

KonX: Cross-Resolution Image Quality Assessment

Oliver Wiedemann^{1*}[†], Vlad Hosu^{1†}, Shaolin Su^{1,2} and Dietmar Saupe¹

¹Department of Computer and Information Science, University of Konstanz, Germany.

²School of Computer Science and Engineering, Northwestern Polytechnical University, China.

*Corresponding author(s). E-mail(s): oliver.wiedemann@uni-konstanz.de;

Contributing authors: vlad.hosu@uni-konstanz.de; shaolin_su@mail.nwpu.edu.cn;
dietmar.saupe@uni-konstanz.de;

[†]These authors contributed equally to this work.

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 errors 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 DNN architecture that improves performance over previous state-of-the-art IQA models for this task, including recent transformers. We thus both raise and address a novel research problem in image quality assessment.

Keywords: Image Quality Assessment, Cross-Resolution Quality Prediction, IQA Models and Databases

1 Introduction

The discipline of image quality assessment (IQA) aims to model how humans perceive the quality of digital images. 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 as 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 their severity. 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 *image size in pixels*, which is to

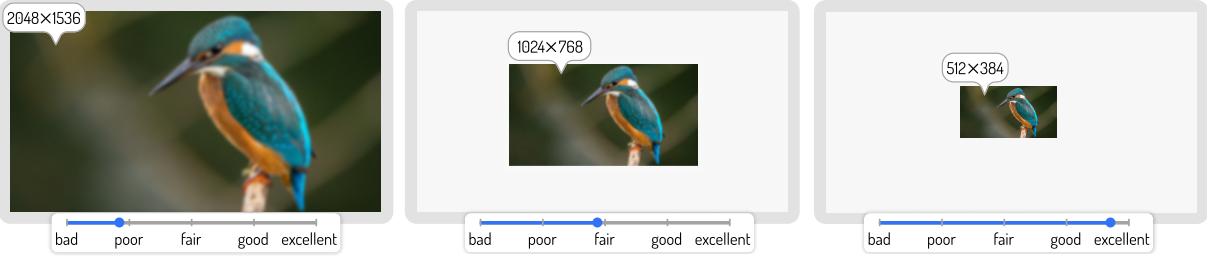


Fig. 1: Scaling affects both human perception and influences IQA model predictions.

be distinguished from *resolution as pixel densities*. On a display these are expressed in terms of dots or pixels per inch (DPI/PPI), whereas 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 across resolutions, we get only a 0.93 Spearman rank-order correlation coefficient (SRCC) between their mean opinion scores (MOS) when the scale ratio is 4:1, compared to a 0.97 SRCC when it is 2:1. 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 the prevalent approach is to train and test them 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.

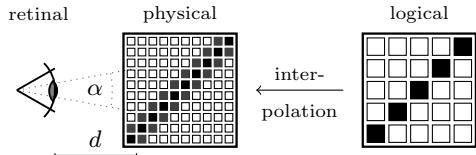


Fig. 2: The term **resolution** can be ambiguous. In this paper we use it for the **logical image size of $w \times h$ pixels**. Presenting an image on a screen, possibly interpolated, yields a *physical* resolution, which defines the image’s spatial dimensions and pixel density. What matters most for the human visual system is the perceivable *angular* resolution, which depends on the physical pixel density on the screen, the distance d to the screen, 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.

1.1 Contributions of this Work

We introduce a novel problem, create a database that allows us to approach it for the first time, propose a DNN architecture that surpasses the state of the art, and add validation considerations that allow proper comparisons of cross-resolution model performances. In greater detail:

1.1.1 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 and particularly for crowdsourced studies it is often unclear how well the annotation resolution was controlled for in the actual studies [1, 6–8].

1.1.2 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 multiple measures to achieve precise annotations:

- i) By inviting expert freelancers as participants.
- ii) By conducting a longitudinal study in which all items were rated twice, which provides valuable information about participant reliability, self-consistency and attention levels.
- iii) By controlling the presentation size. Our interface renders logical pixels 1:1 to screen pixels, which was not ensured for any previous NR-IQA dataset.

1.1.3 A DNN Architecture Proposal

In multi-column architectures, weights are usually shared between columns to limit the capacity and prevent overfitting. We employ a transfer-learning backbone in a multi-column architecture *with individual weights* that still does not overfit. The key is to feed different resolutions to each column and create a bottleneck before combining per-column features. We also integrate information from multiple levels of the network, i.e., from all pseudo-repeated modules of the EfficientNet backbone. These scale-variant features further improve the cross-resolution performance.

1.1.4 Validation Considerations

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

2 Related Work

2.1 IQA Models

Perceptual quality prediction evolved from statistical methods [11, 12] to an application area of deep learning. Most approaches crop or scale their input to a fixed, usually small resolution [3, 4, 13–18]. 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, 19]. In comparison to *full-reference* IQA scenarios, where one has access to both the distorted image and a usually pristine original, the performance of NR-IQA methods in cross-resolution and cross-database tests is significantly reduced, especially on [6, 7]. This is due to a more general problem in computer vision: scale variance [20], which in this case manifests itself as the *cross-resolution problem*.

