Re-IQA : Unsupervised Learning for Image Quality Assessment in the Wild

Sandeep Mishra* Avinab Saha* Alan C. Bovik
University of Texas at Austin
{sandy.mishra, avinab.saha}@utexas.edu, bovik@ece.utexas.edu

Abstract

Automatic Perceptual Image Quality Assessment is a challenging problem that impacts billions of internet, and social media users daily. To advance research in this field, we propose a Mixture of Experts approach to train two separate encoders to learn high-level content and low-level image quality features in an unsupervised setting. The unique novelty of our approach is its ability to generate low-level representations of image quality that are complementary to high-level features representing image content. We refer to the framework used to train the two encoders as Re-IQA. For Image Quality Assessment in the Wild, we deploy the complementary low and high-level image representations obtained from the Re-IQA framework to train a linear regression model, which is used to map the image representations to the ground truth quality scores, refer Figure 1. Our method achieves state-of-the-art performance on multiple large-scale image quality assessment databases containing both real and synthetic distortions, demonstrating how deep neural networks can be trained in an unsupervised setting to produce perceptually relevant representations. We conclude from our experiments that the low and high-level features obtained are indeed complementary and positively impact the performance of the linear regressor. A public release of all the codes associated with this work will be made available on GitHub.

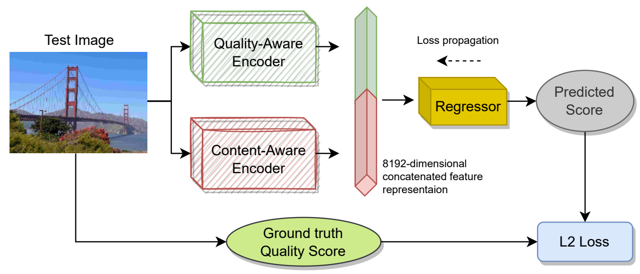


Figure 1. IQA score prediction uses two encoders trained for complementary tasks of learning content and quality aware image representations. The encoders are frozen while the regressor learns to map image representations to quality predictions.

1. Introduction

Millions of digital images are shared daily on social media platforms such as Instagram, Snapchat, Flickr, etc. Making robust and accurate Image Quality Assessments (IQA) that correlate well with human perceptual judgments is essential to ensuring acceptable levels of visual experience. Social media platforms also use IQA metrics to decide parameter settings for post-upload processing of the images, such as resizing, compression, enhancement, etc. In addition, predictions generated by IQA algorithms are often used as input to recommendation engines on social media platforms to generate user feeds and responses to search queries. Thus, accurately predicting the perceptual quality of digital images is a high-stakes endeavor affecting the way billions of images are stored, processed, and displayed to the public at large.

IQA metrics can be simply categorized into Full-Reference (FR) and No-Reference (NR) algorithms. FR-IQA algorithms like SSIM [31], FSIM [39], and LPIPS [40] require both reference (undistorted) and distorted version of an image to quantify the human-perceivable quality. This requirement limits their applicability for the "Images in the Wild" scenario, where the reference image is unavailable. On the contrary, NR-IQA algorithms like BRISQUE [16], PaQ-2-PiQ [36], and CONTRIQUE [13] do not require a reference image nor any knowledge about the kind of present distortions to quantify human-perceivable quality in a test image, paving the way for their use in "Images in the Wild" scenarios.

No-Reference IQA for "Images in the Wild" presents exciting challenges due to the complex interplay among the various kinds of distortions. Furthermore, due to the intricate nature of the human visual system, image content affects quality perception. In this work, we aim to learn low-level quality-aware image representations that are complementary to high-level features representative of image content. Figure 2 illustrates some of the challenges encountered

*Equal Contribution

(a) JPEG Compressed : 1 (b) JPEG Compressed : 2 (c) Motion Blur - Camera Shake (d) Overlaid Film Grain/ Noise

Figure 2. Exemplar Synthetically and "In the Wild" distorted pictures. (a), (b) are two images captured on iPhone an 13 Pro and then JPEG compressed using the same encoding parameters. (c), (d), were taken from KonIQ and AVA datasets respectively, and exhibit typical "Images in the Wild" distortions. Best viewed when zoomed in.

