Image Quality Assessment Using Contrastive Learning

Pavan C. Madhusudana, Neil Birkbeck, Yilin Wang, *Member, IEEE*, Balu Adsumilli, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*— We consider the problem of obtaining image quality representations in a self-supervised manner. We use prediction of distortion type and degree as an auxiliary task to learn features from an unlabeled image dataset containing a mixture of synthetic and realistic distortions. We then train a deep Convolutional Neural Network (CNN) using a contrastive pairwise objective to solve the auxiliary problem. We refer to the proposed training framework and resulting deep IQA model as the CONTRastive Image QUality Evaluator (CONTRIQUE). During evaluation, the CNN weights are frozen and a linear regressor maps the learned representations to quality scores in a No-Reference (NR) setting. We show through extensive experiments that CONTRIQUE achieves competitive performance when compared to state-of-the-art NR image quality models, even without any additional fine-tuning of the CNN backbone. The learned representations are highly robust and generalize well across images afflicted by either synthetic or authentic distortions. Our results suggest that powerful quality representations with perceptual relevance can be obtained without requiring large labeled subjective image quality datasets. The implementations used in this paper are available at https://github.com/pavancm/CONTRIQUE.**

***Index Terms*— No reference image quality assessment, blind image quality assessment, self-supervised learning, deep learning.**

I. Introduction

IMAGE Quality Assessment (IQA) pertains to the problem of quantifying and predicting human perceptual judgments of image quality. No-Reference (NR) or blind IQA is focused on estimating the quality of degraded images with no information about any pristine reference images or of the types of distortions that are present. The goal of NR-IQA models is to make robust and accurate quality predictions that correlate well with subjective judgments. The typical presence of multiple types of artifacts, as well as the influence of image content on perceived quality makes NR-IQA an interesting and challenging problem. NR-IQA has become a central technology for social media platforms such as Facebook, Instagram, Flickr etc. where millions of digital user-generated content (UGC) images are uploaded everyday. It is necessary to be able to objectively determine and control the quality of these digital photographs, and to guide subsequent processing tasks, such as compression [1]. Additionally, IQA models can also be employed as objectives when training image enhancement models for image denoising, super-resolution etc. [2]–[4].

NR-IQA has been a topic of intense interest among the research community for more than a decade, resulting in a variety of IQA datasets and objective models. Legacy IQA databases such as LIVE-IQA [5], CSIQ-IQA [6] etc. have been influential in advancing the field of image quality prediction. These early datasets contain images with synthetic distortions, whereby a pristine high quality reference is artificially corrupted by commonly observed distortions such as blur, white noise, compression artifacts etc. However, a shortcoming of these datasets is that in most instances, a 'single' distortion type is applied on each image, whereas in reality images commonly are degraded by a combination of multiple distortions. To address this, various recent databases have been introduced that contain real, authentically distorted images [7]–[10], typically captured by casual users with handheld camera devices. From the perspective of objective NR model design, it is desirable to obtain a model that can perform well on both synthetic and authentic distortions, so that it is applicable to any image regardless of the type of impairments it is afflicted with.

Well established NR-IQA models typically rely on parametric or learned approaches. Natural scene statistics (NSS) based models [11]–[14] use features which are derived from statistical observations, and use them to predict visual quality. These kinds of algorithms have been very successful at analyzing synthetic artifacts, but their performance has proven to be limited when evaluated on images afflicted by unknown, often commingled authentic distortions. Over the last decade, the many successes of deep Convolutional Neural Networks (CNN) [15]–[17] trained on large databases has motivated the development of many CNN based, data-driven IQA models have been proposed [18]–[21].

One barrier to the development of CNN based IQA models is the lack of availability of sufficiently large labeled IQA datasets. Annotating IQA datasets is an expensive and labor intensive process. Most available IQA datasets are too small to effectively train deep CNN models from scratch. Because of this, most CNN based IQA models utilize transfer

Manuscript received October 27, 2021; revised April 6, 2022 and May 13, 2022; accepted May 26, 2022. Date of publication June 14, 2022; date of current version June 20, 2022. This work was supported by the National Science Foundation Artificial Intelligence (AI) Institute for Foundations of Machine Learning (IFML) under Grant 2019844. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel L. Lau. *(Corresponding author: Pavan C. Madhusudana.)*

Pavan C. Madhusudana and Alan C. Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: pavancm@utexas.edu; bovik@ece.utexas.edu).

Neil Birkbeck, Yilin Wang, and Balu Adsumilli are with Google Inc., Mountain View, CA 94043 USA (e-mail: birkbeck@google.com; yilin@google.com; badsumilli@google.com).

Digital Object Identifier 10.1109/TIP.2022.3181496

learning, where the CNN is pretrained on a large dataset like ImageNet [22], then fine-tuned end-to-end on images with subjective quality judgments. Although fine-tuned models achieve impressive performances on both synthetic and authentic distortions, fine-tuning requires carefully chosen hyper-parameters that can vary with different IQA databases. Moreover, excessive fine-tuning can overfit the model on the training data limiting its generalizability.

Here we introduce a contrastive learning based IQA training framework aimed towards obtaining efficient image quality representations using *unlabeled* datasets. Our ideas are motivated by the successes of unsupervised/self-supervised pretraining methods [23]–[26] originally proposed for image classification problems. We refer to the new model as **CONTR**astive **I**mage **QU**ality **E**valuator (CONTRIQUE). The salient characteristics of CONTRIQUE are as follows:

1) We use prediction of distortion type and degree as an auxiliary task to train a deep CNN from scratch. Training is done on an unlabeled dataset containing both synthetic and authentic distortions, using a contrastive objective function.
2) To learn robust representations, multiscale and quality preserving transformations are performed on the unlabeled data during training.
3) During testing, the weights of the deep CNN are frozen, and features from this network are mapped to quality scores using a simple linear regressor. Quality predictions produced by CONTRIQUE are shown to be competitive with those of state-of-the-art (SOTA) IQA models across multiple databases. This is accomplished with no additional fine-tuning of the CNN backbone.
4) The CONTRIQUE training framework is simple, and results in in highly generalizable representations that perform well on both synthetic and realistic distortions. Additionally, we show that the CONTRIQUE features can be easily extended to the Full-Reference (FR) IQA problem with no additional training of the CNN backbone.

The rest of the paper is organized as follows: In Section II we discuss prior methods related to IQA and self-supervised learning. In Section III we provide a detailed description of the design of CONTRIQUE. Section IV analyzes and compares various experimental results of CONTRIQUE, and we conclude in Section VI.

II. Related Work

In this section we review related work from the literature concerning NR-IQA and self-supervised learning.