Regarding model architectures we took inspiration from successful and recent works, 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 for example has shown success in image aesthetics assessment (IAA) [21, 22]. 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 even in comparison to transformer-based architectures [18, 23], in this case with SRCCs of 0.921 (KonCept-512) vs. 0.916 (MUSIQ) 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. Furthermore, it is unclear if transformers perform better in IQA than traditional CNNs, especially so for cross-resolution tasks.

Some works on full-reference IQA [24, 25] integrate information from downsampled versions of their input internally. However, they're only evaluated on predictions for a single fixed resolution, so they don't 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 [23, 26] have been made. We considered adding explicit information about the scale similar to [23], but [27] has shown that CNNs can infer the input dimensions by using the 0-padding that is added to images before convolution kernels 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, 23] and scale-free



Fig. 3: The cross-resolution problem: Grad-CAM [9] heatmaps depict aberrant regions-of-interest for the top predicted class of an InceptionResNetV2 [10]. Analogous difficulties in CNN-based IQA methods are even more delicate, as perceptual quality varies with scale, unlike object class labels.

rankings rather than absolute ratings [5, 28]. This work aims 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 alternatives even when the evaluation metric is Spearman’s rank correlation coefficient between predictions and ground-truth ratings [1], 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, and though it appears to converge faster in the early epochs, there is no clear overall advantage compared to the MSE.

2.2 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 newer architectures to generally improve performance, but multiple factors play a role. ImageNet CNNs are usually trained at small resolutions, many at 224×224 pixels, up to 800×800 for EfficientNet-L2 [29]. Pre-training on such small resolutions might limit the performance in large-resolution IQA. InceptionResNet-v2 was applied successfully in IAA [22] on AVA [30], an aesthetics database that contains images of various resolutions (up to 800×800). It outperformed other proposals in the past years since its introduction, which raises the question: *what*

makes this particular architecture more suitable for cross-resolution tasks?

Recent quality and aesthetics models [5, 22, 31] combine activations from multiple layers of pre-trained backbone models. Later-stage layers of ImageNet models usually represent abstract, scale-invariant concepts [27], 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 integrating information from multiple layers of an object classification network for IQA.

CNNs trained on a single resolution [27, 29] exhibit scale-wise overfitting, which can be mitigated by multi-resolution ensembles [32]. Multi-column architectures have shown success in crowd-counting [33–36], which involves varying object scales within single images. Again, this integrates information from multiple scales: [33] feed rescaled images to a shared-weight CNN column. Most crowd-counting works use directly trained custom architectures for the task, but we consider pre-trained networks as columns in hopes that they can jointly handle different scales.

2.3 Databases

IQA datasets are classified into two types: 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 single or in combinations [8, 37–39]. 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 on it have poor generalization to new impairments [40]. On the other hand, authentically distorted

IQA databases are usually sampled directly from online photography communities. The images are affected by mixtures of naturally occurring distortions. The state of the art for general authentically distorted IQA databases is currently *KonIQ-10k* [19], with 10,073 images. *SPAQ* [41] 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* [14] and extended through *Paq-2-Piq* [6]. They allow to compare the quality of patches with that 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* [41] 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 why IQA models perform poorly in cross-resolution (and cross-dataset) tests is studied thoroughly.

Our proposed dataset, *KonX*, allows 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 the cross-resolution testing. We anticipate that our work will pave the way for new directions in image quality research.

2.4 Subjective Factors in QoE

Previous studies in which existing IQA databases were annotated did not consider well-known aspects of *quality of experience* (QoE). Reiter et al. [42] 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 [43–49] report on the influence of the display device (System IF) on the perceived quality, especially regarding device characteristics.

The *visual resolution* [50] 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 [51].
- 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. [43] presented videos centered on mobile screens, while Gong et al. [44] resized images to ensure a constant physical size. On the other hand, Zou et al. [46] and Kara et al. [48] opted for full-screen, rescaled as needed. The source images were not always the same size as the screen resolution.

