curve indicates the 95% confidence interval for a. The parabolas for *KonX* are lower than those of the other datasets, indicating lower ratings variances

Table 1 *KonX* cross-resolution IQA benchmark

Sources	Flickr (*KonIQ-10k*) and Pixabay
#Images	210 from each source
Resolutions	2048 × 1536px, 1024 × 768px, 512 × 384px
Participants	19 in the full study
Annotations	2 per image at each resolution, 45360 in total

deployed a custom web interface that ensures a 1:1 rendering of virtual image pixels to physical screen pixels, thus displaying the lower-resolution images at a smaller spatial size. The experiment resulted in 45360 annotations of 420 image sources at three resolutions. We now explain and justify the choices behind *KonX* in detail. The most important facts are summarized below in Table 1.

Content preparation

One of the main goals when creating an IQA database is to reduce potentially unknown biases, which stem from shared characteristics among images. This problem can be mitigated by enforcing *diversity* through adequate sampling strategies. Similar goals have been set for previous IQA [1] and VQA [55] datasets. We incorporated several means to diversify *KonX* with respect to perceptual quality as the primary attribute as well as with regards to auxiliary aspects such as image content, camera type and exposure parameters.

Data sources

We sampled from two online photography platforms: Flickr[6] and Pixabay.[7] All candidate images from Flickr were already included in *KonIQ-10k* [1], which provides preexisting MOSes for comparison. This set was augmented with content from Pixabay, which offers mostly high-resolution images. The goal was to supplement the high-quality range in which *KonIQ-10k* is lacking.

Resolution and aspect ratio

Candidate images from both sources had be larger than 2048 × 1536px with aspect ratios between [1.315, 1.785] to retain similarity. We extracted image content at 2048 × 1536 px, 1024 × 768px and 512 × 384px. The first step was to crop the original images to an aspect ratio of 4:3. We selected the center part of the image for Pixabay, and used the smart-cropping [20] procedure for *KonIQ-10k*. The crops were then downsampled to 2048 × 1536px and the aforementioned lower resolutions using Lanczos interpolation. On the Flickr subset, this enforced identical image portions as published in the *KonIQ-10k* dataset at 1024 × 768px.

[6] https://flickr.com.

[7] https://pixabay.com.

Stratified attribute sampling

The sampling strategy relied on stratified discrete attributes, for which Flickr and Pixabay provided different tags and metadata. Our aim was to achieve uniform occurrence frequencies of unique attribute values (e.g., the camera model) and binned continuous attributes (e.g. *KonIQ-10k* MOS). We additionally included machine tags from [56] for the Flickr candidates. The pre-existing MOS from *KonIQ-10k* was quantized into 10 equal-width bins to fit into our discrete approach. For the Pixabay candidates, we considered the camera model, user-assigned tags and incorporated *normalized favorites* $\widetilde{F}(I)$. This measure is calculated as follows, where $F(I)$ is the number of "*favorites*" that an image I received on Pixabay by users of the platform, and $V(I)$ is the total number of times that image I was *viewed*:

$$\widetilde{F}(I) = \ln(F(I) + e) / \ln(V(I) + e) \tag{1}$$

On the admissible 7818 Flickr and 757.016 Pixabay images, we iterated the following procedure, thereby sampling 210 images from each source:

i) Randomly select a metadata attribute.
ii) Randomly select one of its manifestations.
iii) Keep the images corresponding to this choice.
iv) On this subset, continue alike with step i)

After all attributes have been considered, the procedure either returns a single image or a set of images from which we chose one image at random.

Subjective annotation study

In order to establish a benchmark that allows meaningful comparisons across resolutions, we had to design a reliable subjective study protocol, which we ensured by several means. Similar to the work presented in [57], we invited participants on freelancer.com. The candidates were pre-filtered based on their previous experience, mostly in photography or graphic design, and finally evaluated with regard to their *practical abilities to rate the quality of images*. They had to pass multiple tests in order to qualify for our main study.