A. NR-IQA Models

Blind image quality prediction is a challenging problem due to the diverse types of artifacts involved. The influence of image content on different distortion types adds additional complexity to the problem. Over the past decade, considerable research effort has been expended on designing NR-IQA models, with the goal of obtaining quality predictions that have high correlations against human judgements. NR models can be broadly categorized based on the design methodology - traditional/hand-crafted models, and deep CNN based models. Most prior models pursue a design philosophy of having a feature extraction framework followed by a regressor to map features to quality values. In traditional models, feature extraction is accomplished by modeling the image artifacts, Natural Scene Statistics (NSS) based models are a popular example employing this approach. NSS models extract features from a transform domain, where deviations from expected statistical regularities due to distortions are predictive of quality. NSS models include DIIVINE [11], which employs steerable pyramids, BLIINDS [12], which uses DCT coefficients, and BRISQUE [13] and NIQE [14], which use mean subtracted contrast normalized coefficients (MSCN) to obtain quality aware features. In CORNIA [27] and HOSA [28], a visual codebook constructed from local patches is used to obtain quality representative features. Although traditional models achieve impressive performances when evaluated on images with synthetic distortions, their capabilities are often limited when tested on images containing realistic distortions and combinations of them.

The successes of deep learning on many computer vision tasks [15]–[17] has inspired a large number of CNN-based NR-IQA models. The motivation behind using CNN is to obtain reliable semantic features from deep architectures, then perform appropriate modifications to adapt them for quality prediction. Due to a lack of large scale data pertaining to image quality, the majority of CNN-based models use transfer learning techniques, whereby a pretrained model (usually pretrained on ImageNet [22]) is fine-tuned using ground-truth image quality labels. In [29], it was shown that features obtained from pretrained CNN architectures like Resnet [15] can be particularly effective in capturing authentic distortions. Ma *et al*. [30] used a multi-task model containing two subnetworks, a distortion identification network, and a quality prediction network, where both networks shared features from early layers. In [18], two separate CNNs are employed to account for synthetic and authentic artifacts, respectively. Kim and Lee [19] employed FR-IQA maps as intermediate regression targets during training. Zeng *et al*. [20] used a statistical distribution of subjective scores when training which led to faster convergence and resulted in superior quality estimates. In [31], [32] the earth mover's distance (EMD) loss was used to train a CNN to predict the distribution of human opinion scores. Su *et al*. [21] proposed an adaptive hyper network architecture to separate quality prediction from content understanding. Talebi *et al*. [33] employ a rank-smoothed loss where pairwise probabilities are regularized with aggregated rankwise probabilities, and show that this modified loss yielded superior correlations against human judgments. In [34] meta-learning was employed to extract prior knowledge that is shared among different types of distortions. Ying *et al*. [9] demonstrated that training with both image and patch quality scores can significantly boost model performance. The PaQ-2-PiQ algorithm developed by these authors also benefited by the availability of an unusually large

subjective database of realistically distorted images. All these models rely on specific supervised fine-tuning mechanisms in order to achieve improved performance. In contrast, our work focuses on *unsupervised* feature learning with no fine-tuning procedures.

The transformer [35] which was initially introduced for natural language processing (NLP) tasks, has gathered significant interest in the computer vision community for various vision tasks [36]–[38]. For example, the Vision Transformer (ViT) [38] employs a pure transformer based architecture directed towards the image classification task by treating images as a sequence of patches. Ke *et al*. [39] proposed a multi-scale image quality (MUSIQ) transformer for processing images with varying resolutions and aspect ratios, demonstrating superior performances on multiple IQA datasets. Transformer architectures are typically complex and often require a significant amount of data and computational resources for training. Here, our focus is instead directed towards self-supervised feature learning, and having model complexity similar to CNN based IQA models. Thus, we only use CNN based architectures.

B. Self-Supervised Learning

Self-supervised learning or unsupervised pretraining aims at obtaining representations using unlabeled data. These techniques derive useful representations by exploiting existing structural information available in the image data. Recent SOTA methods rely on instance discrimination task, in which each image and augmented versions of it are treated as a single class [23], [25], [26]. Another form of self-supervision involves learning features through auxiliary tasks (different but related to the original task) for which data is abundant, and which requires no annotations. Examples of these self-supervised tasks include rotation prediction [40], obtaining color images from grayscale and vice versa [41], [42], and inpainting [43]. Liu *et al*. [44] proposed an NR-IQA model using image ranking as an auxiliary task, and achieved competitive performance on datasets with synthetic artifacts. Here we use discrimination of distortion types and degrees, which is related to quality assessment, as a self-supervision task and then we use the learned representations for image quality prediction.

III. Method

Our method is a transform domain approach where a transformation $f : \mathbb{R}^{3\times H\times W} \mapsto \mathbb{R}^d$ maps an image x to a representation h. Bandpass transformations such as wavelet-like decompositions are often used to model the responses of visual neurons in primary visual cortex that are tuned to visual stimuli having specific spatial locations, frequencies, and orientations. Traditional NR-IQA models have been based on band-pass transformations such as the DCT [12], steerable pyramids [11], local mean-subtraction [13], [14] and so on, have been highly effective at predicting perceptual quality. Recently, transformations induced by deep CNNs have demonstrated remarkable efficiency at capturing perceptual image artifacts [18], [20], [21].

Here, our goal is to learn robust representations that can be used to predict image quality, without employing any ground-truth quality scores during training. Our proposed training pipeline is illustrated in Fig. 1. In the following sections each module present in the framework is discussed in detail.

A. Auxiliary Task

An auxiliary task to learning problem is an alternate but closely related task, for which the ground-truth labels are known or can easily be obtained. In this approach, model is trained to solve an auxiliary problem, then during the inference stage, the trained model is evaluated on the original task. In the case of IQA, the goal is to obtain discriminative representations that can distinguish different types of distortions, as well as the degrees of degradations. Thus, we transform the IQA representation learning problem to a classification problem, where each class consists of images having a similar type of distortion, as well as similar degree of quality degradation. The goal of the auxiliary task is to learn features that can differentiate images into distortion dependent classes, similar to [18], [45], which employ a cross-entropy objective during training to achieve this.