Rehman et al. [45] did not state what the presentation size was, 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 virtual resolutions that are smaller 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 [48] deemed it essential to be controlled, less so those on mobile and desktop devices [43, 46]. 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 scenarios, e.g., 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

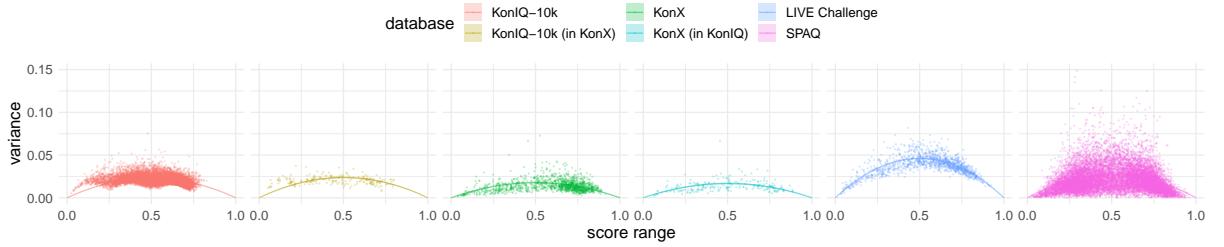


Fig. 4: Variance versus MOS of authentically distorted, crowdsourced datasets. The SOS-hypothesis a values for KonX, KonX scores at 1024×768 for the subset of images sampled from KonIQ-10k, KonIQ-10k, KonIQ-10k scores for the subset of images sampled for KonX, Live Challenge, and SPAQ are 0.071, 0.067, 0.091, 0.095, 0.184, and 0.107 respectively. The 95% confidence interval for a is indicated by the shaded region around the main curve.

this in crowdsourcing, but feeling uncomfortable might reduce the participants' ability to focus on the assessment task and negatively affect their judgments.

3 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 for the first time to investigate the relationship between perceived quality and scale.

3.1 Introduction Overview

KonX consists of 210 images from *Flickr*¹, which were already included in *KonIQ-10k* [1], and another 210 images from *Pixabay*² to supplement the high-quality range. The images were sampled using a stratified approach based on discretized metadata and other image properties. We aimed to diversify both their perceptual quality levels and contents. We center-cropped all Pixabay candidates, and smart-cropped [19] the KonIQ-10k original images to an aspect ratio of 4:3. These were then downsampled using the Lanczos-interpolation to three resolutions: 2048×1536 px, 1024×768 px and 512×384 px.

Eighteen freelancers³ 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 deployed a custom web interface that ensures a 1:1 rendering of virtual image pixels to physical screen pixels without scaling, thus displaying the lower-resolution images at a smaller spatial size. This experimental setup 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 in Table 1.

Table 1: KonX: A Cross-Res. IQA Benchmark

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

3.2 Content Preparation

When creating an image database, one of the main goals is to reduce potentially unknown biases, which stem from shared characteristics among images. This can be mitigated by enforcing *diversity* through adequate sampling strategies. Similar goals have been set for previous IQA [1] and VQA [52] datasets. We incorporated several means to diversify *KonX* with respect to perceptual quality as the primary attribute as well as auxiliary aspects such as image content.

¹<https://flickr.com>

²<https://pixabay.com>

³<http://freelancer.com>

3.2.1 Data Sources

We sampled from two online photography platforms: *Flickr*¹ and *Pixabay*². 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.

3.2.2 Resolution and Aspect Ratio

Candidate images from both sources had to be larger than 2048×1536 px and have aspect ratios between $[1.315, 1.785]$ to retain similarity. We extracted image content at 2048×1536 px, 1024×768 px and 512×384 px by cropping the original images to an aspect ratio of 4:3. We cropped the center part of the image for *Pixabay*, and used the smart-cropping [19] procedure for *KonIQ-10k*. The crops were then downsampled to 2048×1536 px 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 1024×768 px.

3.2.3 Stratified Attribute Sampling

Our sampling strategy relies on stratified discrete attributes, for which *Flickr* and *Pixabay* provide different tags and metadata. The occurrence frequencies of unique values were treated as “levels”, over which we aimed for uniformity. We additionally included machine tags from [53] for the *Flickr* candidates. The pre-existing MOSEs from *KonIQ-10k* were quantized into equal-length bins to fit into our discrete approach. For the *Pixabay* candidates, we considered the camera model, user-assigned tags and incorporated *normalized favorites* $\tilde{F}(I)$. This measure is calculated as follows, where $F(I)$ is the number of “favorites” that image I received on the *Pixabay* platform and $V(I)$ is the total number of times it was viewed:

$$\tilde{F}(I) = \ln(F(I) + e) / \ln(V(I) + e) \quad (1)$$

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 an attribute.
- ii) Randomly select one of its available “levels”.

- 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. In the latter case, we chose one image at random.

3.3 Subjective Annotation Study

In order to establish a benchmark that allows meaningful comparisons across resolutions, we had to design a *reliable* subjective study, which we ensured by several means. Similar to the work presented in [54], we invited participants on [freelancer.com](https://www.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.

3.3.1 Quality Assessment UI

We developed a custom web interface that allows to control the image presentation scale and thus 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 below in Fig. 5.



Fig. 5: Image quality assessment viewer (IQAVi).

