From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality

Zhenqiang Ying[1*], Haoran Niu[1*], Praful Gupta[1], Dhruv Mahajan[2], Deepti Ghadiyaram[2†], Alan Bovik[1†]
[1]University of Texas at Austin, [2]Facebook AI
{zqying, haoranniu, praful_gupta}@utexas.edu, {dhruvm, deeptigp}@fb.com, bovik@ece.utexas.edu

Abstract

Blind or no-reference (NR) perceptual picture quality prediction is a difficult, unsolved problem of great consequence to the social and streaming media industries that impacts billions of viewers daily. Unfortunately, popular NR prediction models perform poorly on real-world distorted pictures. To advance progress on this problem, we introduce the largest (by far) subjective picture quality database, containing about $40,000$ *real-world distorted pictures and* $120,000$ *patches, on which we collected about* $4M$ *human judgments of picture quality. Using these picture and patch quality labels, we built deep region-based architectures that learn to produce state-of-the-art global picture quality predictions as well as useful local picture quality maps. Our innovations include picture quality prediction architectures that produce global-to-local inferences as well as local-to-global inferences (via feedback). The dataset and source code are available at* `https://live.ece.utexas.edu/research.php`.

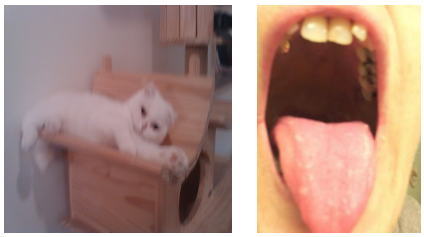
1. Introduction

Digital pictures, often of questionable quality, have become ubiquitous. Several hundred billion photos are uploaded and shared annually on social media sites like Facebook, Instagram, and Tumblr. Streaming services like Netflix, Amazon Prime Video, and YouTube account for 60% of all downstream internet traffic [1]. Being able to understand and predict the perceptual quality of digital pictures, given resource constraints and increasing display sizes, is a high-stakes problem.

It is a common misconception that if two pictures are impaired by the same amount of a distortion (e.g., blur), they will have similar perceived qualities. However, this is far from true because of the way the vision system processes picture impairments. For example, Figs. 1(a) and 1(b) have identical amounts of JPEG compression applied, but Fig. 1(a) appears relatively unimpaired perceptually, while Fig. 1(b) is unacceptable. On the other hand, Fig. 1(c) has had spatially uniform white noise applied to it, but its perceived distortion severity varies across the picture. The complex interplay between picture content and distortions (largely determined by masking phenomena [2]), and the way distortion artifacts are visually processed, play an important role in how visible or annoying visual distortions may present themselves. Perceived quality correlates poorly with simple quantities like resolution and bit rate [7]. Generally, predicting perceptual picture quality is a hard, long-standing research problem [8, 2, 7, 9, 10], despite its deceptive simplicity (we sense distortion easily with little, if any, thought).

(a) (b) (c)

Fig. 1: **Challenges in distortion perception:** Quality of a (distorted) image as perceived by human observers is *perceptual quality*. Distortion perception is highly content-dependent. Pictures (a) and (b) were JPEG compressed using identical encode parameters, but present very different degrees of perceptual distortion. The spatially uniform noise in (c) varies in visibility over the picture content, because of contrast masking [2].



(a) (b)

Fig. 2: **Aesthetics vs. perceptual quality** (a) is blurrier than (b), but likely more aesthetically pleasing to most viewers.

*†Equal contribution

Table 1: **Summary of popular IQA datasets.** In the legacy datasets, pictures were synthetically distorted with different types of single distortions. "In-the-wild" databases contain pictures impaired by complex mixtures of highly diverse distortions, each as unique as the pictures they afflict.

Database	# Unique contents	# Distortions	# Picture contents	# Patch contents	Distortion type	Subjective study framework	# Annotators	# Annotations
LIVE IQA (2003) [3]	29	5	780	0	single, synthetic	in-lab		
TID-2008 [4]	25	17	1700	0	single, synthetic	in-lab		
TID-2013 [4]	25	24	3000	0	single, synthetic	in-lab		
CLIVE (2016) [5]	1200	-	1200	0	in-the-wild	crowdsourced	8000	350K
KonIQ (2018) [6]	10K	-	10K	0	in-the-wild	crowdsourced	1400	1.2M
Proposed database	$39,810$	-	$39,810$	$119,430$	in-the-wild	crowdsourced	7865	$3,931,710$

It is important to distinguish between the concepts of *picture quality* [2] and *picture aesthetics* [11]. Picture quality is specific to perceptual distortion, while aesthetics also relates to aspects like subject placement, mood, artistic value, and so on. For instance, Fig. 2(a) is noticeably blurred and of lower perceptual quality than Fig. 2(b), which is less distorted. Yet, Fig. 2(a) is more aesthetically pleasing than the unsettling Fig. 2(b). While distortion can detract from aesthetics, it can also contribute to it, as when intentionally adding film grain [12] or blur (bokeh) [13] to achieve photographic effects. While both concepts are important, picture quality prediction is a critical, high-impact problem affecting several high-volume industries, and is the focus of this work. Robust picture quality predictors can significantly improve the visual experiences of social media, streaming TV and home cinema, video surveillance, medical visualization, scientific imaging, and more.

In many such applications, it is greatly desired to be able to assess picture quality at the point of ingestion, to better guide decisions regarding retention, inspection, culling, and all further processing and display steps. Unfortunately, measuring picture quality without a pristine *reference* picture is very hard. This is the case at the output of any camera, and at the point of content ingestion by any social media platform that accepts user-generated content (UGC). *No-reference* (NR) or blind picture quality prediction is largely unsolved, though popular models exist [14, 15, 16, 17, 18, 19, 20]. While these are often predicated on solid principles of visual neuroscience, they are also simple and computationally shallow, and fall short when tested on recent databases containing difficult, complex mixtures of real-world picture distortions [5, 6]. Solving this problem could affect the way billions of pictures uploaded daily are culled, processed, compressed, and displayed.