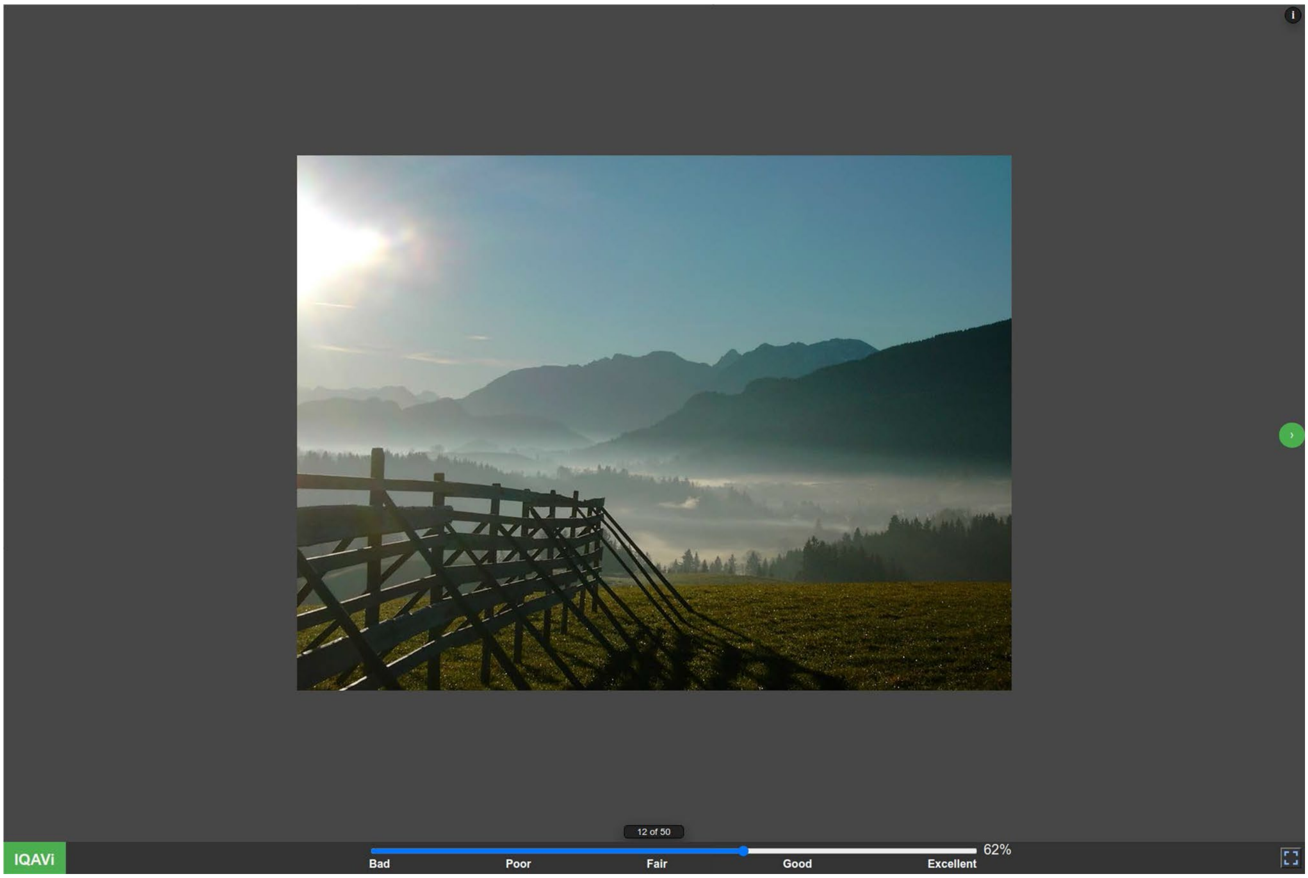

Fig. 5 The "Image Quality Assessment Viewer" (IQAVi) interface enforces a one-to-one rendering from image to screen pixels, thus presenting the lower-resolution versions on a smaller screen area. Ratings were given on a fine-grained scale from 1 to 100 with the slider below, which was labeled according to the standard ACR scale for easier orientation. Moving back to previous images was prohibited

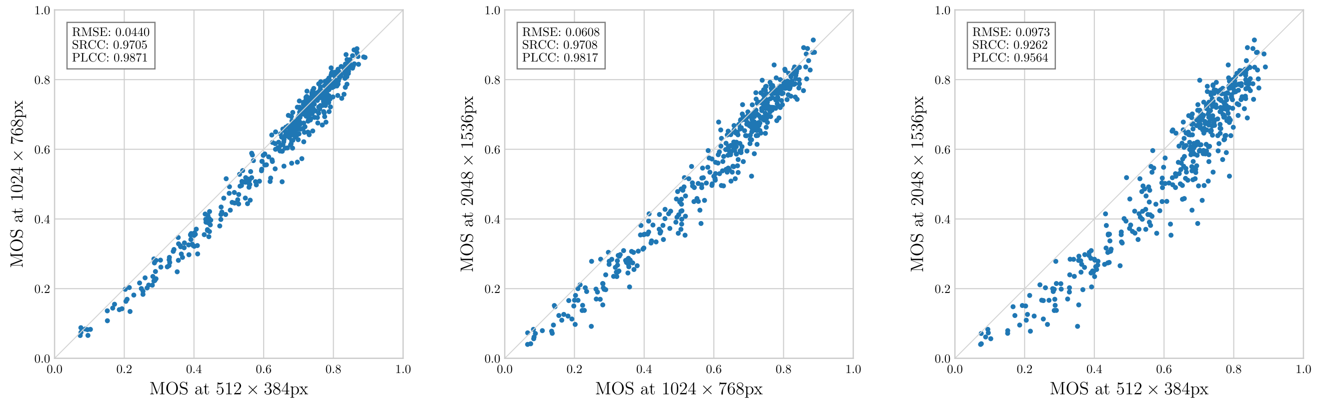

Fig. 6 Scatterplots of *KonX* MOS scores across annotation resolutions. The curvature is particularly pronounced in the right plot, where the resolution difference is the largest. The right-shift is higher for images in the middle of the quality range, indicating that the down-scaling had an overall positive effect. Lower-quality images in the lower-left corners of the plots were less affected. On the high-quality end, a few examples seem to excel when presented at the higher resolution and thus lie above the diagonal line

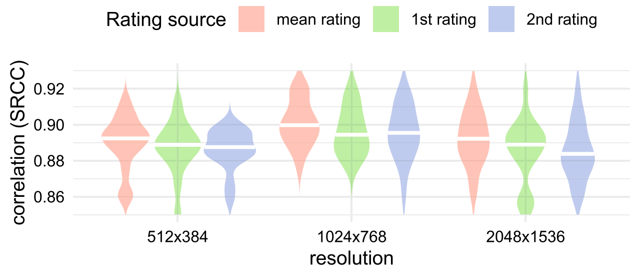

Fig. 7 Density of SRCCs of the *KonX* scores vs. the *KonIQ-10k* MOS. The horizontal white lines indicate their median. Overall, the mean of both *KonX* ratings outperforms the individual ratings

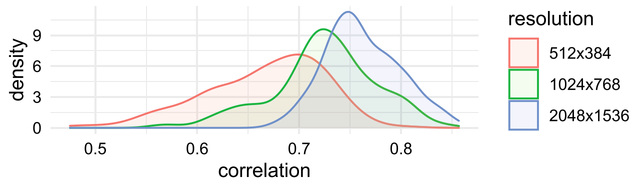

Fig. 8 Distribution of SRCCs between all participants in our study and how they depend on the presented image resolution. Agreements increase with the resolution, indicating that rating the quality of an image is easier at a larger resolution

Quality assessment UI

We developed a custom web interface that allows us to control the image presentation scale and thereby enables reproducible studies. It ensures that virtual image pixels are displayed as physical screen pixels in a 1:1 fashion. We account for devices where the *virtual resolution* used in the rendering stage differs from the actual *physical resolution* of the screen. Ratings were assigned through a slider on a scale from 1 to 100, which showed labels according to the standard absolute category rating (ACR) scheme. A depiction of our interface is given in Fig. 5.