Let a pristine high quality image x be degraded by a distortion $d^i, i \in \{1, \ldots, D\}$ with degradation degree $l^{ij}, j \in \{1, \ldots, L^i\}$ resulting in a distorted image $\tilde{x}_i^j$. Here, D and L^i correspond to the number of distortion types and degradation degrees, respectively. For a given $\tilde{x}_i^j$, the task of the model is to identify d^i and l^{ij}. This task translates to a classification problem having $\sum_{i=1}^{D} L^i + 1$ classes (total number of degradation levels + one pristine image). Motivated by the successes of using contrastive loss [25], [26] for learning representations, we incorporate a similar technique into the CONTRIQUE framework. To extract embeddings, we define a deep model consisting of two parts : an encoder and projector. The encoder can be any popular CNN architecture such as VGG [46], Resnet [15] etc., with any fully connected terminal layer removed. The projector is a multi-layer perceptron (MLP) that reduces the dimensionality of the representation produced by the encoder. Let $f(.)$ and $g(.)$ denote the deep encoder network and the projector network respectively. For a given image $x \in \mathbb{R}^{3\times H\times W}$

$$h = f(x), \quad z = g(h) = g(f(x)) \; h \in \mathbb{R}^B, \; z \in \mathbb{R}^K \tag{1}$$

where h is the B-dimensional output from the encoder. Similar to [25], [26] the encoder output h is L_2 normalized before being fed to the projector network. Note that the output of the entire model z is a K-dimensional vector (where K is a hyperparameter in this design). The goal is to obtain similar representations z of images belonging to the same class. The similarity between a pair of representations is measured using the dot product $\phi(a, b) = a^T b/||a||_2||b||_2$. The loss function is a normalized temperature-scaled cross entropy (NT-Xent), and for image x_i is defined as

$$\mathcal{L}_i^{syn} = \frac{1}{|P(i)|} \sum_{j\in P(i)} -\log \frac{\exp(\phi(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{k\neq i} \exp(\phi(z_i, z_k)/\tau)}, \tag{2}$$

where N is the number of images present in the batch, $\mathbb{1}$ is the indicator function, τ is the temperature parameter, $P(i)$ is a set containing image indices belonging to the same class as x_i (but

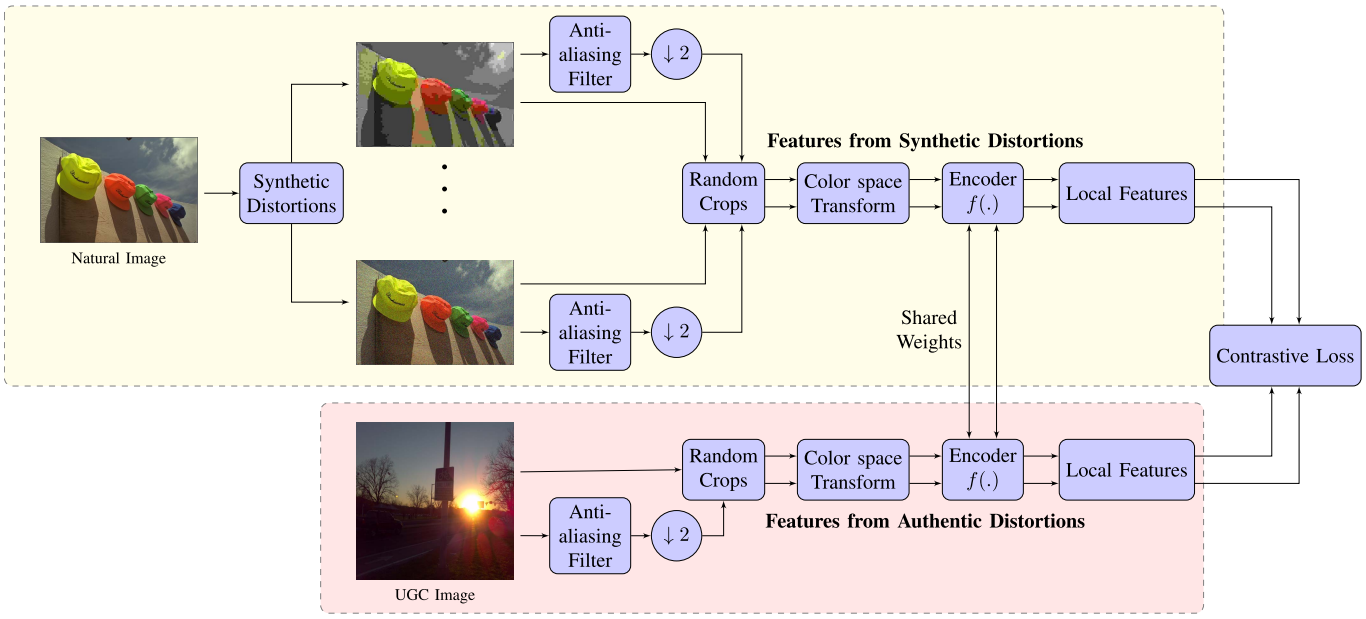


Fig. 1. Illustration of training pipeline of the CONTRIQUE Framework.

excluding the index i) and $|P(i)|$ is its cardinality. For example if x_i is an image corrupted with additive white Gaussian noise (AWGN) with $\sigma = 10$, $P(i)$ will contain indices of all images (excluding index i) present in the batch corrupted with AWGN $\sigma = 10$. There exists transformed versions of image x_i (image transformations are discussed in Sec. III-B and Sec. III-C) ensuring that $P(i)$ contains atleast one sample. The objective function (2) is similar to the supervised contrastive loss proposed in [47]. However, in [47] it was employed in the context of image classification with ground-truth labels, while in our design we incorporate prior knowledge of synthetic distortions as class labels. Another observation that can be made about the objective described in (2) is that it measures pairwise similarities between every pair of images in a batch. This pairwise loss computation is a key characteristic that differentiates it from the traditional cross-entropy loss.

B. Multiscale Learning and Cropping

Images are inherently multi-scale, as are distortions of them, and perceived image quality is influenced by both local characteristics as well as global details. Prior IQA models [11], [13], [14], [48] have attempted to simulate the functionality of front-end visual processing in the brain by employing multi-scale representations when predicting quality. CNN based IQA models [21], [49], which use multi-scale features, are able to achieve remarkable efficiency in capturing visual quality. In CONTRIQUE, we employ two scales : native/full resolution, and half-scale resolution obtained by downsampling by a factor of two along both dimensions. To avoid aliasing artifacts, an anti-aliasing filter is used before downsampling as shown in Fig. 1. Note that the aspect ratio is preserved in this resizing operation, since modifying this ratio can affect the quality of the underlying image.

The images are then subjected to random cropping where the input images are cropped to a random fixed size $M \times M$. A simplifying assumption we make here is that the cropped version inherits the same distortion class as the original version. Although the cropped version need not represent the same perceived quality as the original image, we presume that the distortion class remains nearly the same and is unaffected by the cropping operation. For each input image, two random crops are obtained, one each at full-scale and half-scale. For cases where the size of the image was smaller than $M \times M$, the entire image was employed with zero padding to maintain the same resolution. Additionally, cropping provides images of fixed resolution in a batch, which is essential when training deep networks, since training with variable resolutions can be challenging and unstable [9].

C. Quality Preserving Transformations/Augmentations

The goal of the objective function in (2) is to learn image embeddings that demonstrate discriminative behavior among images belonging to different classes, and at the same time exhibit invariance to quality preserving transformations. Image operations that do not modify image quality we collectively refer to as quality preserving transforms. In the CONTRIQUE framework, we employ two transforms: horizontal flipping and color space conversion.