Towards advancing progress on this high-impact unsolved problem, we make several new contributions.

- **We built the largest picture quality database in existence**. We sampled hundreds of thousands of open source digital pictures to match the feature distributions of the largest use-case: pictures shared on social media. The final collection includes about $40,000$ real-world, unprocessed (by us) pictures of diverse sizes, contents, and distortions, and about $120,000$ cropped image patches of various scales and aspect ratios (Sec. 3.1, 3.2).
- **We conducted the largest subjective picture quality study to date.** We used Amazon Mechanical Turk to collect about $4M$ human perceptual quality judgments from almost $8,000$ subjects on the collected content, about four times more than any prior image quality study (Sec. 3.3).
- **We collected both picture and patch quality labels to relate local and global picture quality**. The new database includes about $1M$ human picture quality judgments and $3M$ human quality labels on patches *drawn from the same pictures*. Local picture quality is deeply related to global quality, although this relationship is not well understood [21], [22]. This data is helping us to learn these relationships and to better model global picture quality.
- **We created a series of state-of-the-art deep blind picture quality predictors**, that builds on existing deep neural network architectures. Using a modified ResNet [23] as a baseline, we (a) use patch and picture quality labels to train a region proposal network [24], [25] to predict both global picture quality and local patch quality. This model is able to produce better global picture quality predictions by learning relationships between global and local picture quality (Sec. 4.2). We then further modify this model to (b) predict spatial maps of picture quality, useful for localizing picture distortions (Sec. 4.3). Finally, we (c) innovate a local-to-global feedback architecture that produces further improved whole picture quality predictions using local patch predictions (Sec. 4.4). This series of models obtains state-of-the art picture quality performance on the new database, and transfer well – *without finetuning* – on smaller "in-the-wild" databases such as LIVE Challenge (CLIVE) [5] and KonIQ-10K [6] (Sec. 4.5).

2. Background

Image Quality Datasets: Most picture quality models have been designed and evaluated on three "legacy" databases: LIVE IQA [3], TID-2008 [4], and TID-2013 [26]. These datasets contain small numbers of unique, pristine images (~ 30) synthetically distorted by diverse types and amounts of single distortions (JPEG, Gaussian blur, etc.). They contain limited content and distortion diversity, and do not capture complex mixtures of distortions that often occur in real-world images. Recently, "in-the-wild" datasets such as CLIVE [5] and KonIQ-10K [6], have been introduced to attempt to address these shortcomings (Table 1).

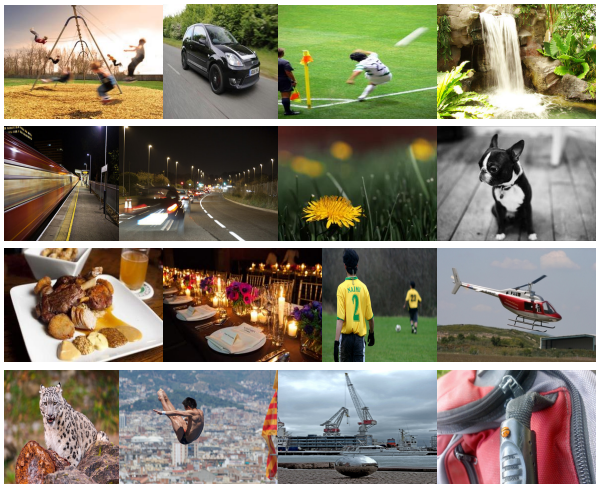


Fig. 3: **Exemplar pictures from the new database**, each resized to fit. Actual pictures are of highly diverse sizes and shapes.

Full-Reference models: Many *full-reference* (FR) perceptual picture quality predictors, which make comparisons against high-quality *reference* pictures, are available [9, 10], [27, 28, 29, 30, 31, 32, 33]. Although some FR algorithms (e.g. SSIM [9], [34], VIF [10], [35, 36]) have achieved remarkable commercial success (e.g. for monitoring streaming content), they are limited by their requirement of pristine reference pictures.

Current NR models arent general enough: *No-reference* or blind algorithms predict picture content without the benefit of a reference signal. Popular blind picture quality algorithms usually measure distortion-induced deviations from perceptually relevant, highly regular bandpass models of picture statistics [2], [37, 38, 39, 40]. Examples include BRISQUE [14], NIQE [15], CORNIA [17], FRIQUEE [16], which use "handcrafted" statistical features to drive shallow learners (SVM, etc.). These models produce accurate quality predictions on legacy datasets having single, synthetic distortions [3, 4, 26, 41], but struggle on recent in-the-wild [5, 6] databases.

Several deep NR models [42, 43, 44, 45, 46] have also been created that yield state-of-the-art performance on legacy synthetic distortion databases [3, 4, 26, 41], by pretraining deep nets [47, 48, 49] on ImageNet [50], then fine tuning, or by training on proxy labels generated by an FR model [45]. However, most deep models struggle on CLIVE [5], because it is too difficult, yet too small to sufficiently span the perceptual space of picture quality to allow very deep models to map it. The authors of [51], the code of which is not made available, reported high results, but we have been unable to reproduce their numbers, even with more efficient networks. The authors of [52] use a pretrained ResNet-101 and report high performance on [5, 6], but later disclosed [53] that they are unable to reproduce their results in [52].

3. Large-Scale Dataset and Human Study

Next we explain the details of the new picture quality dataset we constructed, and the crowd-sourced subjective quality study we conducted on it. The database has about $40,000$ pictures and $120,000$ patches, on which we collected 4M human judgments from nearly $8,000$ unique subjects (after subject rejection). It is significantly larger than commonly used "legacy" databases [3, 4, 26, 41] and more recent "in-the-wild" crowd-sourced datasets [5, 6].