Participant filtering

We conducted a qualifier experiment as a *contest* on `freelancer.com`. Instructions were given on how to identify distortions, judge the overall quality of an image, and interpret and use the rating scale. We carefully explained that judgments should be made independent of the image resolution, as larger presentations are not necessarily better in terms of quality. We required a screen diagonal size above 14 inches with a resolution of at least 1920 × 1080 pixels and rejected participants with smartphones and small tablets.

While most device checks were fully automated, additional information was gathered through participant self-reporting. We stored both the reported and the measured characteristics of all devices that were used in the study. Participation in a training phase was mandatory for all freelancers. It consisted of 50 images for which we had ground-truth ranges of acceptable quality ratings annotated by experts in a lab study. Upon failing to submit a rating within these bounds, we displayed the acceptable range and required the participant to retry until successful.

Training considerations

In the absence of absolute ground truth and given potentially flawed subjective ratings, one faces a bias-variance tradeoff when designing subjective studies (Fig. 6). The question of which votes to accept and which to reject as outliers is not easy to answer (Fig. 7). Strategies for bias compensation or even strict training of the participants might raise criticism from a rather philosophical perspective: how can one guarantee that the results are still representative of the

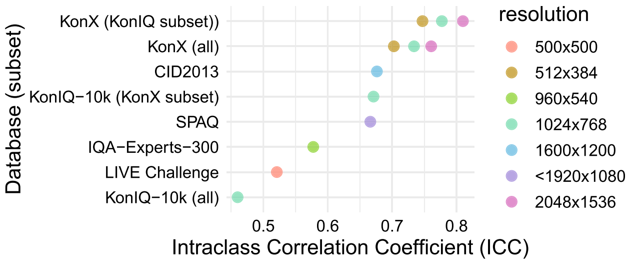

Fig. 9 Intraclass correlation scores (ICCs) [58] for authentically distorted IQA datasets. For *LIVE Challenge* and *SPAQ*, they are approximated based on the MOS and standard deviations and likely overestimated. The ICC is not always easily comparable across datasets, as it measures the fraction of the total variance accounted for by the per-image (intraclass) variance. Thus, the ICC tends to be larger for databases with a larger spread of the MOS

participant's judgments and not tailored to reflect the opinions of the researchers that designed the study? We believe that we found a suitable middle ground with lenient training and outlier detection through enforced self-consistency. This did not hinder participants in expressing their opinions (within reasonable bounds and while being consistent about it) and still led to high agreement among individuals, as shown, e.g., in Fig. 4 and Fig. 8.

We forced the participants to keep their browser window maximized during the study. In IQAVi, panning of the currently displayed image allows us to assess peripheral content if the virtual image resolution exceeds the physical screen resolution, so the participants with only FHD displays could also view the 2048 × 1536px images in their entirety. We logged the image area in view, the timestamps of annotations, and other interactions throughout the experiments for each participant individually.

Main annotation study

The images in the main study were presented in randomly ordered batches of 50. Each batch contained two repetitions of 25 images of a single resolution. Participants could not check their previous annotations to avoid fraudulent positive influences on their self-consistency. We required them to retry batches on which they failed to meet an SRCC of 0.9 between their two ratings, which serves as a consistency-enforcing outlier removal.

It was rarely necessary to repeat a batch, but when that was the case, almost all batches met the requirements after a single repetition. A participant was asked to repeat a specific batch at most once. The mean of both ratings for an image usually performs better than a single score, as confirmed by computing the correlation to the *KonIQ-10k* MOS scores, as shown in Fig. 7.

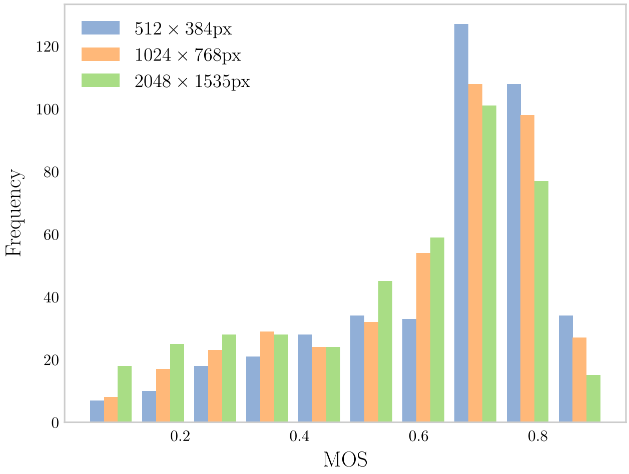

Fig. 10 Histogram of *KonX* MOS by resolution

Data analysis

Reliable, thus reproducible annotations are important for IQA datasets in general, but especially so for *KonX* due to its primary purpose as a benchmark. To characterize *KonX* and to compare it to other datasets, we consider a number of measures.

We plot the distribution of inter-user correlations in Fig. 8, measure the intraclass correlation coefficient (ICC) in Fig. 9, and investigate the SOS-hypothesis [54] in Fig. 4. The SOS hypothesis provides an indicator of reliability that accounts for the distribution of MOSes within a dataset. The central idea is that the range of possible ratings constrains their variance. If an image MOS is closer to the boundaries of the rating scale, its variance should be smaller than for a MOS at the center of the scale. The a coefficient of a parabola fitted to the variance vs. MOS plot indicates reliability. Larger a means a larger *SOS-normalized variance*, which implies less agreement between ratings. Figure 4 shows SOS plots for several databases, including subsets of *KonX* and *KonIQ-10k*.

The ICC(1, 1) coefficient, a one-way random effects single score model [58, 59], measures the absolute agreement between participants. This is reasonable, as we have to compare to datasets with only partial observations. The ICC is proportional to the variance of the image scores, which is related to the variance of the per-image MOS and roughly inversely proportional to the total variance of all ratings. It is thus sensible to compare ICCs on the same image subset. For the shared 210 images at 1024 × 768px, this indicates improved reliability for *KonX* over *KonIQ-10k* (Fig. 9).