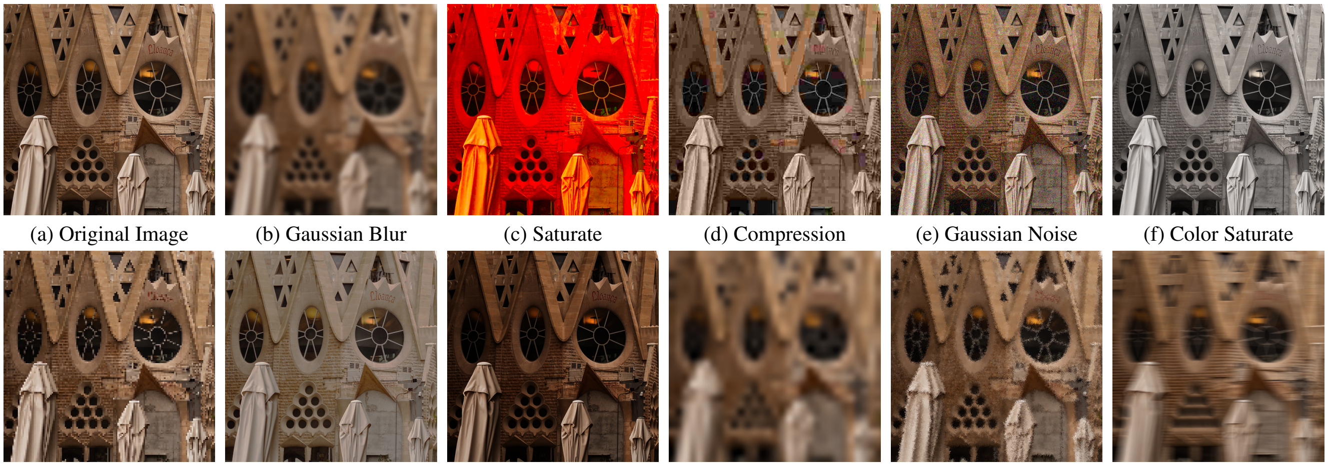
in the development of NR-IQA algorithms. Figures 2 (a-b) show two images captured by the authors on an iPhone 13 Pro and compressed using the same encoding parameters. While any distortions are almost negligible in Figure 2 (a), there are artifacts that are clearly visible in Figure 2 (b). As in these examples, it is well known that picture distortion perception is content dependent, and is heavily affected by content related perceptual processes like masking [1]. Figures 2 (c-d) illustrates a few distorted pictures "In the Wild". Figures 2 (c-d) show two exemplar distorted pictures, one impaired by motion blur (Figure 2 (c)) and the other by film grain noise (Figure 2 (d)). It is also well established that perceived quality does not correlate well with image metadata like resolution, file size, color profile, or compression ratio [32]. Because of all these factors and the essentially infinite diversity of picture distortions, accurate prediction of the perception of image quality remains a challenging task, despite its apparent simplicity, and hence research on this topic remains quite active [13, 14, 16, 25, 31, 35, 38–40].

Our work is inspired by the success of momentum contrastive learning methods [2, 5] in learning unsupervised representations for image classification. In this work, we engineer our models to learn content and quality-aware image representations for NR-IQA on real, authentically distorted pictures in an unsupervised setting. We adopt a Mixture of Experts approach to independently train two encoders, each of which accurately learns high-level content and low-level image quality features. We refer to the new framework as Re-IQA. The key contributions we make are as follows:

- We propose an unsupervised low-level image quality representation learning framework that generates features complementary to high-level representations of image content. We demonstrate how the "Mixture" of the two enables Re-IQA to produce image quality predictions that are highly competitive with existing state-of-the-art traditional, CNN and Transformer based NR-IQA models, developed in both supervised and unsupervised settings, across several contemporary databases.
- We demonstrate the superiority of high-level representations of image content for the NR-IQA task, obtained from the unsupervised pre-training of the ResNet-50 [6] encoder over the features obtained from supervised pre-trained ResNet-50 on the ImageNet database [3]. We learn these high-level representations of image content using the unsupervised training framework proposed in MoCo-v2 [2]
- Inspired by the principles of visual distortion perception we propose a novel Image Augmentation and Intra-Pair Image Swapping scheme to enable learning of low-level image quality representations. The dynamic nature of the image augmentation scheme prevents the learning of discrete distortion classes, since it is applied to both pristine and authentically distorted images, enforcing learning of perceptually relevant image-quality features.

2. Related Work

As discussed in Section 1, perceptual image quality prediction for "Images in the Wild" is a challenging task due to the complex distortions that arise, and the combinations of them, and how they are perceived when they affect different kinds of pictorial content. Over the last few decades, a great deal of effort has been invested in the development of NR-IQA models that are able to accurately predict human judgement of picture quality. In recent years, NR-IQA models have evolved from using hand-crafted perceptual features, feeding shallow learners, into Deep Learning based approaches trained on large subjective databases. Traditional NR-IQA models generally have two components: a feature extractor, which generates quality-relevant features, and a low-complexity regression model, which maps the extracted features to quality scores. Most prior models have focused on improving the feature extractor and thus improving the performance of the overall IQA algorithm. A common practice in traditional NR-IQA methods is to



(a) Original Image (b) Gaussian Blur (c) Saturate (d) Compression (e) Gaussian Noise (f) Color Saturate

(g) Pixelate (h) Contrast (i) Darken (j) Resizing (k) Jitter (l) Motion Blur

Figure 3. Some samples of distortions available in the Image Augmentation Scheme. There are a total of 26 distortions available in the bank with 5 levels of distortion for each. We make sure the chosen levels generate distortions similar to authentic distortions.

model image artifacts using statistical information extracted from a test image. Natural Scene Statistics (NSS) models and distorted versions of them are popular, where features are extracted from transformed domains, on which statistical measurements of deviations due to distortions are used as features for NR-IQA. For example, the NSS-based BRISQUE [16] and NIQE [17] models obtain features that capture in a normalized bandpass space [23]. DIIVINE [18] uses steerable pyramids, and BLIINDS [24] uses DCT coefficients, both to measure statistical traces of distortions. Other methods like CORNIA [35] and HOSA [34] utilize codebooks constructed from local patches, which are applied to obtain quality-aware features. Most of the methods discussed above often obtain acceptable results when evaluated on synthetically distorted images, but their performances significantly degrades when applied on for “Images in the Wild”. This is because the above-discussed methods focus primarily on modeling the distortions present in a test image as statistical deviations from naturalness, while completely ignoring the high-level content present in the image.