3.1. UGC-like picture sampling

Data collection began by sampling about 40K highly diverse contents of diverse sizes and aspect ratios from hundreds of thousands of pictures drawn from public databases, including AVA [11], VOC [54], EMOTIC [55], and Blur Detection Dataset [56]. Because we were interested in the role of local quality perception as it relates to global quality, we also cropped three patches from each picture, yielding about 120K patches. While internally debating the concept of "representative," we settled on a method of sampling a large image collection so that it would be substantially "UGC-like." We did this because billions of pictures are uploaded, shared, displayed, and viewed on social media, far more than anywhere else.

We sampled picture contents using a mixed integer programming method [57] similar to [6], to match a specific set of UGC feature histograms. Our sampling strategy was different in several ways: firstly, unlike KonIQ [6], no pictures were down sampled, since this intervention can substantially modify picture quality. Moreover, including pictures of diverse sizes better reflects actual practice. Second, instead of uniformly sampling feature values, we designed a picture collection whose feature histograms match those of 15M that were randomly selected from unprocessed internal uploads to Facebook. This in turn resulted in a much more realistic and difficult database to predict features on, as we will describe later. Lastly, we did not use a pre-trained IQA algorithm to aid the picture sampling, as that could introduce *algorithmic bias* into the data collection process.

To sample and match feature histograms, we computed the following diverse, objective features on both our picture collection and the 15M UGC pictures:

- *absolute brightness* $L = R + G + B$.
- *colorfulness* using the popular model in [58].
- *RMS brightness contrast* [59].
- *Spatial Information(SI)*, the global standard deviation of Sobel gradients [60], a measure of complexity.
- *pixel count*, a measure of picture size.
- number of *detected faces* using [61].

In the end, we arrived at about 40K pictures. Fig. 3 shows 16 randomly selected pictures and Fig. 4 highlights the diverse sizes and aspect ratios of pictures in the new database.

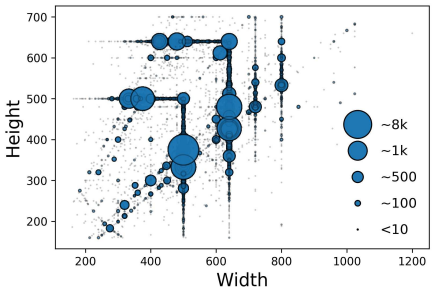


Fig. 4: **Scatter plot of picture width versus picture height** with marker size indicating the number of pictures for a given dimension in the new database.

3.2. Patch cropping

We applied the following criteria when randomly cropping out patches: (a) **aspect ratio:** patches have the same aspect ratios as the pictures they were drawn from. (b) **dimension:** the linear dimensions of the patches are 40%, 30%, and 20% of the picture dimensions. (c) **location:** every patch is entirely contained within the picture, but no patch overlaps the area of another patch cropped from the same image by more than 25%. Fig. 5 shows two exemplar pictures, and three patches obtained from each.

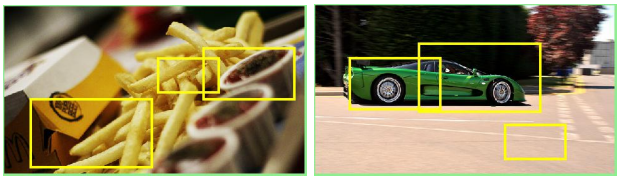


Fig. 5: Sample pictures and 3 randomly positioned crops (20%, 30%, 40%).

3.3. Crowdsourcing pipeline for subjective study

Subjective picture quality ratings are true psychometric measurements on human subjects, requiring 10-20 times as much time for scrutiny (per photo) as for example, object labelling [50]. We used the Amazon Mechanical Turk (AMT) crowdsourcing system, well-documented for this purpose [5, 6, 62, 63], to gather human picture quality labels.

We divided the study into two separate tasks: picture quality evaluation and patch quality evaluation. Most subjects (7141 out of 7865 workers) only participated in one of these, to avoid biases incurred by viewing both, even on different dates. Either way, the crowdsource workflow was the same, as depicted in Fig. 6. Each worker was given instructions, followed by a training phase, where they were shown several contents to learn the rating task. They then viewed and quality-rated N contents to complete their human intelligent task (HIT), concluding with a survey regarding their experience. At first, we set $N = 60$, but as the study accelerated and we found the workers to be delivering consistent scores, we set $N = 210$. We found that workers performed as well when viewing the larger number of pictures.

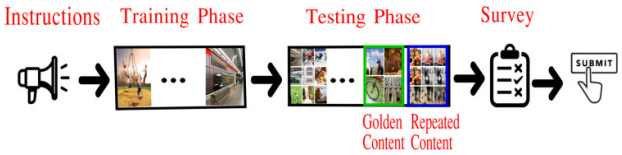


Fig. 6: **AMT task:** Workflow experienced by crowd-sourced workers when rating either pictures or patches.

3.4. Processing subjective scores

Subject rejection: We took the recommended steps [5, 63] to ensure the quality of the collected human data.

- We only accepted workers with **acceptance rates** $> 75\%$.
- **Repeated images:** 5 of the N contents were repeated randomly per session to determine whether the subjects were giving consistent ratings.
- **"Gold" images:** 5 out of N contents were "gold" ones sampled from a collection of 15 pictures and 76 patches that were separately rated in a controlled lab study by 18 reliable subjects. The "gold" images are not part of the new database.

We accepted or rejected each raters scores within a HIT based on two factors: the difference of the repeated content scores compared with overall standard deviation, and whether more than 50% of their scores were identical. Since we desired to capture many ratings, workers could participate in multiple HITs. Each content received at least 35 quality ratings, with some receiving as many as 50.

The labels supplied by each subject were converted into normalized Z scores [3], [5], averaged (by content), then scaled to [0, 100] yielding **Mean Opinion Scores (MOS)**. The total number of human subjective labels collected after subject rejection was $3,931,710$ ($950,574$ on images, and $2,981,136$ on patches).