The motivation behind using different color spaces is to extract complementary quality information that can be present across different domains. In our proposed framework, we employ 4 color spaces: RGB, LAB, HSV and grayscale. Each of these color spaces have different types of perceptual relevance and have earlier been used in NSS based models [50]–[52] to obtain quality features. We also employ a band-pass transform, obtained using local Mean-Subtraction (MS). MS coefficients have been shown to capture statistical deviations arising due to distortions in images [53]–[55]. In the

training pipeline shown in Fig. 1, the color space is randomly chosen for each crop of the input image. By employing different color spaces during training, as we show in Sec. IV-G that using any color space during testing results in similar representations, making CONTRIQUE invariant to color spaces. Note that we avoid employing aggressive augmentation techniques such as color jitter, Gaussian blur, random-resize, MixUp [56], AutoAugment [57] etc. as these methods modify distortion information and hence are not quality preserving.

D. Realistic Distortions

Prior knowledge about synthetic distortions was employed in the contrastive objective (2) to learn image quality embeddings. However, for images containing realistic distortions, such as User Generated Content (UGC) images, information regarding the distortion types is usually not available. Being able to handle authentic distortions is quite important since several hundred billion images are uploaded and shared to social media sites like Facebook, Instagram, YouTube etc. every year. UGC images, which are often afflicted by diverse mixtures of unknown distortions. Thus, the synthetic distortion classes assumed in (2) are not applicable to UGC images. In the CONTRIQUE framework, each UGC image is treated as a unique class obtained by a distinctive combination of multiple distortions, separate and distinct from other UGC images, as well as from images with synthetic artifacts. Thus, for a given UGC image x_i, only its scaled (and transformed) version x_j belongs to the same class. To reflect this modification, we redefine the contrastive objective as

$$\mathcal{L}_i^{UGC} = -\log \frac{\exp(\phi(z_i, z_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{k \neq i} \exp(\phi(z_i, z_k)/\tau)}. \quad (3)$$

This objective is similar to the one used in [25], [26] for the instance discrimination task. As detailed in Sec. III-C, for each image there exists two transformed versions, at full-scale and half-scale. Thus, there are at least two datasamples belonging to the same class making the objective (3) non-zero. The expression described in (3) can also be considered as the special case of (2) where $P(i) = \{j\}$, *i.e.* in a given batch only image x_j belongs to the same class as x_i. The overall training objective is then

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(x_i \notin UGC)} \mathcal{L}_i^{syn} + \mathbb{1}_{(x_i \in UGC)} \mathcal{L}_i^{UGC}, \quad (4)$$

where N is the number of images present in the batch, and $\mathbb{1}$ is the indicator function determining whether the input image is non-synthetically distorted (UGC). During training, to avoid bias, we randomly sampled equal numbers of synthetic and authentically distorted images to form each batch, at each iteration.

E. Patch Features

Local details present in image patches play a significant role in determining global picture quality. Several patch-wise learning based models have been proposed in the literature [29], [58] and shown to be effective for quality prediction. In order to capture distortion and image quality characteristics in a more granular fashion, we partitioned each input image into non-overlapping patches of size $P \times P$. These patches were then fed to the encoder module to obtain local features, and these representations are used in the contrastive objective function (4). Similar to cropping operation, we assume the patches inherit the distortion class labels from the original image for both the synthetic as well as the realistically distorted images. Note that patches need not inherit the perceived quality of the original version, only the distortion class is presumed to be same. In addition to capturing local spatial neighborhood information, including patches provides increased number of data samples for every class, which can be beneficial for gradient descent based learning schemes.

F. Evaluating Representations

We evaluate the learned representations by applying them to the quality prediction problem, using the correlations of human judgements against predicted quality scores as a proxy for representation quality. Once the training is complete, the projector network $g(.)$ is discarded and the outputs of encoder network $h = f(x)$ are used as image representations. We use a regularized linear regressor (ridge regression) trained on top of the frozen encoder network. This is similar to the linear evaluation protocol used in [41], [60], [61] to evaluate the classification accuracy of self-supervised models. The regression weights are learned on a suitable IQA database containing ground-truth quality scores. The expression for ridge regression is given by

$$y = Wh, \quad W^* = \underset{W}{\operatorname{argmin}} \sum_{i=1}^{N} (GT_i - y_i)^2 + \lambda \sum_{j=1}^{M} W_j^2, \quad (5)$$

where GT denotes ground-truth quality scores, y predicted scores, W is a trainable vector having same dimensions as h, λ is the regularization parameter, M is number of dimensions of h, and N is the number of images present in the training set. Similar to training, we follow multiscale convention, and features are computed at two resolutions : full-scale and half-scale, and the final representation is a concatenation of both scales. During evaluation, all the representations are calculated at the native resolution of the input image, and no additional data augmentations are performed. Note that we do not perform any additional *fine-tuning* whereby encoder weights would have been modified using the supervision of ground-truth quality scores. Although fine-tuning can potentially yield better performance, we avoid it as it alters the learned encoder weights, and it would not be a true indicator of the efficiency of the unsupervised training process. Additionally, we show in Sec. IV-B that even without fine-tuning, CONTRIQUE achieves competitive performance as compared with SOTA IQA models.

IV. Experiments and Results

In this section we evaluate the performance of CONTRIQUE by conducting a series of experiments.

We will first describe the experimental settings, evaluation protocol and compared methods. Then we explain how we evaluated CONTRIQUE against SOTA IQA models on multiple IQA databases. We perform a variety of ablation experiments to analyze the significance of distortion types present in the pretraining data, importance of using different color spaces during pretraining, multiscale learning as well as the impact of training batch size and crop size. Additionally, we study the generalizability of the CONTRIQUE features by performing cross-dataset testing. At the end, we also highlight some of the limitations of the CONTRIQUE representations.

A. Experimental Settings

1) Pretraining Data: The pretraining data contains a combination of images impaired by synthetic and authentic distortions.

- Synthetic Distortions : We utilized the KADIS dataset [62] to learn synthetic artifacts. The KADIS dataset contains 700k distorted images obtained from 140k pristine images and contains no subjective quality scores. There are 25 different types of distortions with each distortion spanning 5 degrees of degradation. The distortion types include compression, white noise, blur etc. Interested readers can refer to [62] for more details about the distortions present in this dataset. As there are $D = 25$ distortions and $L^i = 5$ degrees for each distortion type, a total of $25 \times 5 + 1$ (pristine image) $= 126$ synthetic classes are used in the contrastive objective (4).
- Authentic Distortions : We use a combination of 4 datasets aimed at capturing realistic distortions. a) The AVA dataset [63] contains 255k images originally designed for aesthetic visual analysis. (b) The COCO dataset [64] contains 330k images designed to assist learning the detection and segmentation of objects occurring in common contexts. (c) The CERTH-Blur dataset [65] contains 2450 images captured with realistic blur. (d) The VOC [66] contains 33k images initially proposed for object recognition task. We discarded all the labels (if any) present in these datasets before training.