The majority of deep learning approaches utilize pretrained CNN backbones as feature extractors. This is done since end-to-end supervised training of NR-IQA models is difficult given the limited sizes of existing perceptual quality databases. These models typically use CNN backbones trained for ImageNet classification to extract features, combining them with low-complexity regression models like Support Vector Regression or Linear Regression to map the features to human-labeled scores. A few models use labeled scores from IQA databases to fine-tune the CNN backbone. The authors of the RAPIQUE model [29] show that features obtained from pretrained ResNet-50 [6] could be effectively predict quality scores on “In the Wild” content. In [41], authors adopt a two-path technique, where one CNN branch generates distortion-class and distortion-level features, while the other CNN branch provides high-level image content information. These are then combined using bilinear pooling. PQR [37] achieved faster convergence and better quality estimates by using the statistical distributions of subjective opinion scores, instead of just scalar mean opinion scores (MOS) during training. BIECON [8] trains a CNN model on patches of distorted images, using proxy quality scores generated by FR-IQA models as labels. The authors of [27] proposed an adaptive hyper-network architecture that takes content comprehension into account during perceptual quality prediction. Very recent works on NR-IQA includes PaQ-2-PiQ [36], CONTRIQUE [13] and MUSIQ [7]. PaQ-2-PiQ benefits from a specially designed dataset wherein the authors not only collected subjective quality scores on whole pictures, but also on large number of image patches. The dataset is also large enough to train deep models in a supervised setting, and PaQ-2-PiQ achieves state-of-the-art performance. However, although the authors use both patch-level and image-level quality information is used during training, the training process may be susceptible to dataset sampling errors, since only a few patches were extracted from each whole image and annotated with quality scores. MUSIQ uses a transformer-based architecture [30] for pre-trained on the ImageNet classification dataset. The method benefits significantly by the use of transformer architecture and the fine-tuning the transformer backbone on the IQA test databases.

Here, we utilize ResNet-based architectures [6], althought the Re-IQA framework. As our proposed framework is generalizable enough to be implemented using other CNN and transformer-based architectures, we plan to extend it to transformer-based architectures in the future. CONTRIQUE is a closely related work that aims to learn quality-aware representations in a self-supervised setting. CONTRIQUE learns how to group images having similar types and degrees of distortion into classes on an independent dataset. In this way is able to learn quality-aware image representations. Our method which is also completely unsupervised, does not learn representations based on distortion class labels, which can be inaccurate when asserted on "In the Wild" data. Instead, our model, unsupervised model which is based on the fundamental principles of visual distortion perception learns high-level semantic image content as well as low-level image quality features. These image representations features are mapped directly to subjective scores using a low-complexity regression model, without any fine-tuning of the deep neural network.

3. Rethinking-IQA

The Re-IQA model framework is embodied by three processing phases. The first and second phases consist of training two ResNet50 encoders using contrastive learning of high-level and low-level image information. We then use the pre-trained encoders with frozen weights as image representation generation backbones, which supply features to a low-complexity regression model that is trained to conduct image quality prediction as shown in Figure 1.

To learn high-level content-aware information we deploy MoCo-v2 [2] ImageNet pre-trained model and adopt the design considerations developed in the original paper. Further discussed in section 3.1.

To learn quality-aware representations we develop a contrastive learning framework that deploys a novel augmentation protocol and an intra-pair image-swapping scheme to facilitate model convergence towards learning robust image quality-aware features. Further discussed in section 3.2.

3.1. Re-IQA : Content Aware

The primary objective in the MoCo-v2 framework [2] is to assign a 'similar' label to two crops from a single image while assigning a 'different' label to two crops taken from two different images. Although, content aware Re-IQA based completely on the original MoCo-v2 framework performs well in the image quality prediction problem (refer Table 1), it still suffers from a critical design problem: two crops from a same image can be given significantly different quality scores by human viewers. Hence we only use, the original MoCo-v2 framework to generate content-aware image representations. We make appropriate changes, discussed next, to the MoCo-v2 framework to enable accurate learning of quality-aware representations that complement content-aware representations.