Inter-subject consistency: A standard way to test the consistency of subjective data [3], [5], is to randomly divide subjects into two disjoint equal sets, compute two MOS on each picture (one from each group), then compute the Pearson linear correlation (LCC) between the MOS values of the two groups. When repeated over 25 random splits, the average LCC between the two groups MOS was $\mathbf{0.48}$, indicating the difficulty of the quality prediction problem on this realistic picture dataset. Fig. 7 (left) shows a scatter plot of the two halves of human labels for one split, showing a linear relationship and fairly broad spread. We applied the same process to the patch scores, obtaining a higher LCC of $\mathbf{0.65}$. This is understandable: smaller patches contain less spatial diversity; hence they receive more consistent scores. We also found that nearly all the non-rejected subjects had a positive Spearman rank ordered correlation (SRCC) with the golden pictures, validating the data collection process.

Relationships between picture and patch quality: Fig. 7 (right) is a scatter plot of the entire database of picture MOS against the MOS of the largest patches cropped from them.

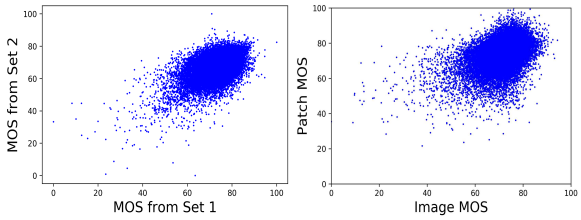


Fig. 7: **Scatter plots descriptive of the new subjective quality database**. Left: Inter-subject scatter plot of a random 50% divisions of the human labels of all 40K+ pictures into disjoint subject sets. Right: Scatter plot of picture MOS vs MOS of largest patch (40% of linear dimension) cropped from each same picture.

The linear correlation coefficient (LCC) between them is 0.43, which is strong, given that each patch represents only 16% of the picture area. The scatter plots of the picture MOS against that of the smaller (30% and 20%) patches are quite similar, with somewhat reduced LCC of 0.36 and 0.28, respectively (supplementary material).

An outcome of creating highly realistic "in-the-wild" data is that it is much more difficult to train successful models on. Most pictures uploaded to social media are of reasonably good quality, largely owing to improved mobile cameras. Hence, the distribution of MOS in the new database is narrower and peakier as compared to those of the two previous "in the wild" picture quality databases [5], [6]. This is important, since it is desirable to be able to predict small changes in MOS, which can be significant regarding, for example, compression parameter selection [64]. As we show in Sec. 4, the new database, which we refer to as the LIVE-FB Large-Scale Social Picture Quality Database, is very challenging, even for deep models.

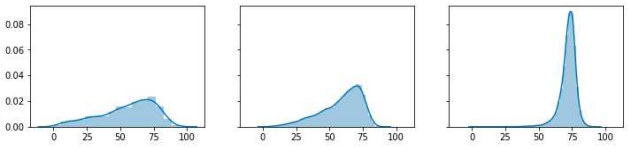


Fig. 8: **MOS (Z-score) histograms of three "in-the-wild" databases**. Left: CLIVE [5]. Middle: KonIQ-10K [6]. Right: The LIVE-FB database introduced here.

4. Learning Blind Picture Quality Predictors

With the availability of the new dataset comprising pictures and patches associated with human labels (Sec. 3), we created a series of deep quality prediction models, we collectively refer to as PaQ-2-PiQ, that exploit its unique characteristics. We conducted four picture quality learning experiments, evolving from a simple network into models of increasing sophistication and perceptual relevance which we describe next.

4.1. P2P-BM: A baseline picture-only model

To start with, we created a simple model that only processes pictures and the associated human quality labels. We will refer to this hereafter as the PaQ-2-PiQ Baseline Model, or P2P-BM for short. The basic network that we used is the well-documented pre-trained ResNet-18 [23], which we modified (described next) and fine-tuned to conduct the quality prediction task.

Input image pre-processing: Because picture quality prediction (whether by human or machine) is a psychometric prediction, it is crucial to not modify the pictures being fed into the network. While most visual recognition learners augment input images by cropping, resizing, flipping, etc., doing the same when training a perceptual quality predictor would be a psychometric error. Such input pre-processing would result in perceptual quality scores being associated with different pictures than they were recorded on.

The new dataset contains thousands of unique combinations of picture sizes and aspect ratios (see Fig. 4). While this is a core strength of the dataset and reflects its realism, it also poses additional challenges when training deep networks. We attempted several ways of training the ResNet on raw multi-sized pictures, but the training and validation losses were not stable, because of the fixed sized pooling and fully connected layers.

In order to tackle this aspect, we white padded each training picture to size 640×640, centering the content in each instance. Pictures having one or both dimensions larger than 640 were moved to the test set. This approach has the following advantages: (a) it allows supplying constant-sized pictures to the network, causing it to stably converge well, (b) it allows large batch sizes which improves training, (c) it agrees with the experiences of the picture raters, since AMT renders white borders around pictures that do not occupy the full webpage's width.

Training setup: We divided the picture dataset (and associated patches and scores) into training, validation and testing sets. Of the collected $39,810$ pictures (and $119,430$ patches), we used about 75% for training (30K pictures, along with their 90K patches), 19% for validation (7.7K pictures, 23.1K patches), and the remaining for testing (1.8K pictures, 5.4K patches). When testing on the validation set, the pictures fed to the trained networks were also white bordered to size 640×640. As mentioned earlier, the test set is entirely composed of pictures having at least one linear dimension exceeding 640. Being able to perform well on larger pictures of diverse aspect ratios was deemed as an additional challenge to the models.