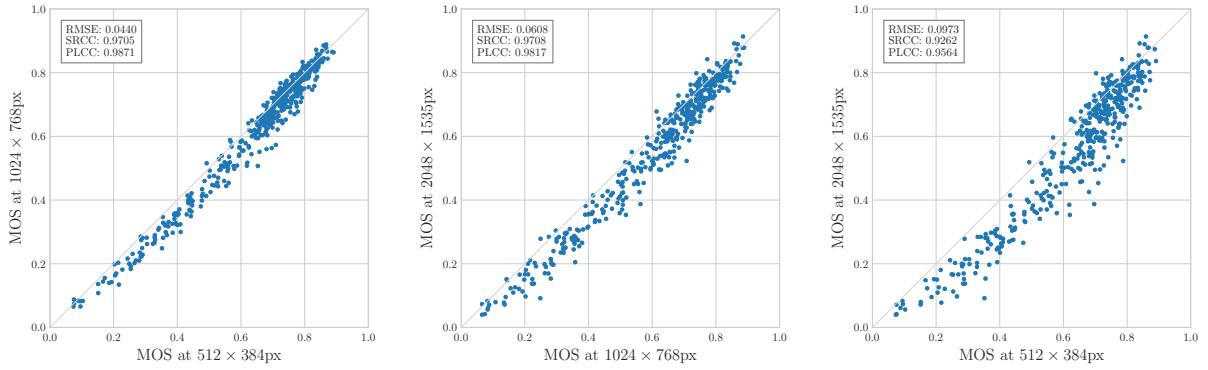


Fig. 6: Scatterplots of KonX MOS scores by annotation resolution.

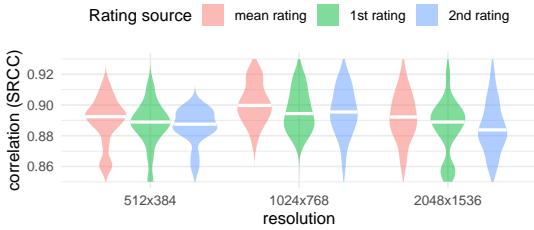


Fig. 7: Density of SRCCs of the KonX participant's scores vs the KonIQ-10k MOS. The horizontal white lines indicate their median.

3.3.2 Participant Filtering

We conducted a qualifier experiment as a *contest* on [freelancer.com](#). Instructions were given on how to identify distortions, how to judge the overall quality of an image and how to use the rating scale correctly. 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 self-reporting from the participants. 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 quality ratings. Upon failing to submit a rating within these bounds we displayed the range of *acceptable* values and users were required to retry until successful. We forced them to keep their browser window

maximized during the study. In IQAVi, panning of the currently displayed image allows assessing peripheral content if the image resolution exceeds that of the screen, so those with FHD displays could view the 2048×1536 images in their entirety. We logged the image area in view, as well as the timestamps of annotations and other interactions throughout the experiments for each participant individually.

3.3.3 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 effects on their self-consistency. We required them to retry batches on which they failed to meet a SRCC of 0.9 between their two ratings.

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 *KonIQ-10k* MOSes (Fig. 7).

3.4 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.

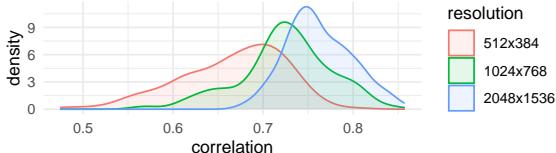


Fig. 8: Distribution of SRCCs between all participants in our study, and how they depend on the presented image resolution. Agreements increase with the resolution, indicating that rating the quality of an image is easier at a larger resolution.

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 [55] in Fig. 4. The SOS-hypothesis [55] provides an indicator of reliability that accounts for the distribution of MOSes within a dataset. The central point is that the variance of the ratings is constrained by their possible range. 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 serves as an indicator of 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 *KoniIQ-10k*. The ICC(1, 1) coefficient, a one-way random effects single score model [56, 57], measures the absolute agreement between

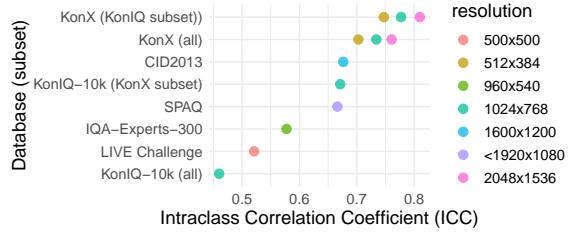


Fig. 9: ICCs [56] for authentically distorted IQA datasets. For *LIVE Challenge* and *SPAQ* they are approximated based on the MOS and standard-deviations and likely overestimated. The ICC is not always easily comparable across datasets, as it measures the fraction of the total variance accounted for by the per-image (intraclass) variance. Thus, the ICC tends to be larger for databases with a larger spread of the MOS.

participants. This is reasonable, as we have to compare datasets with partial observations. The ICC is proportional to the variance of the image scores, which is related to the variance of per-image MOSes 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 *KoniIQ-10k*, as shown in 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 resolutions. This probably is related with the larger difference in quality between the best and the worst images at high resolutions.