3.2. Re-IQA : Quality Aware

Our quality-aware contrastive learning framework uses an Image Augmentation method and an Intra-pair image Swapping scheme to training a ResNet-50 encoder within the MoCo-v2 framework [2] with a goal of modelling a feature space wherein all images having similar degrees of perceptual quality fall closer to one another than to images having different perceptual qualities. The MoCo-v2 framework simultaneously processes a query image through the query encoder and a key image through the key encoder. In a single batch, a positive sample occurs when features are generated using any paired query and key, the pair being labeled 'similar.' A negative sample occurs when the query and the key do not belong to the same pair, hence they are marked 'different'.

To train a contrastive network we need paired images, such that for any sample index k, we have image pairs $[{i_1}^k, {i_2}^k]$ that can be assigned the 'similar-quality' label, and for any j, k; where $k \neq j$ we have image pairs $[{i_1}^k, {i_2}^j]$ that can be assigned the 'different-quality' label. From here on we shall refer to perceptual quality-aware features as PQAF. To define the decision boundary between 'similar-quality' and 'different-quality' labels we assume the following three hypotheses to be true:

H1: PQAF varies within an image itself. If we assign PQAF to an image patch x and denote it as $PQAF_x$, then $PQAF_x$ varies only a small amount between neighboring patches. However, $PQAF_x$ may have significantly between two distant patches.
H2: The PQAF of any two randomly selected images are 'different', which assumes that the scenes depicted in the images to be different. However, this does not enforce any restrictions on the quality scores of the two images.
H3: Two different distorted versions of the same image have different PQAF.

3.2.1 Quality-Aware Image Augmentation Scheme

To conduct quality-aware image feature extraction we deploy a novel bank of image quality distortion augmentations, as elaborated in the Supplemental material §S.1. The augmentation bank is a collection of 25 distortion methods each realized at 5 levels of severity. For any source image i^k from the training set, where $k \in \{1, 2...K\}$ and K is the total number of images in the training data, a randomly chosen subset of the augmentations available in the bank are applied to each image resulting in a mini-batch of distorted images. We combine each source image with its distorted versions to form $chunk^k$:

$$chunk^k = [i^k, {i_1}^k, {i_2}^k, ..., {i_n}^k] \tag{1}$$

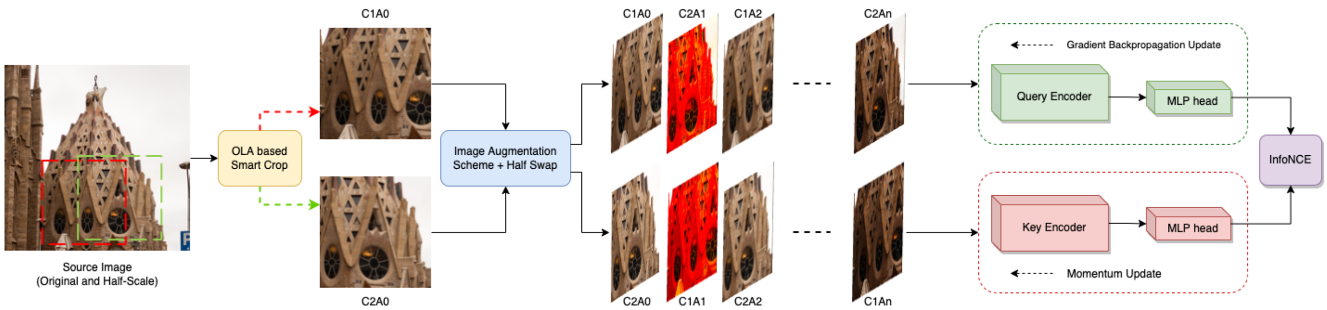


Figure 4. Learning Quality Aware Representations: The OLA based cropping, Image Augmentation scheme and Half-swapping enable the generation of appropriate 'similar' and 'different' image pairs which can be used to learn quality-aware features. Note that A_0 has no augmentation, while $A_1..A_n$ are randomly sampled from the augmentation bank. During loss calculation, representations generated using the key encoder for the previous 65536 samples are also used as negative keys, following MoCo-V2 settings.

where ${i_j}^k$ is the j^{th} distorted version of i^k, and n is the number of augmentations drawn from the bank. We then generate two random crops of $chunk^k$, namely $chunk^{k_{c1}}$ and $chunk^{k_{c2}}$, using an overlap area based smart cropping mechanism. We choose these crop locations such that the overlapping area (OLA) in the two crops falls within a minimum and maximum bounds. We make sure that the crop location is the same over all images in each chunk and different between chunks, resulting in:

$$\begin{aligned} chunk^{k_{c1}} &= [i^{k_{c1}}, {i_1}^{k_{c1}}, {i_2}^{k_{c1}}, ..., {i_n}^{k_{c1}}] \\ chunk^{k_{c2}} &= [i^{k_{c2}}, {i_1}^{k_{c2}}, {i_2}^{k_{c2}}, ..., {i_n}^{k_{c2}}] \end{aligned} \quad (2)$$

When training, by choosing an augmented image a_{th} from both $chunk^{k_{c1}}$ and $chunk^{k_{c2}}$, form the pair $[i_a^{k_{c1}}, i_a^{k_{c2}}]$. Image $i_a^{k_{c1}}$ and $i_a^{k_{c2}}$ are neighboring patches because of OLA-based cropping and hence are marked 'similar-quality' as stated in **H1**. Similarly, for any image k and distortion a, b, where $a \neq b$, the pair $[i_a^{k_{c1}}, i_b^{k_{c2}}]$ are labelled as 'different-quality' as in **H1**. Finally, for any two different image samples k, j, label the pair $[i_-^{k_{c1}}, i_-^{j_{c2}}]$ as 'different-quality', following **H2**.