Implementation Details: We used the PyTorch implementation of ResNet-18 [66] pre-trained on ImageNet and retained only the CNN backbone during fine-tuning. To this, we added two pooling layers (adaptive average pooling and adaptive max pooling), followed by two fully-connected (*FC*) layers, such that the final *FC* layer outputs a single score. We used a batch size of 120 and employed the MSE loss when regressing the single output quality score. We employed the Adam optimizer with $\beta_1 = .9$ and $\beta_2 = .99$, a

Table 2: **Patch quality predictions:** Results on (a) the largest patches (40% of linear dimensions), (b) middle-size patches (30% of linear dimensions) and (c) smallest patches (20% of linear dimensions) in the validation and test sets. Same protocol as used in Table 3.

	(a)				(b)				(c)			
	Validation		Test		Validation		Test		Validation		Test	
Model	**SRCC**	**LCC**	**SRCC**	**LCC**	**SRCC**	**LCC**	**SRCC**	**LCC**	**SRCC**	**LCC**	**SRCC**	**LCC**
NIQE [15]	0.109	0.106	0.251	0.271	0.029	0.011	0.217	0.109	0.052	0.027	0.154	0.031
BRISQUE [14]	0.384	0.467	0.433	0.498	0.442	0.503	0.524	0.556	0.495	0.494	0.532	0.526
CNNIQA [65]	0.438	0.400	0.445	0.373	0.522	0.449	0.562	0.440	0.580	0.481	0.592	0.475
NIMA [46]	0.587	0.637	0.688	0.691	0.547	0.560	0.681	0.670	0.395	0.411	0.526	0.524
P2P-BM (Sec. 4.1)	0.561	0.617	0.662	0.701	0.577	0.603	0.685	0.704	0.563	0.541	0.633	0.630
P2P-RM (Sec. 4.2)	0.641	0.731	0.724	0.782	0.686	0.752	0.759	0.808	0.733	0.760	0.769	0.792
P2P-FM (Sec. 4.4)	**0.658**	**0.744**	**0.726**	**0.783**	**0.698**	**0.762**	**0.770**	**0.819**	**0.756**	**0.783**	**0.786**	**0.808**

weight decay of .01, and do a full fine-tuning for 10 epochs. We followed a discriminative learning approach [67], using a lower learning rate of $3e^{-4}$, but a higher learning rate of $3e^{-3}$ for the head layers. These settings apply to all the models we describe in the following.

Evaluation setup: Although the P2P Baseline Model was trained on whole pictures, we tested it on both pictures and patches. For comparison with popular shallow methods, we also trained and tested BRISQUE [14] and the "completely blind" NIQE [15], which does not involve any training. We reimplemented two deep picture quality methods - NIMA [46] which uses a Mobilenet-v2 [68] (except we replaced the output layer to regress a single quality score), and CNNIQA [65], following the details provided by the authors. All of the compared models were trained over the same number of epochs on the LIVE-FB training set. As is the common practice in the field of picture quality assessment, we report two metrics: Spearman Rank Correlation Coefficient (SRCC) and Linear Correlation Coefficient (LCC).

Results: From Table 3, the first thing to notice is the level of performance attained by popular shallow models (NIQE [15] and BRISQUE [14]), which have the same feature sets. The unsupervised NIQE algorithm performed poorly, while BRISQUE did better, yet the reported correlations are far below desired levels. Despite being CNN-based, CNNIQA [65] performed worse than BRISQUE [14]. Our Baseline Model outperformed most methods and competed very well with NIMA [46]. The other entries in the table (the RoIPool and Feedback Models) are described later.

Table 3: **Picture quality predictions:** Performance of picture quality models on the full-size validation and test pictures in the LIVE-FB database. A higher value indicates superior performance. NIQE is not trained.

	Validation Set		**Testing Set**	
Model	**SRCC**	**LCC**	**SRCC**	**LCC**
NIQE [15]	0.094	0.131	0.211	0.288
BRISQUE [14]	0.303	0.341	0.288	0.373
CNNIQA [65]	0.259	0.242	0.266	0.223
NIMA [46]	0.521	0.609	0.583	0.639
P2P-BM (Sec. 4.1)	0.525	0.599	0.571	0.623
P2P-RM (Sec. 4.2)	0.541	0.618	0.576	0.655
P2P-FM (Sec. 4.4)	**0.562**	**0.649**	**0.601**	**0.685**

Table 2 shows the performances of the *same* trained, unmodified models on the associated picture patches of three reduced sizes (40%, 30% and 20% of linear image dimensions). The P2P Baseline Model maintained or slightly improved performance across patch sizes, while NIQE continued to lag, despite the greater subject agreement on reduced-size patches (Sec. 3.4). The performance of NIMA suffered as patch size decreased. Conversely, BRISQUE and CNNIQA improved as patch size decreased, although they were trained on whole pictures.

4.2. P2P-RM: A picture + patches model

Next, we developed a new type of picture quality model that leverages both picture and patch quality information. Our "RoIPool Model", or P2P-RM, is designed in the same spirit as Fast/Faster R-CNN [24, 25], which was originally designed for object detection. As in Fast-RCNN, our model has an *RoIPool* layer which allows the flexibility to aggregate at both patch and picture-sized scales. However, it differs from Fast-RCNN [24] in three important ways. First, instead of regressing for detecting bounding boxes, we predict full-picture and patch quality. Second, Fast-RCNN performs multi-task learning with two separate heads, one for image classification and another for detection. Our model instead shares a single head between patches and images. This was done to allow sharing of the "quality-aware" weights between pictures and patches. Third, while both heads of Fast-RCNN operate solely on features from RoI-pooled region proposals, our model pools over the entire picture to conduct global picture quality prediction.

Implementation details: As in Sec. 4.1, we added an RoIPool layer followed by two fully-connected layers to the pre-trained CNN backbone of ResNet-18. The output size of the RoIPool unit was fixed at 2×2. All of the hyperparameters are the same as detailed in Sec. 4.1.