3.4.1 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 effects of down-scaling from Section 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 few samples at the low-quality end, but the plots indicate that there is a smaller difference in perceptual quality here, i.e. the images look bad regardless of their resolution.

We additionally plot the histograms for the MOS scores per resolution in Fig. 10. To formally

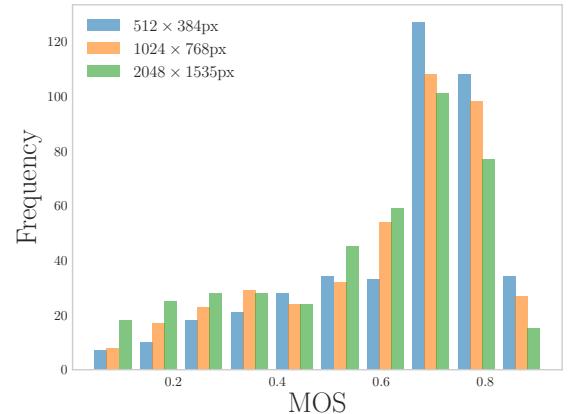


Fig. 10: Histogram of KonX MOS by resolution.

confirm that there exists 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, which is a non-parametric alternative to the popular t-test. The results were significant with $p < 0.005$ for all pairs.

3.4.2 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 fine-grained annotation scale instead of the traditional five-point ACR.
- 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.

4 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 that they can only detect patterns of a fixed size. 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 [27]. 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 [58].

Scale-Agnostic Features: Following the observations of Graziani et al. [27] 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, 22] 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.

4.1 NR-IQA Model Architecture

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

- An EfficientNet-B7 [59] pre-trained at 600×600 px serves as a backbone, which is close to our targeted resolutions and has been shown to be tweakable regarding input scales [29].
- The Inception-MLSP approach from [22] 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 [33–36], at different input resolutions. This enables the deep integration of column-wise MLSP-type features, synergizing with the proposed shallow-binding fix.

The proposed Effnet-2C-MLSP is depicted in Fig. 11. It consists of two columns (2C) of MLSP [22] blocks based on independent-weights EfficientNet-B7 backbones. These were pre-trained on ImageNet-1000 at 600×600 px as a middle ground for the fine-tuning at 512×384 px and 1024×768 px.

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 [22, 39] 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

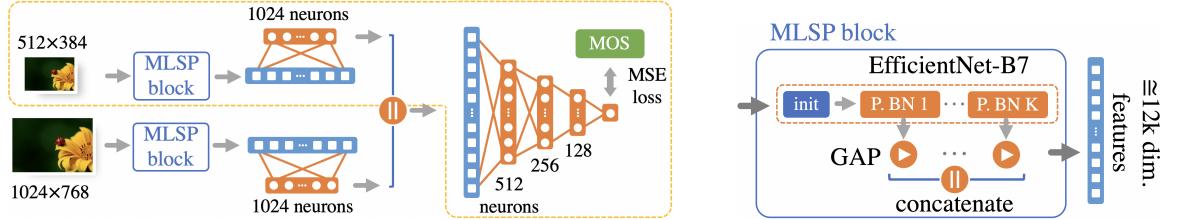


Fig. 11: The proposed Effnet-2C-MLSP two-column NR-IQA architecture. The yellow-dotted section on the left figure describes the single-column (1C) variants, P.BN K refers to the project_bn layers.

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.

4.2 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, 19, 41], for which each image was presented for rating at a single resolution⁴ 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*⁵ 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 indicators for each *KonX* subset in Table 3 and show cross-test results on other datasets in Table 2.

4.2.1 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 [60] optimizer with Nesterov momentum.

⁴Paq-2-Piq [6] patches have to be considered as entirely different images because the placement of the patch sampling affects their perceptual quality.

⁵KonIQ-10k was annotated at 1024×768 px.