3.2.2 Intra-Pair Image Swapping Scheme

Given a spatial arrangement of $chunk^{k_{c1}}$ and $chunk^{k_{c2}}$:

$$\begin{array}{cccccccc} i^{k_{c1}} & {i_1}^{k_{c1}} & ... & {i_m}^{k_{c1}} & {i_{m+1}}^{k_{c1}} & ... & {i_{n-1}}^{k_{c1}} & {i_n}^{k_{c1}} \\ \updownarrow & \updownarrow & & \updownarrow & \updownarrow & & \updownarrow & \updownarrow \\ i^{k_{c2}} & {i_1}^{k_{c2}} & ... & {i_m}^{k_{c2}} & {i_{m+1}}^{k_{c2}} & ... & {i_{n-1}}^{k_{c2}} & {i_n}^{k_{c2}} \end{array}$$

form the following types of image-pairs and corresponding labels:

$$\begin{aligned} [i_m^{k_{c1}}, i_m^{k_{c2}}] &\mapsto similar - quality \\ [i_m^{k_{c1}}, i_l^{k_{c2}}] &\mapsto different - quality \end{aligned}$$

Then apply intra-pair image swapping on the generated chunks to obtain the following arrangement:

$$\begin{array}{cccccccc} i^{k_{c1}} & {i_1}^{k_{c2}} & ... & {i_m}^{k_{c1}} & {i_{m+1}}^{k_{c2}} & ... & {i_{n-1}}^{k_{c1}} & {i_n}^{k_{c2}} \\ \updownarrow & \updownarrow & & \updownarrow & \updownarrow & & \updownarrow & \updownarrow \\ i^{k_{c2}} & {i_1}^{k_{c1}} & ... & {i_m}^{k_{c2}} & {i_{m+1}}^{k_{c1}} & ... & {i_{n-1}}^{k_{c2}} & {i_n}^{k_{c1}} \end{array}$$

By swapping images within each pair over half the pairs, (referred to as Half Swap), the network is introduced to samples having the following configuration: $[i_a^{k_{c1}}, i_b^{k_{c1}}]$ where $a, b; a \neq b$ are two different distortions. Note that the crops $[i_a^{k_{c1}}, i_b^{k_{c1}}]$ are exactly the same, except for the distortion applied, and thus contain the same essential visual content. Despite this, we mark such samples as 'different-quality' as stated in **H3**, thus forcing the network to look beyond content-dependent features. With this we finally end up with the following image pairs and labels:

$$\begin{aligned} [i_m^{k_{c1}}, i_m^{k_{c2}}] &\mapsto similar - quality \\ [i_m^{k_{c1}}, i_l^{k_{c2}}] &\mapsto different - quality \\ [i_m^{k_{c1}}, i_l^{k_{c1}}] &\mapsto different - quality \\ [i_m^{k_{c1}}, i_l^{j_{c2}}] &\mapsto different - quality \end{aligned}$$

3.2.3 Quality-Aware Training

Define two identical encoders 1) Online Encoder (query encoder) and 2) Momentum Encoder (key encoder). Both encoders have ResNet-50 backbones and an MLP head to generate the final output embeddings from the ResNet features. Split the pairs designed in the previous step, passing the first image from each pair through the query encoder, and the other through the key encoder. To calculate the loss between the representation generated by query and key encoder, we use the InfoNCE [20] loss function:

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q.k^+/\tau)}{\exp(q.k^+/\tau) + \sum_{k^-} \exp(q.k^-/\tau)} \quad (3)$$

Here q is the query image, k^+ is a positive sample (similar-quality), k^- represnt negative samples (different-quality), and τ is a temperature hyper-parameter. This loss is then used to update the weights of the online encoder by back-propagation. The weights of the momentum encoder are updated using the weighted sum of its previous weights and the new weights of the online encoder. Formally denoting the parameters of the query encoder by θ_q and the parameters of the key encoder as θ_k, update θ_k as:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \quad (4)$$

Here $m \in [0, 1)$, is the momentum coefficient. Once the encoder pre-training has saturated the frozen ResNet-50, the encoder can be used as a backbone for any downstream task associated with perceptual image quality.

3.3. IQA Regression

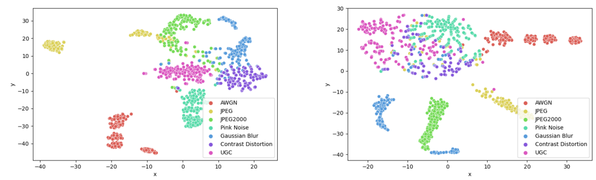
We combine the image representations obtained from the content and quality-aware encoders in the previous steps to train a regressor head to map the obtained features to the final perceptual image quality scores as shown in Figure 1. In our experiments, we use a single-layer perceptron as the regressor head. It is important to note that we train only the low-complexity regressor head while evaluating our Re-IQA framework across multiple databases. Our method does not require us to fine-tune the feature extraction backbone(s) separately for each evaluation database as required in MUSIQ [7].