Train and test setup: Recall that we sampled 3 patches per image and obtained picture and patch subjective scores (Sec. 3). During training, the model receives the following input: (a) image, (b) location coordinates `(left, top, right, bottom)` of all 3 patches and, (c) ground truth

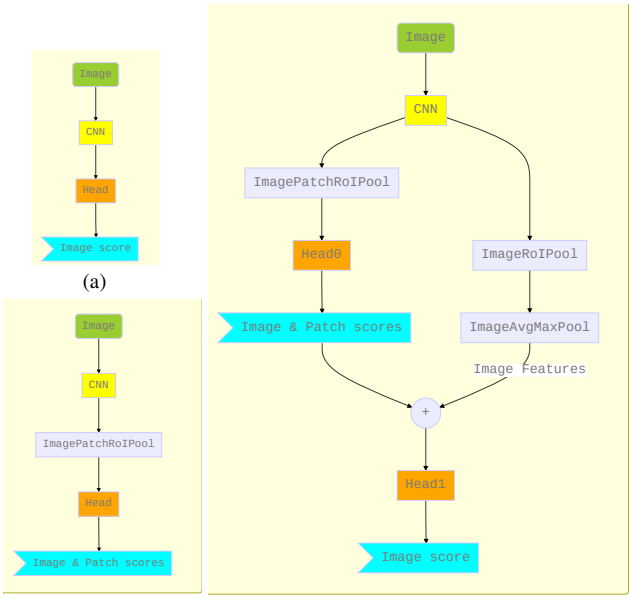


Fig. 9: Illustrating the different deep quality prediction models we studied. (a) **P2P Baseline Model:** ResNet-18 with a modified head trained on pictures (Sec. 4.1). (b) **P2P RoIPool Model:** trained on both picture and patch qualities (Sec. 4.2). (c) **P2P Feedback Model:** where the local quality predictions are fed back to improve global quality predictions (Sec. 4.4).

quality scores of the image and patches. At test time, the RoIPool Model can process both pictures and patches of any size. Thus, it offers the advantage of predicting the qualities of patches of any number and specified locations, in parallel with the picture predictions.

Results: As shown in Table 3, the RoIPool Model yields better results than the Baseline Model and NIMA on whole pictures on both validation and test datasets. When the same trained RoIPool Model was evaluated on patches, the performance improvement was more significant. Unlike the Baseline Model, the performance of the RoIPool Model increased as the patch sizes were reduced. This suggests that: (i) the RoIPool Model is more scalable than the Baseline Model, hence better able to predict the qualities of pictures of varying sizes, (ii) accurate patch predictions can help guide global picture prediction, as we show in Sec. 4.4, (iii) this novel picture quality prediction architecture allows computing local quality maps, which we explore next.

4.3. Predicting perceptual quality maps

Next, we used the P2P RoIPool Model to produce patch-wise quality maps on each image, since it is flexible enough to make predictions on any specified number of patches. This unique picture quality map predictor is the first deep model that is learned from true human-generated picture and patch labels, rather than from proxy labels delivered by an algorithm, as in [45]. We generated picture quality maps in the following manner: (a) we partitioned each picture into a 32×32 grid of non-overlapping blocks, thus preserving aspect ratio (this step can be easily extended to process denser, overlapping, or smaller blocks) (b) Each block's boundary coordinates were provided as input to the RoIPool to guide learning of patch quality scores (c) For visualization, we applied bi-linear interpolation to the block predictions, and represented the results as magma color maps. We α-blended the quality maps with the original pictures ($\alpha = 0.8$). From Fig. 10, we observe that the RoIPool Model is able to accurately distinguish regions that are blurred, washed-out, or poorly exposed, from high-quality regions. Such spatially localized quality maps have great potential to support applications like image compression, image retargeting, and so on.

4.4. P2P-FM: A local-to-global feedback model

As noted in Sec. 4.3, local patch quality has a significant influence on global picture quality. Given this, how do we effectively leverage local quality predictions to further improve global picture quality? To address this question, we developed a novel architecture referred to as the PaQ-2-PiQ Feedback Model, or P2P-FM (Fig. 9(c)). In this framework, the pre-trained backbone has two branches: (i) an RoIPool layer followed by an FC-layer for local patch and image quality prediction (`Head0`) and (ii) a global image pooling layer. The predictions from `Head0` are concatenated with the pooled image features from the second branch and fed to a new FC layer (`Head1`), which makes whole-picture predictions.

Fig. 10: **Spatial quality maps** generated using P2P-RM (Sec. 4.2). Left: Original Images. Right: Quality maps blended with the originals using magma color.

From Tables 2 and 3, we observe that the performance of the P2P Feedback Model on both pictures and patches is improved even further by the unique local-to-global feedback architecture. This model consistently outperformed all shallow and deep quality models. The largest improvement is made on the whole-picture predictions, which was the main goal. The improvement afforded by the Feedback Model is understandable from a perceptual perspective, since, while quality perception by a human is a low-level task involving low-level processes, it also involves a viewer casting their foveal gaze at discrete localized patches of the picture being viewed. The overall picture quality is likely an integrated combination of quality information gathered around each fixation point, similar to the Feedback Model.

Failure cases: While our model attains good performance on the new database, it does make errors in prediction. Fig 11(a) shows a picture that was considered of a very poor quality by the human raters (MOS=18), while the Feedback Model predicted an overrated score of 57, which is moderate. This may have been because the subjects were less forgiving of the blurred moving object, which may have drawn their attention. Conversely, Fig 11(b) is a picture that was underrated by our model, receiving a predicted score of 68 against the subject rating of 82. It may have been that the subjects discounted the haze in the background in favor of the clearly visible waterplane. These cases further reinforce the difficulty of perceptual picture quality prediction and highlight the strength of our new dataset.