Comparing *KonX* subsets by resolution suggests that larger images are rated more reliably with better agreement. Furthermore, the inter-user correlations in Fig. 8 also indicate that quality assessment might indeed be easier at higher

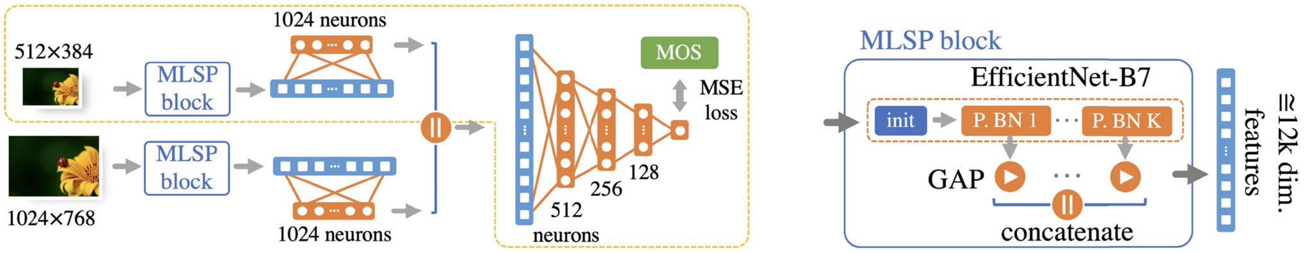

Fig. 11 The proposed `Effnet-2C-MLSP` two-column NR-IQA architecture. The yellow-dotted section on the left figure describes the single-column (`1C`) variants, `P.BN K` refers to the `project_bn` layers

resolutions. This probably is related to the larger difference in quality between the best and the worst images at high resolutions.

Label shifts

We display scatter plots of the MOSes of the same image contents compared by resolution in Fig. 6. They show curved trends, which match our hypotheses about the effects of down-scaling from Sect. 2.4 quite well. We observe a pronounced preference for the lower resolution in medium-quality images, resulting in the shift to the right. There are only a few samples at the low-quality end, but the plots indicate that there is a smaller difference in perceptual quality here, suggesting the images look bad regardless of their resolution.

We additionally plot the histograms for the MOS scores per resolution in Fig. 10. To formally confirm a statistically significant difference between the resolution-wise mean opinion scores in KonX, we conducted a Wilcoxon signed-rank test for all pairs of resolutions, a non-parametric alternative to the popular *t*-test. The results were significant with $p < 0.005$ for all pairs.

Summary

We conclude from this analysis that KonX is reliably annotated, especially in contrast to previous works. This is likely due to multiple factors, including the following design choices we made:

i) Usage of a quasi-continuous 100-point annotation scale instead of the traditional five-point absolute category rating (ACR) scale.
ii) Consistency checks of the participants, as all items were repeated twice in the study.
iii) Noise-reduction by averaging the repetitions for each participant individually.
iv) A high(er) level of control, especially by rendering image pixels 1:1 to screen pixels.

Cross-resolution prediction

Our model architecture is inspired by several observations from the literature regarding the properties of features from different CNN layers, their scale dependence, and their effect on transfer learning. Scale dependence is obvious for individual filters, meaning they can only detect fixed-size patterns. This is less evident for groups of filters or the usual cascades of convolutions used in deep CNNs. ImageNet models, for example, achieve a certain degree of scale-invariance of object classes only close to the last layers [28]. We considered multiple aspects:

Train-test scale discrepancy Object classification models that were trained closer to the test resolutions perform better after fine-tuning, which we expect to hold for IQA as well [60].

Scale-agnostic features Following the observations of Graziani et al. [28] on scale-invariance, the prevalent use of late-stage features could be suboptimal for quality assessment.

Multi-level binding The connection between the backbone and head network is traditionally based on the outputs of a single late-stage layer. Cross-task learning might be limited by this, as the success of multi-level features in well-performing architectures [5, 23] suggests.

Resolution overfitting Modern DNN architectures for NR-IQA accept one input size at a time. We found in our limited experiments that training such models on multiple resolutions did not improve their cross-resolution performance, on the contrary, it often decreased it. Learning scale-specific features on only one common network architecture seems to be a limitation of this approach, at least in practice, with limited time and training data.

NR-IQA model architecture

To get around these difficulties with our architecture, we made the following design choices:

- An EfficientNet-B7 [9] pre-trained at 600 × 600px serves as a backbone, which is close to our targeted resolutions

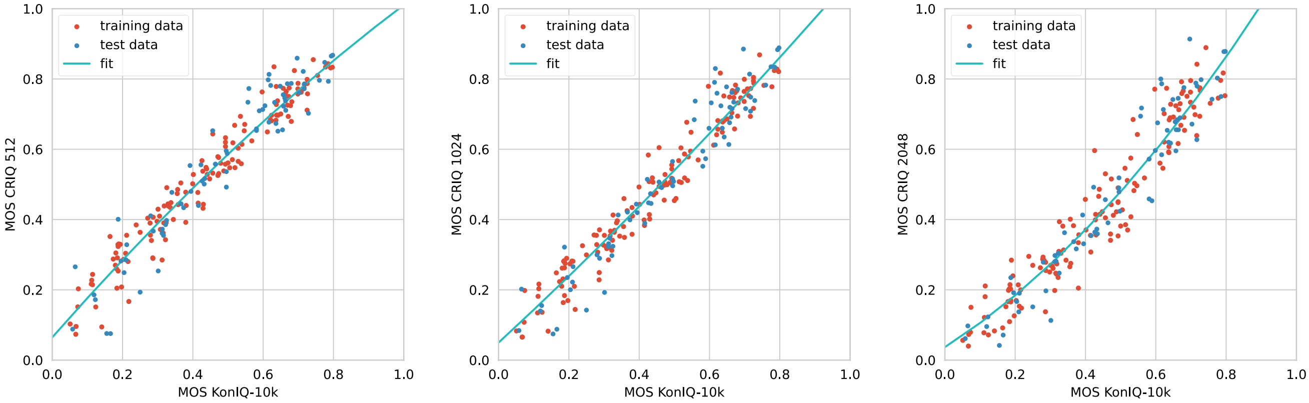

Fig. 12 Quadratic mapping from *KonIQ-10k* MOS to *KonX* at all three resolutions to align the scores for training at different resolutions on *KonIQ-10k* and evaluation on *KonX*. The blue markers were kept as a test set to determine the quality of the fit. At 1024 × 768px the scores are essentially just shifted

and has been shown to be tweakable regarding input scales [30].