4. Experimental Results

4.1. Training Datasets

In the Re-IQA framework, two ResNet-50 encoders are trained to obtain high-level image content features and low-level image quality features. The encoder that learns earn high-level image content features was trained on a subset of the ImageNet database [3] containing approximately 1.28 million images across 1000 classes. When training the encoder in an unsupervised setting, we discard the class label information and only use images without labels during the training process.

To learn the low-level image quality features, we use a combination of pristine images and authentically distorted images as training data. The augmentation scheme (applied to all images in the dataset) ensures that the network learns how to differentiate between distortions when the semantic



(a) Re-IQA Quality-Aware (b) CONTRIQUE

Figure 5. Comparison of 2D TSNE Visualization of learned representations of 1016 images sampled from KonIQ (UGC - #150) and CSIQ (Synthetic Distortions - #866) between Re-IQA Quality Aware sub-module and CONTRIQUE. Best viewed zoomed in.

content in the image is the same. The presence of authentically distorted images in the dataset helps tune the model to accurately predict the quality of "In the Wild" pictures.

- Pristine Images: We used the 140K pristine images in the KADIS dataset [10]. We do not use the 700K distorted images available in the same dataset. The authors of KADIS did not provide subjective quality scores for any image in the dataset.
- Authentically Distorted Images: We used the same combination of datasets as proposed in CONTRIQUE [13] to form our distorted image set: (a) AVA [19] - 255K images, (b) COCO [11] - 330K images, (c) CERTH-Blur [15] - 2450 images, d) VOC [4] - 33K images

4.2. Evaluation Datasets

Many previous IQA methods used legacy databases like LIVE IQA [26], TID-2008 [22], TID-2013 [21], CSIQ-IQA [9], and KADID [10] for development and evaluation purposes. However, these datasets contain only a small number ($\sim 25-100$) of pristine images synthetically distorted by various levels and types of single distortions. Hence, these datasets lack diversity and realism of content and distortion. Recently many "In the Wild" datasets like KonIQ, CLIVE, FLIVE, and SPAQ have been developed and used by visual quality researchers since they address the shortcomings of the legacy datasets. The newer breed of perceptual quality datasets contains many authentically distorted images.

The characteristics of each of the above-mentioned "In the Wild" datasets are provided:

- KonIQ-10K: 10K images sampled from the publicly available large-scale multimedia database - YFCC100M [28].
- CLIVE: 1162 authentically distorted images captured by diverse mobile devices.
- FLIVE: 40K images sampled from open-source images to statistically mimic the feature distributions of real social media images.

- SPAQ: 11K images captured using 66 mobile devices, along with a variety of annotations (brightness, content labels, EXIF data, etc) about each image. We only use the image and its quality score for our experiments.

We also evaluated our method on four legacy synthetically distorted datasets: LIVE-IQA, TID-2013, CSIQ-IQA, and KADID. We provide short descriptions of each of the databases below.

- LIVE IQA: 779 distorted images generated by applying 5 distortion types at 4 levels on 29 pristine images.
- TID-2013: 3000 distorted images generated by applying 24 distortion types on 25 pristine images. Each distortion type was applied with 5 different levels of degradation.
- CSIQ-IQA: 866 distorted images generated by applying 6 distortion types on 30 pristine images.
- KADID: 10125 distorted images generated by applying 25 different distortion types on 81 pristine images.

4.3. Training Configurations

Our content-aware encoder is pre-trained on ImageNet database following the configuration proposed in MoCo-v2. Due to time and resource constraints, we train the content-aware encoder for 200 epochs.

For the quality-aware encoder, we used ResNet-50 as feature extractor, and a 2-layer MLP head to regress contrastive features of dimension 128. The hidden dimension of the MLP head has 2048 neurons. In each forward pass, the OverLap Area (OLA) based cropping mechanism chooses two crops (C_1 and C_2) from each image, such that the percentage of overlap between the crops is maintained within a minimum and a maximum bound. The performance variation of Re-IQA against changes in percentage OverLap Area is depicted in Table 2. We also run ablation on the patch size chosen during training and report the same in Table 2. We achieved the highest IQA scores when the percentage OverLap Area bound was chosen as $10 - 30\%$ with a patch size of 160. The Image Augmentation scheme generates augmented versions of these crops which are then Half-swapped. The number of augmentations chosen for our best model was 11, and its impact on performance can be seen in Table 2.

The processed chunks are passed through the query and key encoders respectively in the MoCo-v2 framework, followed by an adaptive pooling layer to compress the output of the ResNet-50 into a 1D feature vector. The generated feature vector is then fed to the MLP head to generate the contrastive feature vectors required for loss computation. Our design of the Re-IQA model is inspired by previous works [13, 33] that use images both at their original and half-scale. Therefore, during the training phase of the Re-IQA model we use all images in a database both at original and half-scale, thereby doubling the training dataset.