4.5. Cross-database comparisons

Finally, we evaluated the P2P Baseline (Sec. 4.1), RoIPool (Sec. 4.2), and Feedback (Sec. 4.4) Models, and other baselines – all trained on the proposed dataset – on two other smaller "in-the-wild" databases CLIVE [5] and KonIQ-10k [6] without any fine-tuning. From Table 4, we may observe that all our three models, trained on the proposed dataset, transfer well to other databases. The Baseline, RoIPool, and Feedback Models all outperformed the shallow and other deep models [46, 65] on both datasets. This is a powerful result that highlights the representativeness of our new dataset and the efficacy of our models. The best reported numbers on both databases [69] uses a Siamese ResNet-34 backbone by training and testing on the same datasets (along with 5 other datasets). While this model reportedly attains 0.851 SRCC on CLIVE and 0.894 on KonIQ-10K, we achieved the above results by directly applying pre-trained models, thereby not allowing them to adapt to the distortions of the test data. When we also trained and tested on these datasets, our picture-based P2P Baseline Model also performed at a similar level, obtaining an SRCC of 0.844 on CLIVE and 0.890 on KonIQ-10K.

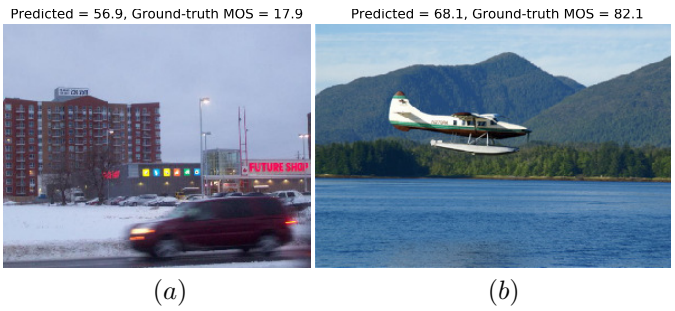


Fig. 11: **Failure cases:** Examples where the Feedback Model's predictions differed the most from the ground truth predictions.

Table 4: **Cross-database comparisons:** Results when models trained on the LIVE-FB database are applied on CLIVE [5] and KonIQ [6] **without fine-tuning.**

	Validation Set			
	CLIVE [5]		**KonIQ** [6]	
Model	**SRCC**	**LCC**	**SRCC**	**LCC**
NIQE [15]	0.503	0.528	0.534	0.509
BRISQUE [14]	0.660	0.621	0.641	0.596
CNNIQA [65]	0.559	0.459	0.596	0.403
NIMA [46]	0.712	0.705	0.666	0.721
P2P-BM (Sec. 4.1)	0.740	0.725	0.753	0.764
P2P-RM (Sec. 4.2)	0.762	**0.775**	0.776	0.794
P2P-FM (Sec. 4.4)	**0.784**	0.754	**0.788**	**0.808**

5. Concluding Remarks

Problems involving perceptual picture quality prediction are long-standing and fundamental to perception, optics, image processing, and computational vision. Once viewed as a basic vision science modelling problem to improve on weak Mean Squared Error (MSE) based ways of assessing television systems and cameras, the picture quality problem has evolved into one that demands the large-scale tools of data science and computational vision. Towards this end we have created a database that is not only substantially larger and harder than previous ones, but contains data that enables global-to-local and local-to-global quality inferences. We also developed a model that produces local quality inferences, uses them to compute picture quality maps, and global image quality. We believe that the proposed new dataset and models have the potential to enable quality-based monitoring, ingestion, and control of billions of social-media pictures and videos.

Finally, examples in Fig. 11 of competing local vs. global quality percepts highlight the fundamental difficulties of the problem of no-reference perceptual picture quality assessment: its subjective nature, the complicated interactions between content and myriad possible combinations of distortions, and the effects of perceptual phenomena like masking. More complex architectures might mitigate some of these issues. Additionally, mid-level semantic side-information about objects in a picture (e.g., faces, animals, babies) or scenes (e.g., outdoor vs. indoor) may also help capture the role of higher-level processes in picture quality assessment.

References

[1] *Sandvine*. The Global Internet Phenomena Report September 2019. [Online] Available: https://www.sandvine.com/global-internet-phenomena-report-2019.

[2] A. C. Bovik. Automatic prediction of perceptual image and video quality. *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008-2024, Sep. 2013.

[3] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440-3451, Nov 2006.

[4] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[5] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372-387, Jan 2016.

[6] H. Lin, V. Hosu, and D. Saupe. Koniq-10K: Towards an ecologically valid and large-scale IQA database. *arXiv preprint arXiv:1803.08489*, March 2018.

[7] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98-117, Jan 2009.

[8] J. Mannos and D. Sakrison. The effects of a visual fidelity criterion of the encoding of images. *IEEE Trans. Inf. Theor.*, vol. 20, no. 4, pp. 525–536, July. 1974.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004.

[10] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430-444, Feb 2006.

[11] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Int'l Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, June 2012.

[12] A. Norkin and N. Birkbeck. Film grain synthesis for AV1 video codec. In *Data Compression Conf. (DCC)*, Mar. 2018.

[13] Y. Yang, H. Bian, Y. Peng, X. Shen, and H. Song. Simulating bokeh effect with kinect. In *Pacific Rim Conf. Multimedia*, Sept. 2018.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[15] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "Completely blind" image quality analyzer. *IEEE Signal Processing Letters*, vol. 20, pp. 209-212, 2013.

[16] D. Ghadiyaram and A. C. Bovik. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, vol. 17, no. 1, art. 32, pp. 1-25, January 2017.

[17] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Int'l Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, pages 1098–1105, June 2012.

[18] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444-4457, Sep. 2016.

[19] K. Gu, G. Zhai, X. Yang, and W. Zhang. Using Free Energy Principle For Blind Image Quality Assessment. *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50-63, Jan 2015.

[20] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In *IEEE Int'l Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, pages 995–1002, June 2013.

[21] A. K. Moorthy and A. C. Bovik. Visual importance pooling for image quality assessment. *IEEE J. of Selected Topics in Signal Process.*, vol. 3, no. 2, pp. 193-201, April 2009.