Thus, a total of 1.3 million images were used to train CONTRIQUE.

2) Pretraining Details: We used a Resnet-50 [15] architecture as the encoder network $f(.)$ and included 2 layers of MLP as the projector network $g(.)$. The hidden layers of MLP contained 2048 neurons each. The dimension of the final output z was chosen to be $K = 128$. The CONTRIQUE framework is fairly generic in nature, and can easily be extended to other CNN based architectures. The pretraining was done using a batch size of $N = 1024$, with 512 images randomly chosen from the synthetic distortion set and the rest authentically distorted. The sampled images were cropped to square blocks of size $M = 256$. These crops constitute approximately 50% of the original dimensions of the images present in the training data. When extracting patch features, patches of size $P = 64$ were used, resulting in 4 patches from each input image. Patch features were computed by using an adaptive average pooling layer at the end of the encoder. During training the images were loaded in RGB format, and on each image a colorspace transform was applied resulting in a 3-channel image matrix. For each image the choice of colorspace was chosen in a stochastic manner. The resulting 3-channel matrix was normalized to lie in the range $[0, 1]$ before being given as input to the encoder. The temperature parameter used in (2) and (3) was fixed at $\tau = 0.1$. The model was trained from scratch for 25 epochs using a stochastic gradient descent (SGD) optimizer with initial learning rate of 1.2 for a batch size of $N = 1024$. Furthermore, the learning rate was subjected to a linear warmup for the first two epochs followed by a cosine decay schedule without restarts [69]. All the implementations were done in Python using the PyTorch[1] framework.

3) Evaluation Datasets: We ran experiments on 8 large IQA databases spanning both synthetic and authentic distortions.

- Authentic Distortions
  - KonIQ [8] : contains 10k images sampled from the public media database YFCC100M [70].
  - CLIVE [7] : contains 1162 images captured from many diverse mobile devices.
  - FLIVE [9] : contains 40k real-world images and 120k patches along with respective quality scores. We only used images (and their corresponding scores) for analysis, and did not include patch information.
  - SPAQ [10] : contains 11k images captured using 66 smartphones. We only used images and their corresponding scores, and did not utilize the additional tag information available. Similar to [10], we resized the images before evaluation such that the shorter side is 512.
- Synthetic Distortions
  - LIVE-IQA [5] : contains 779 distorted images obtained from 29 pristine images using 5 synthetic distortion types.
  - CSIQ-IQA [6] : contains 866 distorted images obtained from 30 source contents with 6 types of distortions.
  - TID2013 [67] : contains 3000 distorted images obtained from 25 natural images with 24 distortion types, each having 5 levels of degradation.
  - KADID [68] : contains 10125 distorted images from 81 source contents spanning 25 different types of distortions.

4) Compared Methods: We compare the performance of CONTRIQUE against 14 SOTA NR IQA models. The compared methods can be categorized into 4 categories : (a) Traditional/hand-crafted features - BRISQUE [13] and NIQE [14]. (b) Codebook-based features - CORNIA [27] and HOSA [28]. Except NIQE, the rest use a support vector regressor (SVR) for quality prediction. (c) CNN based models - DB-CNN [18], MEON [30], WaDIQaM [59], PQR [20], BIECON [19], PaQ-2-PiQ [9], NIMA [31], HyperIQA [21] and MetaIQA [34]. (d) Transformer based

[1] https://pytorch.org/

TABLE I

PERFORMANCE COMPARISON OF CONTRIQUE AGAINST DIFFERENT NR MODELS ON IQA DATABASES CONTAINING **AUTHENTIC** DISTORTIONS. MODELS ARE CATEGORIZED BASED ON THE TYPE OF FEATURE EXTRACTION USED. IN EACH COLUMN, THE THREE BEST MODELS ARE BOLDFACED. ENTRIES MARKED '-' DENOTE THAT THE RESULTS ARE NOT AVAILABLE

Method	Model Type	KonIQ [8]		CLIVE [7]		FLIVE [9]		SPAQ [10]	
		SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
BRISQUE [13]	Traditional/ Handcrafted Features	0.665	0.681	0.608	0.629	0.288	0.373	0.809	0.817
NIQE [13]		0.531	0.538	0.455	0.483	0.211	0.288	0.700	0.709
CORNIA [27]	Codebook-based Features	0.780	0.795	0.629	0.671	-	-	0.709	0.725
HOSA [28]		0.805	0.813	0.640	0.678	-	-	0.846	0.852
DB-CNN [18]	Supervised pretraining and supervised fine-tuning	0.875	0.884	**0.851**	**0.869**	0.554	**0.652**	0.911	0.915
WaDIQaM [59]		0.797	0.805	0.671	0.680	-	-	-	-
PQR [20]		0.880	0.884	**0.857**	**0.882**	-	-	-	-
PaQ-2-PiQ [9]		0.870	0.880	0.840	0.850	**0.571**	0.623	-	-
HyperIQA [21]		**0.906**	**0.917**	**0.859**	**0.882**	0.535	0.623	**0.916**	**0.919**
MetaIQA [34]		0.850	0.887	0.802	0.835	-	-	-	-
MUSIQ [39]		**0.916**	**0.928**	-	-	-	-	**0.917**	**0.921**
Resnet-50 [15]	Supervised pretraining and Linear Regression	0.888	0.904	0.781	0.809	**0.595**	**0.648**	0.904	0.909
CONTRIQUE	Unsupervised pretraining and Linear Regression	**0.894**	**0.906**	0.845	0.857	**0.580**	**0.641**	**0.914**	**0.919**

model - MUSIQ [39]. For objective comparison of the above IQA models, we copied the numbers as reported by the respective authors or as available in the literature. For PaQ-2-PiQ, we consider the baseline model, since patch quality scores are not employed for training. We also included a Resnet-50 [15] model pretrained on Imagenet [22], using a similar linear regression module as CONTRIQUE to predict quality. This comparison enabled us to compare the effect of supervised and unsupervised pretraining techniques.

5) Evaluation Protocol: Two commonly used evaluation metrics Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) were employed to evaluate and compare the IQA models. Before computing PLCC, the quality predictions were passed through a four-parameter logistic non-linearity as described in [71].

Each dataset was randomly divided into 70%, 10% and 20% corresponding to training, validation and test sets, respectively. The validation set was used to determine the regularization coefficient of the regressor using grid search. On datasets with synthetic distortions, the splits were implemented based on reference images, to ensure no overlap of contents. To avoid any bias towards the choice of training set, we repeated the train/test split operation 10 times and reported the median performance. On FLIVE, due to the large size of the dataset, we used a single split as reported by the authors in [9].