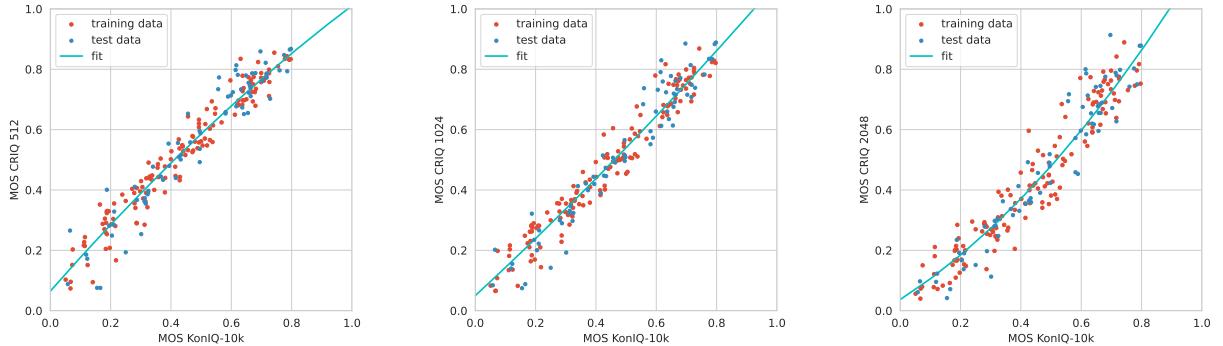


Fig. 12: Quadratic mapping from KonIQ-10k MOS to KonX at all three resolutions to align the scores for training at different resolutions on KonIQ-10k and evaluation on KonX. The blue markers were kept as a test-set to determine the quality of the fit. At $1024 \times 768\text{px}$ the scores are essentially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

4.2.2 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 $512 \times 384\text{px}$. The other column received the original image size. When testing on e.g. $2048 \times 1536\text{px}$ KonX images, a downsampled $512 \times 384\text{px}$ version was presented to the low-resolution column, and the $2048 \times 1536\text{px}$ 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 [41]. Table 3 shows correlations per subset, split by training and test resolution as well as 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 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 $1920 \times 1080\text{px}$) 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 on both metrics.

Models	KonIQ-10k		Live Challenge		SPAQ	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
LinearityIQA	0.9299	0.9415	0.8114	0.8404	0.8442	0.8422
Effnet-NIMA	0.7635	0.7788	0.6886	0.7269	0.7896	0.7936
IRN-1C-MLSP	0.8601	0.8932	0.8005	0.8310	0.8523	0.8553
Effnet-2C-MLSP	0.9490	0.9596	0.8327	0.8595	0.8641	0.8641

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

Model	Training Resolution	SRCC								PLCC							
		512 × 384px		1024 × 768		2048 × 1536		512 × 384px		1024 × 768		2048 × 1536					
		KoniQ	Pixabay	KoniQ	Pixabay	KoniQ	Pixabay										
KonCept	512	0.8807	0.3047	0.8264	0.2703	0.6821	0.3112	0.8535	0.3049	0.7522	0.2670	0.6016	0.2690	0.8251	0.2658	0.8888	0.4175
	1024	0.8251	0.2658	0.8888	0.4175	0.8165	0.4518	0.6968	0.2658	0.8845	0.4201	0.8420	0.4926				
Effnet-NIMA	512	0.8506	0.3101	0.7648	0.3739	0.5505	0.4010	0.8357	0.3682	0.7664	0.4118	0.5928	0.3972	0.8568	0.2506	0.8840	0.3184
	1024	0.8568	0.2506	0.8840	0.3184	0.8185	0.3925	0.8449	0.3105	0.8849	0.3895	0.8423	0.4503				
LinearityIQA	512	0.9436	0.3818	0.9111	0.3994	0.7611	0.4485	0.9416	0.4681	0.9068	0.4670	0.7933	0.4859	0.9141	0.3849	0.9452	0.4519
	1024	0.9141	0.3849	0.9452	0.4519	0.9023	0.4935	0.9087	0.4311	0.9435	0.4813	0.9115	0.5291				
IRN-1C-MLSP	512	0.9279	0.3197	0.9093	0.3490	0.8072	0.4501	0.9274	0.4155	0.9046	0.4355	0.8326	0.4967	0.8949	0.3117	0.9320	0.4190
	1024	0.8949	0.3117	0.9320	0.4190	0.9076	0.5037	0.8992	0.4003	0.9313	0.4876	0.9160	0.5596				
Effnet-2C-MLSP	512	0.9273	0.3955	0.9056	0.4457	0.7900	0.5149	0.9248	0.4689	0.9035	0.5063	0.8252	0.5391	0.8918	0.3762	0.9358	0.4844
	1024	0.8918	0.3762	0.9358	0.4844	0.9105	0.5415	0.8957	0.4443	0.9361	0.5422	0.9228	0.5857				
	both	0.9234	0.4058	0.9426	0.4715	0.9276	0.5132	0.9251	0.4783	0.9437	0.5220	0.9325	0.5596				

Table 3: Correlations on KonX subsets when training and testing at different resolutions. SRCC and PLCC is the Spearman’s Rank and Pearson linear correlation coefficient.

5 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, the database allows for the study of the effects of cross-resolution independent of cross-content, while also allowing for cross-domain experiments by splitting on 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 also 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.