During training, the following hyper-parameters were fixed throughout all experiments: learning rate $= 0.6$ with cosine annealing [12] scheduler, InfoNCE temperature $\tau = 0.2$, and momentum coefficient $= 0.999$. Our best-performing model required a batch size of 630 (effectively 630 x $(n + 1)$ augmentations x (2) scales) during training and was trained for 25 epochs. Convergence occurs in a relatively shorter number of epochs as the effective dataset size increases drastically due to a large number of augmentations and processing of each image in the dataset at two scales.

All the implementations were done in Python using the PyTorch deep learning framework. We trained the content and quality-aware encoders on a system configured with 18 Nvidia A100-40GB GPUs.

4.4. Evaluation Protocol

We tested our Re-IQA model against other state-of-the-art models on all of the "In the Wild" and synthetically distorted IQA databases described in Section . Each of these datasets is a collection of images labeled by subjective opinions of picture quality in the form of mean of the opinion scores (MOS).

The single-layer regressor head in Re-IQA is trained by feeding the output of the pre-trained encoders and then comparing the output of the regressor, against the ground truth MOS using L2 loss. We use both Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC) as metrics to evaluate the trained model across the different IQA databases.

Following the evaluation protocol used in [13], each dataset was randomly divided into 70%, 10% and 20% corresponding to training, validation and test sets, respectively. We used the validation set to determine the regularization coefficient of the regressor head using a 1D grid search over values in the range $[10^{-3}, 10^{3}]$. To avoid overlap of contents in datasets with synthetic distortions, splits were selected based on source images. We also prevented any bias towards the training set selection by repeating the train/test split operation 10 times and reporting the median performance. On FLIVE, due to the large size of the dataset, we follow the train-test split recommended by the authors in [36].

4.5. Results

Our "Mixture of Experts" approach in Re-IQA enables us to learn robust high-level image content and low-level quality aware representations independently, the benefit of which can be clearly observed in the performance values reported in Table 1. We compared the performance of Re-

Method	Authentic Distortions a.k.a "Images in the Wild"								Synthetic Distortions							
	KonIQ		CLIVE		FLIVE		SPAQ		LIVE-IQA		CSIQ-IQA		TID-2013		KADID	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.665	0.681	0.608	0.629	0.288	0.373	0.809	0.817	0.939	0.935	0.746	0.829	0.604	0.694	0.528	0.567
CORNIA	0.780	0.795	0.629	0.671	-	-	0.709	0.725	0.947	0.950	0.678	0.776	0.678	0.768	0.516	0.558
DB-CNN	0.875	0.884	0.851	**0.869**	0.554	0.652	0.911	0.915	0.968	**0.971**	**0.946**	0.959	0.816	**0.865**	0.851	0.856
PQR	0.880	0.884	**0.857**	**0.882**	-	-	-	-	0.965	**0.971**	0.872	0.901	0.740	0.798	-	-
PaQ-2-PiQ	0.870	0.880	0.840	0.850	0.571	0.623	-	-	-	-	-	-	-	-	-	-
HyperIQA	0.906	0.917	**0.859**	**0.882**	0.535	0.623	0.916	0.919	0.962	0.966	0.923	0.942	0.840	0.858	0.852	0.845
CONTRIQUE	0.894	0.906	0.845	0.857	0.580	0.641	0.914	0.919	0.960	0.961	0.942	0.955	**0.843**	0.857	**0.934**	**0.937**
MUSIQ	**0.916**	**0.928**	-	-	**0.646**	**0.739**	**0.917**	**0.921**	-	-	-	-	-	-	-	-
ImageNet Pretrained (Supervised)	0.888	0.904	0.781	0.809	0.595	0.648	0.904	0.909	0.925	0.931	0.840	0.848	0.679	0.729	0.701	0.677
Re-IQA (content aware)	0.896	0.912	0.808	0.844	0.588	0.699	0.902	0.908	0.867	0.858	0.766	0.824	0.658	0.736	0.601	0.656
Re-IQA (quality aware)	0.861	0.885	0.806	0.824	0.584	0.590	0.900	0.910	**0.971**	**0.972**	0.944	**0.964**	**0.844**	**0.880**	**0.885**	**0.892**
Re-IQA (content + quality)	**0.914**	**0.923**	0.840	0.854	**0.645**	**0.733**	**0.918**	**0.925**	**0.970**	**0.971**	**0.947**	**0.960**	0.804	0.861	0.872	0.885

Table 1. Performance comparison of Re-IQA against various NR-IQA models on IQA databases with authentic distortions. The top 2 best performing models are in bold. Higher SRCC and PLCC scores imply better performance.