B. Correlation Against Human Judgments

We compared the performance of CONTRIQUE against other models on IQA datasets containing authentic distortions in Table I. It may be observed from the table that CONTRIQUE achieves competitive performance when compared to other SOTA models. In the table, we categorized the models based on the type of feature extraction techniques. Notably CONTRIQUE achieves performance comparable to CNN based fine-tuned models even without fine-tuning, highlighting the effectiveness of our proposed self-supervision methodology. Furthermore, it outperformed Resnet-50 features, reinforcing the efficiency of the auxiliary task employed in CONTRIQUE. Note that MUSIQ [39], a transformer based IQA model, has a higher model complexity than CONTRIQUE and other CNN based models, which likely contributes to the better correlations.

In Table II model performances are compared on datasets with synthetic distortions. Here as well, CONTRIQUE achieved superior performance among the compared models, indicating a better generalizability of learned representations across both synthetic and authentic distortions.

C. Cross Dataset Evaluation

We conducted cross dataset evaluations whereby training and testing was performed on different datasets to analyze the dependence of training data, yielding the results reported in Table III. For simplicity we only include 4 datasets for comparison, two each from synthetic and realistic distortion sets. It may be inferred from the table that CONTRIQUE attains performance comparable to other IQA models across both synthetic and authentic distortions. Note that for CONTRIQUE, even for cross-dataset evaluations, only the weights of the linear regressor are modified depending on the training data, while the weights of the encoder backbone were kept intact.

D. Visual Comparison of Representations

The learned representations for CONTRIQUE are visualized in Fig. 2 using t-sne [72]. In the figure, for plotting purposes we used 4 commonly observed synthetic distortions: white noise, Gaussian blur, JPEG, and JPEG200, along with natural and UGC images. Each set contains 150 images, with synthetic distortions taken from the CSIQ-IQA dataset, while natural and UGC images sampled from the KADIS and KonIQ datasets, respectively. For comparison, we also include features from a Resnet-50 (Imagenet pretrained) in Fig. 2. Since the auxiliary task was to learn distortion discriminable embeddings, the learned CONTRIQUE features can be easily clustered depending on the type of distortions, as shown in

TABLE II

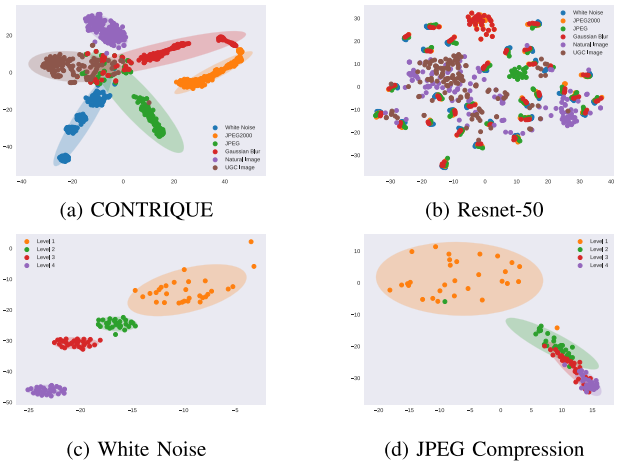
PERFORMANCE COMPARISON OF CONTRIQUE AGAINST DIFFERENT NR MODELS ON IQA DATABASES CONTAINING **SYNTHETIC** DISTORTIONS. MODELS ARE CATEGORIZED BASED ON THE TYPE OF FEATURE EXTRACTION USED. IN EACH COLUMN, THE THREE BEST MODELS ARE BOLDFACED. ENTRIES MARKED '-' DENOTE THAT THE RESULTS ARE NOT AVAILABLE

Method	Model Type	LIVE-IQA [5]		CSIQ-IQA [6]		TID2013 [67]		KADID [68]	
		SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
BRISQUE [13]	Traditional/Handcrafted Features	0.939	0.935	0.746	0.829	0.604	0.694	0.528	0.567
NIQE [14]		0.907	0.901	0.627	0.712	0.315	0.393	0.374	0.428
CORNIA [27]	Codebook-based Features	0.947	0.950	0.678	0.776	0.678	0.768	0.516	0.558
HOSA [28]		0.946	0.950	0.741	0.823	0.735	0.815	0.618	0.653
DB-CNN [18]	Supervised pretraining and supervised fine-tuning	**0.968**	**0.971**	**0.946**	**0.959**	**0.816**	**0.865**	**0.851**	**0.856**
MEON [30]		-	-	0.852	0.864	0.808	0.824	-	-
WaDIQaM [59]		0.954	0.963	-	-	0.761	0.787	-	-
PQR [20]		**0.965**	**0.971**	0.872	0.901	0.740	0.798	-	-
BIECON [19]		0.961	0.962	0.815	0.823	0.717	0.762	-	-
NIMA [31]		0.637	0.698	-	-	0.750	0.827	-	-
HyperIQA [21]		**0.962**	**0.966**	**0.923**	**0.942**	**0.840**	**0.858**	**0.852**	**0.845**
Resnet-50 [15]	Supervised pretraining and Linear Regression	0.925	0.931	0.840	0.848	0.679	0.729	0.701	0.677
CONTRIQUE	Unsupervised pretraining and Linear Regression	0.960	0.961	**0.947**	**0.958**	**0.861**	**0.871**	**0.934**	**0.937**

TABLE III

CROSS DATABASE SROCC COMPARISON OF IQA MODELS. IN EACH ROW TOP PERFORMING MODEL IS HIGHLIGHTED

Training	Testing	DB-CNN	PQR	HyperIQA	CONTRIQUE
CLIVE	KonIQ	0.754	0.757	**0.772**	0.676
KonIQ	CLIVE	0.755	0.770	**0.785**	0.731
LIVE-IQA	CSIQ-IQA	0.758	0.719	0.744	**0.823**
CSIQ-IQA	LIVE-IQA	0.877	0.922	**0.926**	**0.925**



(a) CONTRIQUE (b) Resnet-50

(c) White Noise (d) JPEG Compression

Fig. 2. Visualization of learned representations using t-sne. (c) and (d) are zoomed versions of the white noise and JPEG compression clusters shown in (a). (b) denotes Resnet-50 visualizations for same data shown in (a).

Fig. 2. However, the same is not true of Resnet-50 features, as they appear to be scattered across the space and did not form separable clusters. Fig. 2 also illustrates the degradation level separability of CONTRIQUE features for the white noise and JPEG compression distortions.

E. Significance of Training Data

During training of CONTRIQUE, we employed a mixed dataset containing both synthetic and realistic distortions. We conducted an ablation study whereby the effects of synthetic and authentic distortions were analyzed in isolation.