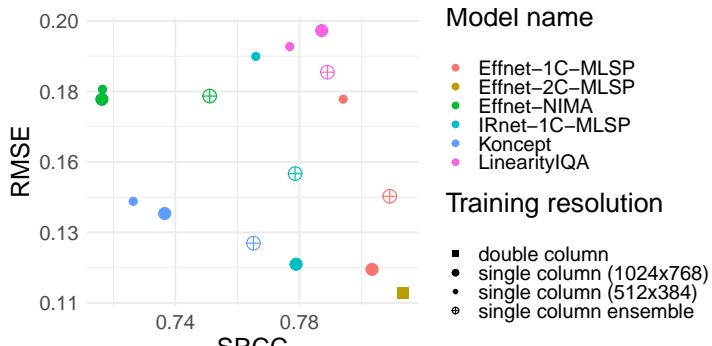


Fig. 13: RMSE vs. rank correlations (SRCC) were calculated jointly over the entirety of KonX on all resolutions. We report averages over all five splits. Through the RMSE, a key indicator of cross-resolution performance, this plot reveals biased but highly correlated predictions. We also report single resolution/column performance and that of ensembles made of two single-column predictors where the individual model’s outputs are averaged. Our Effnet-2C-MLSP has the highest SRCC and lowest RMSE.

Acknowledgements

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

References

- [1] Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* **29**, 4041–4056 (2020)
- [2] Su, S., Hosu, V., Lin, H., Zhang, Y., Saupe, D.: 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) (2021)
- [3] Talebi, H., Milanfar, P.: Nima: Neural image assessment. *IEEE Transactions on Image Processing* **27**(8), 3998–4011 (2018)
- [4] Bosse, S., Maniry, D., Wiegand, T., Samek, W.: A deep neural network for image quality assessment. In: International Conference on Image Processing (ICIP), pp. 3773–3777 (2016). IEEE
- [5] Li, D., Jiang, T., Jiang, M.: 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 (2020). <https://doi.org/10.1145/3394171.3413804>. arXiv: 2008.03889. Accessed 2021-11-07
- [6] Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: 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 (2020)
- [7] Ghadiyaram, D., Bovik, A.: Live in the Wild Image Quality Challenge Database (2015)
- [8] Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication* **30**, 57–77 (2015)
- [9] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
- [10] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)
- [11] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
- [12] Sheikh, H.R., Bovik, A.C.: 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 (2005). sn
- [13] Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1733–1740 (2014)
- [14] Wiedemann, O., Hosu, V., Lin, H., Saupe, D.: Disregarding the big picture: Towards local image quality assessment. In: 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2018). IEEE
- [15] Zhu, H., Li, L., Wu, J., Dong, W., Shi, G.: 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 (2020)

- [16] Yang, S., Jiang, Q., Lin, W., Wang, Y.: 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 (2019)
- [17] Pan, D., Shi, P., Hou, M., Ying, Z., Fu, S., Zhang, Y.: Blind predicting similar quality map for image quality assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6373–6382 (2018)
- [18] Golestaneh, S.A., Dadsetan, S., Kitani, K.M.: 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 (2022)
- [19] Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* **29**, 4041–4056 (2020)
- [20] Van Noord, N., Postma, E.: Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognition* **61**, 583–592 (2017)
- [21] Hii, Y.-L., See, J., Kairanbay, M., Wong, L.-K.: Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs. In: International Conference on Image Processing (ICIP), pp. 1722–1726. IEEE, ??? (2017)
- [22] Hosu, V., Goldlucke, B., Saupe, D.: Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9375–9383 (2019)
- [23] Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5148–5157 (2021)
- [24] Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thirtieth Asilomar Conference on Signals, Systems & Computers, 2003, vol. 2, pp. 1398–1402 (2003). Ieee
- [25] Temel, D., AlRegib, G.: Persim: Multi-resolution image quality assessment in the perceptually uniform color domain. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 1682–1686 (2015). <https://doi.org/10.1109/ICIP.2015.7351087>
- [26] You, J., Korhonen, J.: Transformer for image quality assessment. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 1389–1393 (2021). IEEE
- [27] Graziani, M., Lompech, T., Müller, H., Depeursinge, A., Andrearczyk, V.: On the Scale Invariance in State of the Art CNNs Trained on ImageNet **3**(2), 374–391. <https://doi.org/10.3390/make3020019>. Accessed 2021-11-23
- [28] 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>. <https://dl.acm.org/doi/10.1145/3240508.3240643> Accessed 2022-03-01
- [29] Touvron, H., Vedaldi, A., Douze, M., Jégou, H.: Fixing the train-test resolution discrepancy: Fixefficientnet. arXiv preprint arXiv:2003.08237 (2020)
- [30] Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: Computer Vision and Pattern Recognition (CVPR), pp. 2408–2415 (2012). IEEE
- [31] 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/](https://ieeexplore.ieee.org/document/9156687/) Accessed 2022-03-01
- [32] van Noord, N., Postma, E.: Learning scale-variant and scale-invariant features for deep image classification **61**, 583–592. <https://doi.org/10.1016/j.patcog.2016.06.005>. Accessed 2021-11-02
- [33] Kang, D., Chan, A.B.: Crowd counting by adaptively fusing predictions from an image pyramid. In: British Machine Vision Conference (BMVC), p. 89 (2018)
- [34] 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/> Accessed 2022-02-26
- [35] Walach, E., Wolf, L.: Learning to count with cnn boosting. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) 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. http://link.springer.com/10.1007/978-3-319-46475-6_41 Accessed 2022-02-27
- [36] Oñoro-Rubio, D., López-Sastre, R.J.: Towards perspective-free object counting with deep learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) 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. http://link.springer.com/10.1007/978-3-319-46478-7_38 Accessed 2022-02-27
- [37] Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: Live image quality assessment database release 2 (2005). URL <http://live.ece.utexas.edu/research/quality> (2005)
- [38] Liu, X., Pedersen, M., Hardeberg, J.Y.: Cid: Iq—a new image quality database. In: International Conference on Image and Signal Processing, pp. 193–202 (2014). Springer
- [39] Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–3 (2019). IEEE
- [40] Lin, H., Hosu, V., Saupe, D.: Deepfl-iqa: Weak supervision for deep iqa feature learning. arXiv preprint arXiv:2001.08113 (2020)
- [41] Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3677–3686 (2020)
- [42] 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
- [43] Moorthy, A.K., Choi, L.K., Bovik, A.C., de Veciana, G.: Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies **6**(6), 652–671. <https://doi.org/10.1109/JSTSP.2012.2212417>. Accessed 2021-11-03
- [44] Gong, R., Xu, H.: Impacts of appearance parameters on perceived image quality for mobile-phone displays **125**(11), 2554–2559. <https://doi.org/10.1016/j.ijleo.2013.10.092>. Accessed 2021-11-03
- [45] Rehman, A., Zeng, K., Wang, Z.: Display device-adapted video quality-of-experience assessment. In: Rogowitz, B.E., Papas, T.N., de Ridder, H. (eds.) Human Vision and Electronic Imaging XX, vol. 9394, pp. 27–37. SPIE, ??? (2015). <https://doi.org/10.1117/12.2077917>. International Society for Optics and Photonics. <https://doi.org/10.1117/12.2077917>
- [46] Zou, W., Song, J., Yang, F.: Perceived Image