Evaluation Dataset	n_{aug}					Patch Size					OLA bound (%)			
	2	5	11	15	23	128	160	192	224	256	5-15	10-30	50-80	no bound
KonIQ	0.895	0.901	0.914	0.898	0.895	0.91	0.914	0.911	0.905	0.897	0.903	0.914	0.897	0.905
SPAQ	0.902	0.913	0.918	0.91	0.909	0.916	0.918	0.914	0.908	0.903	0.907	0.918	0.904	0.906
CSIQ	0.932	0.941	0.947	0.937	0.935	0.94	0.947	0.942	0.937	0.932	0.939	0.947	0.935	0.94

Table 2. SRCC performance comparison of Re-IQA while varying one hyper-parameter at a time. While varying n_{aug}, we keep patch size 160 and OLA bound $10-30\%$. When varying patch size, n_{aug} was fixed to 11 and OLA bound to $10-30\%$. When varying OLA bound, n_{aug} was set to 11 and the patch size was set to 160.

IQA along with its sub-modules against other state-of-the-art models on IQA datasets containing authentic and synthetic distortions in Table 1. From the results, we conclude that Re-IQA achieves competitive performance across all tested databases.

Results from Table 1 highlight the impact of content and low-level image quality on the final NR-IQA task. We observe that high-level content-aware features dominate quality-aware features for authentically distorted images, while the quality-aware features dominate the high-level content-aware features for authentically distorted images. We can hypothesize the reason to be high variation in content in the "Images in the Wild" scenario.

Training a simple linear regressor head that is fed with features from both the content and quality-aware encoders, provides flexibility to adjust the final model based on the application dataset. This can be clearly observed in the performance scores achieved by the combined model when compared to the individual sub-modules.

The performance scores of the quality-aware sub-module do not beat other methods when considering the "Images in the Wild" scenario which is primarily due to the heavy impact of content. Despite this, the quality-aware sub-module single-handedly beats most of its competitors when evaluated on synthetic distortion datasets. Thus we conclude that our generated quality-aware representations align very well with distortions present in an image. This is also conclusive from the T-SNE visualizations depicted in Figure 5.

5. Concluding Remarks

We developed a holistic approach to Image quality Assessment by individually targeting the impact of content and distortion on the overall image quality score. NR-IQA for "Images in the Wild" benefits significantly from content-aware image representations, especially when learned in an unsupervised setting. This work aims to demonstrate that complementary content and image quality-aware features can be learned and, when combined, achieve competitive performance across all evaluated IQA databases. For learning quality-aware representations we re-engineer the MoCo-v2 framework to include our proposed novel Image Augmentation, OLA-based smart cropping, and a Half-Swap scheme. The results of experiments on the eight IQA datasets demonstrate that Re-IQA can consistently achieve state-of-the-art performance. Our Re-IQA framework is flexible to changes in design of encoder architectures and can be extended to other CNN architectures and Transformer based models like MUSIQ [7]. Although developed for IQA tasks, Re-IQA can be extended as a spatial feature extraction module in Video Quality Assessment algorithms that currently use supervised pre-trained Resnet-50 features. A public release of all the codes associated with this work will be made available on GitHub.

References

[1] Alan Conrad Bovik. Automatic prediction of perceptual image and video quality. *Proceedings of the IEEE*, 101:2008–2024, 2013. 2

[2] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 2, 4

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 6

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[7] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. *CoRR*, abs/2108.05997, 2021. 3, 6, 8

[8] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1):206–220, 2016. 3

[9] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010. 6

[10] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 6

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[12] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. 7

[13] Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. Image quality assessment using contrastive learning. *CoRR*, abs/2110.13266, 2021. 1, 2, 3, 6, 7

[14] J. Mannos and D. Sakrison. The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on Information Theory*, 20(4):525–536, 1974. 2

[15] Eftichia Mavridaki and Vasileios Mezaris. No-reference blur assessment in natural images using fourier transform and spatial pyramids. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 566–570. IEEE, 2014. 6

[16] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 1, 2, 3

[17] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 3

[18] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing*, 20(12):3350–3364, 2011. 3

[19] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012. 6

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[21] Nikolay Ponomarenko, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Lina Jin, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Color image database tid2013: Peculiarities and preliminary results. In *European workshop on visual information processing (EUVIP)*, pages 106–111. IEEE, 2013. 6

[22] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009. 6

[23] Daniel L Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994. 3

[24] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing*, 21(8):3339–3352, 2012. 3

[25] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. 2

[26] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 6

[27] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3664–3673, 2020. 3

[28] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 6

[29] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Rapique: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021. 3

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 3

[31] Zhou 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*, 13(4):600–612, 2004. 1, 2

[32] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. 2

[33] Zhou Wang, Eero P. Simoncelli, and Alan Conrad Bovik. Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2:1398–1402 Vol.2, 2003. 7

[34] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016. 3

[35] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1098–1105. IEEE, 2012. 2, 3

[36] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. *CoRR*, abs/1912.10088, 2019. 1, 3, 7

[37] Hui Zeng, Lei Zhang, and Alan C Bovik. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*, 2017. 3

[38] Lin Zhang, Ying Shen, and Hongyu Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014. 2

[39] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 1, 2

[40] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1, 2

[41] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 3