TABLE IV

SROCC PERFORMANCE COMPARISON OF DIFFERENT TRAININGS OF CONTRIQUE. *syn* AND *UGC* DENOTE MODELS TRAINED WITH DATA CONTAINING ONLY SYNTHETIC AND AUTHENTIC DISTORTIONS RESPECTIVELY. IN EACH COLUMN, THE TOP PERFORMING MODEL IS BOLDFACED

Model	KonIQ	CLIVE	LIVE-IQA	CSIQ-IQA
CONTRIQUE-syn	0.854	0.756	**0.965**	**0.950**
CONTRIQUE-UGC	**0.900**	0.843	0.918	0.802
CONTRIQUE	0.894	**0.846**	0.960	0.947

In this experiment, CONTRIQUE was trained with data containing either only synthetic or authentic artifacts, with the performance numbers reported in Table IV. From the Table, we can infer that training with only synthetic distortions boosts performance on synthetic IQA datasets, while the same holds true for authentic IQA datasets when trained on UGC data. Employing mixed data achieves better generalization, with negligible loss in performance as compared to the individual trainings.

F. Robustness to Training Data

The CONTRIQUE features are mapped to quality scores using a regularized linear regressor, as described in (5). The weights of the regressor are learned using the human opinion scores present in IQA datasets. In this experiment we studied the performance variation of CONTRIQUE with respect to the proportion of the IQA dataset employed for training. In Table V we report the variation in performance when the training set proportion of the IQA datasets was varied from 20% to 80%. In each case the remaining samples were used in the test set to obtain correlation values. From the table it may be observed that using as little as 20% of the dataset to train the regressor causes a maximum drop of only 8% against the highest attainable SROCC. This experiment highlights the robustness and generalizability of CONTRIQUE representations, and shows that competitive performance is achievable even with limited training samples.

TABLE V

SROCC PERFORMANCE VARIATION OF CONTRIQUE WITH TRAINING SET PROPORTION. IN EACH COLUMN, THE TOP PERFORMING MODEL IS BOLDFACED

Training set proportion	KonIQ	CLIVE	LIVE-IQA	CSIQ-IQA
20%	0.877	0.777	0.937	0.901
40%	0.886	0.804	0.948	0.920
60%	0.891	0.816	0.954	0.932
80%	**0.894**	**0.846**	**0.960**	**0.947**

TABLE VI

SROCC PERFORMANCE VARIATION OF CONTRIQUE FOR THE DIFFERENT COLOR SPACES USED DURING TRAINING. IN EACH COLUMN THE TOP PERFORMING MODEL IS BOLDFACED

Training Color space	KonIQ	CLIVE	LIVE-IQA	CSIQ-IQA
Grayscale	0.837	0.758	0.948	0.861
RGB	0.834	0.757	0.948	0.888
LAB	0.737	0.600	0.819	0.750
HSV	0.766	0.650	0.909	0.823
MS	0.870	0.800	**0.960**	0.903
All	**0.894**	**0.846**	**0.960**	**0.947**

TABLE VII

SROCC PERFORMANCE VARIATION OF CONTRIQUE WHEN EVALUATED ON DIFFERENT COLOR SPACES

Testing Color space	KonIQ	CLIVE	LIVE-IQA	CSIQ-IQA
Grayscale	0.891	0.846	0.960	0.911
RGB	0.894	0.846	0.960	0.947
LAB	0.886	0.838	0.953	0.921
HSV	0.889	0.843	0.960	0.941
MS	0.880	0.821	0.955	0.881

TABLE VIII

SROCC PERFORMANCE COMPARISON OF SINGLESCALE AND MULTISCALE TRAININGS OF CONTRIQUE. IN EACH COLUMN, THE TOP PERFORMING MODEL IS BOLDFACED

	KonIQ	CLIVE	LIVE-IQA	CSIQ-IQA
Singlescale	0.843	0.791	0.932	0.929
Multiscale	**0.894**	**0.846**	**0.960**	**0.947**

TABLE IX

SROCC PERFORMANCE COMPARISON FOR DIFFERENT CROP SIZES USED TO TRAIN CONTRIQUE. IN EACH COLUMN, THE TOP PERFORMING MODEL IS BOLDFACED

Crop Size	KonIQ	CLIVE	LIVE-IQA	CSIQ-IQA
64×64	0.532	0.371	0.136	0.201
128×128	0.886	0.818	0.967	0.944
256×256	**0.894**	**0.846**	**0.960**	**0.947**

TABLE X

DISTORTION SPECIFIC PERFORMANCE COMPARISON OF CONTRIQUE ON TID2013 [67] AND KADID [68] DATASETS. ONLY A SUBSET OF DISTORTION TYPES WITH LOW CORRELATION VALUES ARE REPORTED. THE LAST ROW INDICATES CORRELATION VALUES WHEN ALL DISTORTIONS PRESENT IN THE DATASET ARE INCLUDED

Distortion	TID2013 [67]		KADID [68]	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑
Mean shift	0.381	0.384	0.567	0.585
Non eccentricity	0.587	0.568	0.538	0.489
Color saturation	0.561	0.572	0.399	0.413
Contrast change	0.751	0.768	0.626	0.605
Local block-wise distortions	0.407	0.443	0.549	0.588
Overall	0.861	0.871	0.934	0.937

G. Importance of Different Color Spaces

In the CONTRIQUE, different color spaces were employed in order to extract complementary quality information. In this experiment we investigate the significance of each color space by training CONTRIQUE on each of them individually. The results are reported in Table VI, and it can be observed that combined training yields superior correlations than for any of the individual color spaces highlighting their complementary nature. Note that during evaluation, the images were converted to the respective color spaces on which they were trained.

Another interesting observation we make is the invariance property of the learned CONTRIQUE representations to the different color spaces. In other words, during evaluation, using any color space yielded approximately similar embeddings. This behavior is illustrated in Table VII, where during evaluation the images were converted to multiple color spaces. From the Table it can be inferred that the performances of CONTRIQUE remained approximately same across color spaces. This property is a consequence of using multiple color spaces during training. Furthermore, this property eliminates the need of changing color spaces during evaluation without significantly sacrificing performance.

H. Significance of Multiscale Learning

We included a multi-scale module in the training pipeline of CONTRIQUE, as shown in Fig. 1 to learn representations that can characterize local as well as global details. In this ablation experiment we studied the significance of multiscale learning by removing the downsampling module from the CONTRIQUE pipeline. In this case, two crops of the image from a single scale were employed instead of crops from native and half-resolution images. During the evaluation phase only single-scale features were used to learn the regressor mapping. Table VIII compares the correlation values obtained by single scale and multi-scale training. From the Table we can infer that using multiscale learning offers significant performance gains, underscoring the importance of capturing local and global image characteristics for IQA.

I. Effect of Batch Size and Crop Size

Fig. 3 plots the variation in SROCC performance against the batch size N used to train CONTRIQUE. From the plot it can be seen that larger batch sizes always yielded better correlations across all datasets. This observation is in line with those made in prior works using contrastive loss [25], [26], where larger batch size promotes better convergence.