[22] J. Park, S. Lee, and A.C. Bovik. VQpooling: Video quality pooling adaptive to perceptual distortion severity. *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 610-620, Feb. 2013.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vision and Pattern Recogn.*, pages 770–778, 2016.

[24] R. Girshick. Fast R-CNN. In *IEEE Int'l Conf. on Comput. Vision (ICCV)*, page 10401049, 2015.

[25] R. Girshick. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. Neural Info Process Syst (NIPS)*, 2015.

[26] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. . J. Kuo. Color image database TID2013: Peculiarities and preliminary results. In *European Workshop on Visual Information Processing*, volume vol. 30, pp. 106-111, pages 106–111, June 2013.

[27] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conf. Signals, Systems Comput.*, Pacific Grove, CA, Nov 2003.

[28] E. C. Larson and D. M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imag.*, vol. 19, no. 4, pp. 011006:1011006:21, Jan.Mar. 2010.

[29] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, Aug 2011.

[30] D. M. Chandler and S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284-2298, Sep. 2007.

[31] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684-695, Feb 2014.

[32] A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Process.: Image Comm.*, vol. 61, pp. 33-43, 2018.

[33] L. Zhang, Y. Shen, and H. Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270-4281, Oct 2014.

[34] Honorees announced for the 67th engineering emmy awards — television academy.

[35] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. D. Cock. VMAF: The Journey Continues, *The Netflix Tech Blog*. [Online] Available: `https://medium.com/netflix-techblog/vmaf-the-journey-continues-44b51ee9ed12`.

[36] M. Manohara, A. Moorthy, J. D. Cock, I. Katsavounidis, and A. Aaron. Optimized shot-based encodes: Now streaming!, *The Netflix Tech Blog*. [Online] Available: `https://medium.com/netflix-techblog/optimized-shot-based-encodes-now-streaming-4b9464204830`.

[37] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, vol. 4, no. 12, pp. 2379-2394, Dec 1987.

[38] D. L. Ruderman. The statistics of natural images. *Network: computation in neural systems*, vol. 5, no. 4, pp. 517–548, 1994.

[39] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[40] A.C. Bovik, M. Clark, and W.S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE Trans Pattern Anal. Machine Intell*, vol. 12, no. 1, pp. 5573, 1990.

[41] E. C. Larson and D. M. Chandler. Categorical image quality (CSIQ) database, 2010. [Online] Available: `http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=23`.

[42] D. Ghadiyaram and A. C. Bovik. Blind image quality assessment on real distorted images using deep belief nets. In *IEEE Global Conference on Signal and Information processing*, volume pp. 946–950, pages 946–950.

[43] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130-141, Nov 2017.

[44] S. Bosse, D. Maniry, T. Wiegand, and W. Samek. A deep neural network for image quality assessment. In *2016 IEEE Int'l Conf. Image Process. (ICIP)*, pages 3773–3777, Sep. 2016.

[45] J. Kim and S. Lee. Fully deep blind image quality predictor. *IEEE J. of Selected Topics in Signal Process.*, vol. 11, no. 1, pp. 206-220, Feb 2017.

[46] H. Talebi and P. Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998-4011, Aug 2018.

[47] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, Sept. 2014.

[48] X. Liu, J. van de Weijer, and A. D. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *IEEE Int'l Conf. on Comput. Vision (ICCV)*, page 10401049, 2017.

[49] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202-1213, March 2018.

[50] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vision and Pattern Recogn.*, pages 248–255, June 2009.

[51] S. Bianco, L. Celona, P. Napoletano, and R. Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, Feb 2018.

[52] D. Varga, D. Saupe, and T. Szirányi. DeepRN: A content preserving deep architecture for blind image quality assessment. *IEEE Int'l Conf. on Multimedia and Expo (ICME)*, pages 1–6, 2018.

[53] D. Saupe. `http://www.inf.uni-konstanz.de/~saupe`.

[54] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int'l J. of Comput. Vision*, pp. 303-338, June 2010.

[55] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. EMOTIC: Emotions in context dataset. In *IEEE Conf. Comput. Vision and Pattern Recogn. Workshops (CVPRW)*, July 2017.

[56] Blur detection dataset. [Online] Available: `http://www.cse.cuhk.edu.hk/~leojia/projects/dblurdetect/dataset.html`.

[57] V. Vonikakis, R. Subramanian, J. Arnfred, and S. Winkler. A probabilistic approach to people-centric photo selection and sequencing. *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2609-2624, Nov 2017.

[58] D. Hasler and S. E. Suesstrunk. Measuring colorfulness in natural images. In *SPIE Conf. on Human Vision and Electronic Imaging VIII*, 2003.

[59] Eli Peli. Contrast in complex images. *J. Opt. Soc. Am. A*, vol. 7, no. 10, pp. 2032–2040, Oct 1990.

[60] H. Yu and S. Winkler. Image complexity and spatial information. In *Int'l Workshop on Quality of Multimedia Experience (QoMEX)*, pages 12–17. IEEE, 2013.

[61] Face detection using haar cascades. *OpenCV-Python Tutorials*, [Online] Available: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_objdetect/py_face_detection/py_face_detection.html.

[62] M. J. C. Crump, J. V. McDonnell, and T. M. Gureckis. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLOS ONE*, vol. 8, pp. 1-18, March 2013.

[63] Z. Sinno and A. C. Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612-627, Feb 2019.

[64] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik. Predicting the quality of images compressed after distortion in two steps. *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5757-5770, Dec 2019.

[65] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Int'l Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, pages 1733–1740, June 2014.

[66] Torchvision.models. Pytorch. [Online] Available: https://pytorch.org/docs/stable/torchvision/models.html.

[67] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[68] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Int'l Conf. on Comput. Vision and Pattern Recogn. (CVPR)*, pages 4510–4520, June 2018.

[69] W. Zhang, K. Ma, G. Zhai, and X. Yang. Learning to blindly assess image quality in the laboratory and wild. *arXiv preprint arXiv:1907.00516*, 2019.