- The Inception-MLSP approach from [23] gets adapted to EfficientNet by substituting Inception-module output activations with an inner layer of the EfficientNet-modules.
- We train a two-column network, similar to those used for scale-invariant detection [34–37], at different input resolutions. This enables the deep integration of column-wise MLSP-type features, synergizing with the proposed multi-level binding.

The proposed `Effnet-2C-MLSP` is depicted in Fig. 11. It consists of two columns (2C) of MLSP [23] blocks based on independent-weights EfficientNet-B7 backbones. These were pre-trained on ImageNet-1000 at 600 × 600px as a middle ground for the fine-tuning at 512 × 384px and 1024 × 768 px. Backbones pre-trained at higher resolutions, especially the version at 800 × 800px were not available at the time, and pre-training models ourselves was not feasible due to limits regarding training data and computational constraints.

Both columns feed into a cascaded multi-layer-perceptron (MLP) head. Features are sampled by global average pooling (GAP) the activations of the `project_bn` layers; this is different from Inception-MLSP features [23, 40] which stem from *mixed* layers. Their analog in ResNet-architectures would be the `add` layers at the end of each module, which are redundant due to the residual connections. Since the immediately preceding layers use dropout normalization, we extract the outputs from two layers before. In our preliminary experiments, neither the `add` nor the `dropout` activations performed better.

The `project_bn` features contain about 12000 scalar values, which we downsize to 1024 through separate dense layers for each column before passing them to the MLP head; the downsizing significantly reduces the number of parameters needed. This hierarchical combination allows for a greater level of per-scale differentiation of the column features through backpropagation compared to simply adding the features together. The models are trained to predict a single mean opinion score (MOS) directly, steered by the MSE loss.

Training data

KonX is now available as a test set, but there is no cross-resolution equivalent that is sufficiently large for training. Existing datasets [6, 7, 20, 42], for which each image was presented for rating at a single resolution[8] limit training to this respective annotation resolution. We can mitigate this shortcoming by exploiting a data overlap.

Fitting quadratic functions that map MOS scores from *KonIQ-10k* to each of the resolutions in *KonX* allows to align the scores between datasets and resolutions. We propose this as a better approximation of the underlying ground-truth labels than using the *KonIQ-10k*[9] scores for different resolutions directly. This adaptation reduces the MAE by 12.8% and the MSE by 20.3% over all three resolutions, as determined on a test set of 70 images that were not utilized in the curve fitting, as shown in Fig. 12.

We excluded the 210 images sampled for *KonX* from *KonIQ-10k* and created a 5-fold train/test split with the property that one of the test sets is a subset of the original *KonIQ-10k* test set. Each model under consideration is trained and evaluated on all folds. We report performance

[8] Paq-2-Piq [6] patches have to be considered as entirely different images because the placement of the patch sampling affects their perceptual quality.

[9] *KonIQ-10k* was annotated at 1024 × 768px.

Table 2 Cross database tests: training was conducted on *KonIQ-10k*, testing on the respective datasets

Models	*KonIQ-10k*		*Live challenge*		*SPAQ*	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
LinearityIQA	0.929	0.941	0.811	0.840	0.844	0.842
Effnet-NIMA	0.763	0.778	0.688	0.726	0.789	0.793
IRN-1C-MLSP	0.860	0.893	0.800	0.831	0.852	0.855
HyperIQA	0.906	0.917	0.801	0.805	0.843	0.841
Effnet-2C-MLSP	**0.949**	**0.959**	**0.832**	**0.859**	**0.864**	**0.864**

The results for `HyperIQA` on *KonIQ-10k* are from [32], the cross-tests for this model were performed using the official weights that were also trained on *KonIQ-10k*. Our proposed `Effnet-2C-MLSP` performs best across all databases and metrics, as indicated in bold font

indicators for each *KonX* subset in Table 3 and show cross-test results on other datasets in Table 2.

Training strategy

Training of `Effnet-2C-MLSP` was conducted in two stages. First, we kept the weights of the MLSP blocks fixed and trained just the head. This already achieves close to optimal performance and converges fast. In the second stage, we fine-tuned both columns jointly but did not update the batch normalization layers. Each stage is run for at most 40 epochs, with early stopping in 10 epochs if the validation loss does not improve.

The learning rates for the two stages were 10^{-5} and 10^{-4}, respectively. Incrementally fine-tuning one column at a time resulted in inferior results. The only augmentation we used was horizontal flipping of images, doing this independently per column improved performance marginally. We feed the entire image at a time. In our experiments, cropping the images did not provide a performance improvement.

Initial experiments with the Adam and SGD optimizers lead to unsatisfactory performance. The large resolutions and small batch sizes caused divergence, and the training loss increased rapidly after the first few epochs of the second stage. In order to reduce the effect of large gradients, we used gradient clipping (`clipnorm`=1.0), which worked well. We ultimately switched to the NAdam [61] optimizer with Nesterov momentum.

Model performance evaluation

Our `Effnet-2C-MLSP` was evaluated by feeding each column a different version of the same image: For the low-resolution column, images were always resized to 512 × 384px. The other column received the original image size. When testing on e.g. 2048 × 1536px *KonX* images, a downscaled 512 × 384px version was presented to the low-resolution column, and the 2048 × 1536px original to the other one. We cross-validated on 5-folds. The test sets are non-overlapping. The training database used was the remapped *KonIQ-10k*, after removing the 210 images that are shared with *KonX*. Thus, each set (training, validation, and test) is slightly smaller than the official splits published for *KonIQ-10k*.