- Quality on Mobile Phones with Different Screen Resolution **2016**, 1–17. <https://doi.org/10.1155/2016/9621925>. Accessed 2021-11-03
- [47] Sotelo, R., Joskowicz, J., Anedda, M., Murroni, M., Giusto, D.D.: 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>. <http://ieeexplore.ieee.org/document/7986225/> Accessed 2021-11-08
- [48] Kara, P.A., Robitz, W., Pinter, N., Martini, M.G., Raake, A., Simon, A.: Comparison of HD and UHD video quality with and without the influence of the labeling effect **4**(1), 4. <https://doi.org/10.1007/s41233-019-0027-3>. Accessed 2021-11-08
- [49] Saad, M.A., Pinson, M.H., Nicholas, D.G., Van Kets, N., Van Wallendael, G., Da Silva, R., Jaladi, R.V., Corriveau, P.J.: Impact of camera pixel count and monitor resolution perceptual image quality. In: 2015 Colour and Visual Computing Symposium (CVCS), pp. 1–6 (2015). IEEE
- [50] Rossi, E.A.: The Limits of Visual Resolution. Technical report, University of California, Berkeley (December 2009)
- [51] Kim, Y.J., Luo, M.R., Choe, W., Kim, H.S., Park, S.O., Baek, Y., Rhodes, P., Lee, S., Kim, C.Y.: Factors affecting the psychophysical image quality evaluation of mobile phone displays: the case of transmissive liquid-crystal displays. Journal of the Optical Society of America A **25**(9), 2215 (2008). <https://doi.org/10.1364/JOSAA.25.002215>. Accessed 2021-11-09
- [52] Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., Saupe, D.: The konstanz natural video database (konvid-1k). In: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2017). IEEE
- [53] Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.-J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
- [54] Hosu, V., Lin, H., Saupe, D.: Expertise screening in crowdsourcing image quality. In: QoMEX 2018: Tenth International Conference on Quality of Multimedia Experience (2018)
- [55] Hoßfeld, T., Schatz, R., Egger, S.: Sos: The mos is not enough! In: 2011 Third International Workshop on Quality of Multimedia Experience, pp. 131–136 (2011). IEEE
- [56] Hallgren, K.A.: Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology **8**(1), 23 (2012)
- [57] Shrout, P.E., Fleiss, J.L.: Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin **86**(2), 420 (1979)
- [58] Touvron, H., Vedaldi, A., Douze, M., Jegou, H.: Fixing the train-test resolution discrepancy. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc.
- [59] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR
- [60] Dozat, T.: Incorporating nesterov momentum into adam. In: Proceedings of the 4th International Conference on Learning Representations (ICLR) (2016)