We compare to previous works on *KonX* and the *KonIQ-10k* [1] test set as well as in cross-tests on LIVE-ITW [7] and SPAQ [42]. Table 3 shows correlations per subset, split by training and test resolution and data source. We trained and tested `KonCept-512` [1], `LinearityIQA` [5] and an `EfficientNet`-based derivative (ours) of `NIMA` [3] for an up to date comparison.

An ablation study on the backbone network selection is included in the table. The `EfficientNet-B7` was replaced in `IRN-2C-MLSP` with an `InceptionResNetV2`, which, as previously stated, was successfully used in many IQA-related experiments. As suggested by Fig. 3, this architecture suffers from cross-resolution discrepancies and is indeed outperformed by the `EfficientNet`-based architecture. An overview of the SRCC and MSE performances is given in Fig. 13, which shows that `Effnet-2C-MLSP` is highly performant, with respect to both its accuracy and correlations with the ground truth. `Effnet-2C-MLSP` also performs best when evaluated against the *KonIQ-10k* test set and across test sets on Live-ITW and SPAQ (at 1920 × 1080px) as shown in Table 2. Absolute error metrics (MSE) are crucial on *KonX*. The concentration of images at the top of the quality scale results in lower correlations on the Pixabay subset, making it more difficult to distinguish model performances. Nonetheless, our proposed model excels with regard to both metrics.

Table 3 Correlations on *KonX* subsets when training and testing at different resolutions

Model	Training Resolution	SRCC						PLCC					
		512 × 384px		1024 × 768		2048 × 1536		512 × 384px		1024 × 768		2048 × 1536	
		KonIQ-10k	Pixabay	*KonIQ-10k*	Pixabay	*KonIQ-10k*	Pixabay	*KonIQ-10k*	Pixabay	*KonIQ-10k*	Pixabay	*KonIQ-10k*	Pixabay
KonCept	512	0.880	0.304	0.826	0.270	0.682	0.311	0.853	0.304	0.752	0.267	0.601	0.269
	1024	0.825	0.265	0.888	0.417	0.816	0.451	0.696	0.265	0.884	0.420	0.842	0.492
Effnet-NIMA	512	0.850	0.310	0.764	0.373	0.550	0.401	0.835	0.368	0.766	0.411	0.592	0.397
	1024	0.856	0.250	0.884	0.318	0.818	0.392	0.844	0.310	0.884	0.389	0.842	0.450
LinearityIQA	512	**0.943**	0.381	0.911	0.399	0.761	0.448	**0.941**	0.468	0.906	0.467	0.793	0.485
	1024	0.914	0.384	**0.945**	0.451	0.902	0.493	0.908	0.431	0.943	0.481	0.911	0.529
IRN-1C-MLSP	512	0.927	0.319	0.909	0.349	0.807	0.450	0.927	0.415	0.904	0.435	0.832	0.496
	1024	0.894	0.311	0.932	0.419	0.907	0.503	0.899	0.400	0.931	0.487	0.916	**0.559**
HyperIQA	512	0.921	0.346	0.860	0.347	0.700	0.384	0.914	0.431	0.859	0.410	0.734	0.398
	1024	0.887	0.317	0.877	0.343	0.763	0.374	0.875	0.409	0.870	0.425	0.783	0.424
Effnet-2C-MLSP	512	0.927	0.395	0.905	0.445	0.790	0.514	0.924	0.468	0.903	0.506	0.825	0.539
	1024	0.891	0.376	0.935	**0.484**	0.910	**0.541**	0.895	0.444	0.936	**0.542**	0.922	0.585
	both	0.923	**0.405**	0.942	0.471	**0.927**	0.513	0.925	**0.478**	**0.943**	0.522	**0.932**	**0.559**

SRCC and PLCC are Spearman's rank correlation coefficient and the Pearson linear correlation coefficient. Our `Effnet-2C-MLSP` yields the best performance in most cases and is reassuringly close to `LinearityIQA` in columns 1, 3, and 7. The high performance of `LinearityIQA` likely stems from the correlation-based loss used to train the model, which seems detrimental to its accuracy, as shown in Fig. 13. The highest correlations per column are indicated in bold font

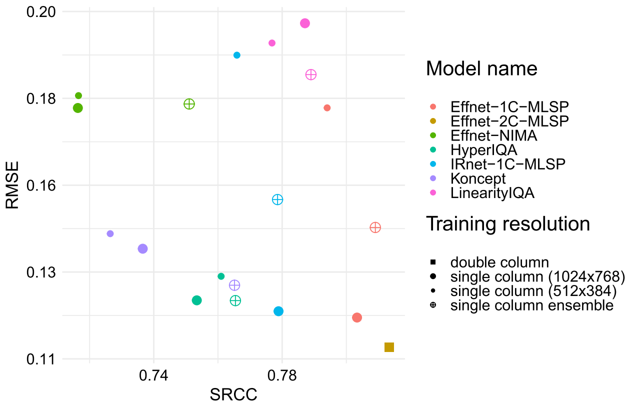

Fig. 13 RMSE vs. Spearman's rank correlation coefficient (SRCC) calculated jointly over the entirety of *KonX* on all resolutions. We report averages over all five splits. Through the RMSE, a key indicator of cross-resolution performance, this plot reveals biased but highly correlated predictions. We also report single resolution/column performance and that of ensembles made of two single-column predictors where the individual model's outputs are averaged. Our proposed `Effnet-2C-MLSP` has the highest SRCC and lowest RMSE as indicated by the square marker in the lower-right corner

Conclusions

This paper introduced the cross-resolution NR-IQA problem, which is a step toward assessing modern high-resolution images with computer vision models. We made significant progress in predicting the quality of authentically distorted images of various sizes. For that purpose, we introduced *KonX*, a benchmark dataset crafted specifically for cross-resolution IQA.

It includes 420 images from two source domains and is reliably annotated at three presentation resolutions through a subjective study. For the first time, this database allows to investigate cross-resolution effects independent of cross-content influences while also allowing for cross-domain experiments by splitting the data source. We additionally established a solid foundation for cross-resolution prediction with our `Effnet-2C-MLSP` model, which achieves state-of-the-art performance not only on *KonX*, but also when tested across databases.

As auxiliary results, we tapped into the importance of the pre-training resolution relative to the post-fine-tuning performance regarding scale-overfitting, the usage of multi-level features with varying levels of scale-variance and the application of column-wise multi-scale training in IQA. Considering these aspects surely helped, but they are far from being completely understood. Our work thus opens up new avenues for research in this field, such as developing computationally less intensive architectures and adapting advances in IQA to video quality assessment.

Acknowledgements Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 251654672 - TRR 161.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Hosu V, Lin H, Sziranyi T, Saupe D (2020) Koniq-10k: an ecologically valid database for deep learning of blind image quality assessment. IEEE Trans Image Process 29:4041–4056
2. Su S, Hosu V, Lin H, Zhang Y, Saupe D (2021) Koniq++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects. In: The 32nd British machine vision conference (BMVC)
3. Talebi H, Milanfar P (2018) NIMA: neural image assessment. IEEE Trans Image Process 27(8):3998–4011
4. Bosse S, Maniry D, Wiegand T, Samek W (2016) A deep neural network for image quality assessment. In: International conference on image processing (ICIP). IEEE, pp. 3773–3777
5. Li D, Jiang T, Jiang M (2020) Norm-in-Norm Loss with faster convergence and better performance for image quality assessment. Proceedings of the 28th ACM International conference on multimedia, 789–797. Accessed 2021-11-07, https://doi.org/10.1145/3394171.3413804. arXiv: 2008.03889
6. Ying Z, Niu H, Gupta P, Mahajan D, Ghadiyaram D, Bovik A (2020) From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3575–3585
7. Ghadiyaram D, Bovik A (2015) Live in the wild image quality challenge database
8. Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, Vozel B, Chehdi K, Carli M, Battisti F (2015) Image database tid2013: peculiarities, results and perspectives. Sign Process: Image Commun 30:57–77
9. Tan M, Le Q (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, pp. 6105–6114. PMLR
10. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618–626

11. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI Conference on artificial intelligence
12. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
13. Sheikh HR, Bovik AC (2005) A visual information fidelity approach to video quality assessment. In: The first international workshop on video processing and quality metrics for consumer electronics, vol. 7
14. Kang L, Ye P, Li Y, Doermann D (2014) Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 1733–1740
15. Wiedemann O, Hosu V, Lin H, Saupe D (2018) Disregarding the big picture: towards local image quality assessment. In: 2018 Tenth international conference on quality of multimedia experience (QoMEX), pp. 1–6. IEEE
16. Zhu H, Li L, Wu J, Dong W, Shi G (2020) Metaiqa: deep meta-learning for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp. 14143–14152
17. Yang S, Jiang Q, Lin W, Wang Y (2019) Sgdnet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment. In: Proceedings of the 27th ACM international conference on multimedia, pp. 1383–1391
18. Pan D, Shi P, Hou M, Ying Z, Fu S, Zhang Y (2018) Blind predicting similar quality map for image quality assessment. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 6373–6382
19. Golestaneh SA, Dadsetan S, Kitani KM (2022) No-reference image quality assessment via transformers, relative ranking, and self-consistency. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 1220–1230
20. Hosu V, Lin H, Sziranyi T, Saupe D (2020) Koniq-10k: an ecologically valid database for deep learning of blind image quality assessment. IEEE Trans Image Process 29:4041–4056
21. Van Noord N, Postma E (2017) Learning scale-variant and scale-invariant features for deep image classification. Patt Recogn 61:583–592
22. Hii Y-L, See J, Kairanbay M, Wong L-K (2017) Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs. In: International conference on image processing (ICIP), pp. 1722–1726. IEEE
23. Hosu V, Goldlucke B, Saupe D (2019) Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9375–9383
24. Ke J, Wang Q, Wang Y, Milanfar P, Yang F (2021) Musiq: multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 5148–5157
25. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The Thrity-seventh asilomar conference on signals, systems & computers, IEEE, vol. 2, pp. 1398–1402
26. Temel D, AlRegib G (2015) Persim: multi-resolution image quality assessment in the perceptually uniform color domain. In: 2015 IEEE International conference on image processing (ICIP), pp. 1682–1686. https://doi.org/10.1109/ICIP.2015.7351087
27. You J, Korhonen J (2021) Transformer for image quality assessment. In: 2021 IEEE International conference on image processing (ICIP), IEEE, pp. 1389–1393
28. Graziani M, Lompech T, Müller H, Depeursinge A, Andrearczyk V (2021) On the scale invariance in state of the art CNNs trained on ImageNet. Mach Learn Knowl Extr 3(2):374–391. https://doi.org/10.3390/make3020019
29. Liu W, Duanmu Z, Wang Z. End-to-end blind quality assessment of compressed videos using deep neural networks. In: ACM International Conference on Multimedia, pp. 546–554. ACM. https://doi.org/10.1145/3240508.3240643
30. Touvron H, Vedaldi A, Douze M, Jégou H (2020) Fixing the train-test resolution discrepancy: Fixefficientnet. arXiv preprint arXiv:2003.08237
31. Murray N, Marchesotti L, Perronnin F (2012) Ava: a large-scale database for aesthetic visual analysis. In: Computer vision and pattern recognition (CVPR). IEEE, pp. 2408–2415
32. Su S, Yan Q, Zhu Y, Zhang C, Ge X, Sun J, Zhang Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Computer vision and pattern recognition (CVPR), pp. 3664–3673. IEEE. https://doi.org/10.1109/CVPR42600.2020.00372. https://ieeexplore.ieee.org/document/9156687/
33. van Noord N, Postma E (2016) Learning scale-variant and scale-invariant features for deep image classification. Patt Recong 61:583–592. https://doi.org/10.1016/j.patcog.2016.06.005
34. Kang D, Chan AB (2018) Crowd counting by adaptively fusing predictions from an image pyramid. In: British machine vision conference (BMVC), p. 89
35. Zhang Y, Zhou D, Chen S, Gao S, Ma Y. Single-image crowd counting via multi-column convolutional neural network. In: Conference on computer vision and pattern recognition (CVPR), pp. 589–597. IEEE. https://doi.org/10.1109/CVPR.2016.70. http://ieeexplore.ieee.org/document/7780439/
36. Walach E, Wolf L (2016) Learning to count with cnn boosting. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) In: European conference on computer vision (ECCV). Lecture notes in computer science, vol. 9906, pp. 660–676. Springer. https://doi.org/10.1007/978-3-319-46475-6_41
37. Oñoro-Rubio D, López-Sastre RJ (2016) Towards perspective-free object counting with deep learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) In: European conference on computer vision (ECCV). Lecture notes in computer science, vol. 9911, pp. 615–629. Springer International Publishing. https://doi.org/10.1007/978-3-319-46478-7_38. Accessed 2022-02-27
38. Sheikh HR, Wang Z, Cormack L, Bovik AC (2005) Live image quality assessment database release 2. http://live.ece.utexas.edu/research/quality
39. Liu X, Pedersen M, Hardeberg JY (2014) Cid: Iq–a new image quality database. In: International conference on image and signal processing, pp. 193–202. Springer
40. Lin H, Hosu V, Saupe D (2019) Kadid-10k: a large-scale artificially distorted iqa database. In: International conference on quality of multimedia experience (QoMEX), IEEE, pp. 1–3
41. Lin H, Hosu V, Saupe D (2020) Deepfl-iqa: weak supervision for deep IQA feature learning. arXiv preprint arXiv:2001.08113
42. Fang Y, Zhu H, Zeng Y, Ma K, Wang Z (2020) Perceptual quality assessment of smartphone photography. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp. 3677–3686
43. Brunnström K, Beker SA, De Moor K, Dooms A, Egger S, Garcia M-N, Hossfeld T, Jumisko-Pyykkö S, Keimel C, Larabi M-C, et al (2013) Qualinet white paper on definitions of quality of experience
44. Reiter U, Brunnström K, De Moor K, Larabi M-C, Pereira M, Pinheiro A, You J, Zgank A. Factors Influencing Quality of Experience. In: Möller, S., Raake, A. (eds.) Quality of Experience. T-Labs Series in Telecommunication Services, pp. 55–72. Springer International Publishing
45. Moorthy AK, Bovik Choi LK, de Veciana ACG (2012) Video quality assessment on mobile devices: subjective, behavioral and

objective studies. IEEE J Select Topics Signal Process 6(6):652–671. https://doi.org/10.1109/JSTSP.2012.2212417

46. Gong R, Xu H (2013) Impacts of appearance parameters on perceived image quality for mobile-phone displays. Optik 125(11):2554–2559. https://doi.org/10.1016/j.ijleo.2013.10.092
47. Rehman A, Zeng K, Wang Z (2015) Display device-adapted video quality-of-experience assessment. In: Rogowitz, B.E., Pappas, T.N., de Ridder, H. (eds.) Human vision and electronic imaging XX, vol. 9394, pp. 27–37. SPIE https://doi.org/10.1117/12.2077917. International Society for Optics and Photonics
48. Zou W, Song J, Yang F (2016) Perceived image quality on mobile phones with different screen resolution. Mobile Inform 2016:1–17. https://doi.org/10.1155/2016/9621925
49. Sotelo R, Joskowicz J, Anedda M, Murroni M, Giusto DD (2017) Subjective video quality assessments for 4K UHDTV. In: 2017 IEEE International symposium on broadband multimedia systems and broadcasting (BMSB), pp. 1–6. IEEE. https://doi.org/10.1109/BMSB.2017.7986225
50. Kara PA, Robitza W, Pinter N, Martini MG, Raake A, Simon A (2019) Comparison of HD and UHD video quality with and without the influence of the labeling effect. Qual User Exper 4(1):4. https://doi.org/10.1007/s41233-019-0027-3
51. Saad MA, Pinson MH, Nicholas DG, Van Kets N, Van Wallendael G, Da Silva R, Jaladi RV, Corriveau PJ (2015) Impact of camera pixel count and monitor resolution perceptual image quality. In: 2015 Colour and Visual Computing Symposium (CVCS). IEEE, pp. 1–6
52. Rossi EA (2009) The Limits of Visual Resolution. Technical report, University of California, Berkeley
53. Kim YJ, Luo MR, Choe W, Kim HS, Park SO, Baek Y, Rhodes P, Lee S, Kim CY (2008) Factors affecting the psychophysical image quality evaluation of mobile phone displays: the case of transmissive liquid-crystal displays. J Opt Soci Am A 25(9):2215. https://doi.org/10.1364/JOSAA.25.002215
54. Hoßfeld T, Schatz R, Egger S (2011) Sos: The mos is not enough! In: 2011 Third International Workshop on Quality of Multimedia Experience, IEEE, pp. 131–136
55. Hosu V, Hahn F, Jenadeleh M, Lin H, Men H, Szirányi T, Li S, Saupe D (2017) The konstanz natural video database (konvid-1k). In: 2017 Ninth international conference on quality of multimedia experience (QoMEX). IEEE, pp. 1–6
56. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li L-J (2016) Yfcc100m: the new data in multimedia research. Commun ACM 59(2):64–73
57. Hosu V, Lin H, Saupe D (2018) Expertise screening in crowdsourcing image quality. In: QoMEX 2018: Tenth international conference on quality of multimedia experience
58. Hallgren KA (2012) Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quantit Meth Psychol 8(1):23
59. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86(2):420
60. Touvron H, Vedaldi A, Douze M, Jegou H (2019) Fixing the train-test resolution discrepancy. In: Advances in neural information processing systems, vol. 32. Curran Associates, Inc
61. Dozat T (2016) Incorporating nesterov momentum into adam. In: Proceedings of the 4th International conference on learning representations (ICLR)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.