In Table IX we study the performance variation of CONTRIQUE with different crop sizes M. From the Table it may be observed that larger crop sizes yield better correlations. This is expected since larger crops might contain more

TABLE XI

FULL REFERENCE PERFORMANCE COMPARISON ACROSS 4 IQA DATABASES. IN EACH COLUMN, THE FIRST AND SECOND BEST MODELS ARE BOLDFACED. ENTRIES MARKED '-' DENOTE THAT THE RESULTS ARE NOT AVAILABLE

Method	LIVE-IQA [5]		CSIQ-IQA [6]		TID2013 [67]		KADID [68]	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
PSNR	0.881	0.868	0.820	0.824	0.643	0.675	0.677	0.680
SSIM [73]	0.921	0.911	0.854	0.835	0.642	0.698	0.641	0.633
FSIM [74]	0.964	0.954	0.934	0.919	0.852	0.875	0.854	0.850
VSI [75]	0.951	0.940	0.944	0.929	0.902	0.903	**0.880**	**0.878**
PieAPP [76]	0.915	0.905	0.900	0.881	0.877	0.850	0.869	0.869
LPIPS [49]	0.932	0.936	0.884	0.906	0.673	0.756	0.721	0.713
DISTS [77]	0.953	0.954	0.942	0.942	0.853	0.873	-	-
DRF-IQA [45]	**0.983**	**0.983**	**0.964**	**0.960**	**0.944**	**0.942**	-	-
CONTRIQUE-FR	**0.966**	**0.966**	**0.956**	**0.964**	**0.909**	**0.915**	**0.946**	**0.947**

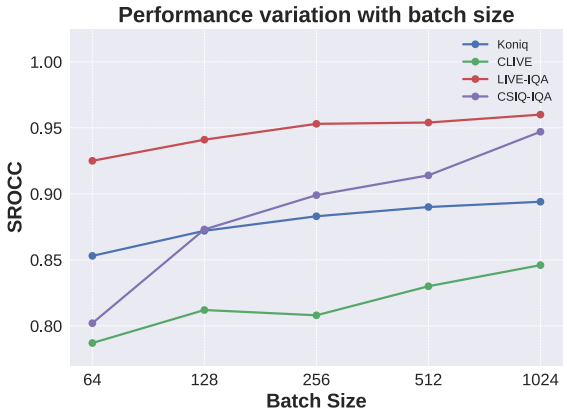


Fig. 3. Variation of SROCC values with training batch size of CONTRIQUE across 4 IQA datasets.

distortion relevant information which can help the learning of better representations.

J. Limitations of the Model

Table I and II show that CONTRIQUE obtains impressive performances on a variety of distortions. However, there exist certain distortion types on which CONTRIQUE features are less effective. For example, Table X reports performance of CONTRIQUE on distortions that are present in the TID2013 [67] and KADID [68] datasets, on which we observed low correlation values. We hypothesize that for these artifacts, the representations learned using distortion type and degree discrimination do not transfer well when used for quality prediction. Distortions such as non eccentricity pattern noise and local block-wise artifacts are highly localized. Thus, our assumption that cropped versions of pictures inheriting the distortion class of the original image used during training could be violated resulting in representations that are less sensitive to these corruptions.

V. CONTRIQUE FULL-REFERENCE MODEL

CONTRIQUE framework offers the flexibility to employ the learned representations on other IQA related tasks. We also propose a simple extension to employ CONTRIQUE representations in a Full (FR) IQA setting, where we have access to both pristine high quality reference images as well as their corresponding distorted versions. To incorporate reference information into the regressor, equation (5) is modified as

$$y = W|h_{ref} - h_{dist}|,$$
$$W^* = \underset{W}{\operatorname{argmin}} \sum_{i=1}^{N} (GT_i - y_i)^2 + \lambda \sum_{j=1}^{M} W_j^2, \tag{6}$$

where absolute difference between the features of reference and distorted images are used to predicting quality. We denote this modified model as CONTRIQUE-FR. Note that we do not perform any additional training or fine-tuning of the encoder network for the FR-IQA task. The same trained encoder obtained from CONTRIQUE was used with only the regressor modified to include reference information.

The performance of CONTRIQUE-FR is compared in Table XI. We followed a similar evaluation protocol of dividing datasets into 70%/10%/20% as train/validation/test sets, respectively based on content, and report the median correlation values over 10 different train/test splits. Since authentic IQA datasets do not contain reference images, we only report performances on the synthetic IQA datasets. For comparison, we include eight SOTA FR-IQA models : (a) Traditional models - PSNR, SSIM [73], FSIM [74] and VSI [75]. (b) Deep learning based models - PieAPP [76], LPIPS [49], DISTS [77] and DRF-IQA [45]. From the Table it can be observed that CONTRIQUE-FR achieves performance comparable to SOTA FR-IQA models, highlighting the flexibility as well as generalizability of the CONTRIQUE training framework. Additionally, comparing the CONTRIQUE correlation values in Table II and XI shows the performance gains due to the knowledge of the high quality reference images.

VI. CONCLUSION

We introduced an unsupervised training framework that learns effective image quality representations. Distinguishing characteristics of the proposed design include learning from unlabeled data, and employing distortion type and degree discrimination as an auxiliary task. We conducted holistic evaluations of our proposed model across multiple IQA databases, and found that CONTRIQUE achieves competitive performance against other, supervised IQA models. The proposed framework is simple, achieves superior performance

with no additional fine-tuning, and generalizes well across synthetic and realistic distortions. We conducted ablation experiments to understand the significance of different color spaces, and found surprisingly complementary quality prediction power among them. We also analyzed the importance of the distortion types present in the training data, and deduced that using a combination of synthetic and authentic artifacts helps achieve better generalization. We also proposed CONTRIQUE-FR, an extension of CONTRIQUE to FR IQA problem, which required no additional training of the CNN backbone. CONTRIQUE-FR also achieved comparable performance against SOTA FR-IQA models. A software release of CONTRIQUE and CONTRiQUE-FR has been made available online.[2]

ACKNOWLEDGMENT

The authors would like to thank YouTube for supporting this research and the Texas Advanced Computing Center (TACC) for providing computational resources that contributed to this work.

REFERENCES

[1] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, "Predicting the quality of images compressed after distortion in two steps," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5757–5770, Dec. 2019.

[2] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1258–1281, Apr. 2021.

[3] H. Talebi and P. Milanfar, "Learned perceptual image enhancement," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, May 2018, pp. 1–13.

[4] H. Zheng, H. Yang, J. Fu, Z.-J. Zha, and J. Luo, "Learning conditional knowledge distillation for degraded-reference image quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10242–10251.

[5] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Mar. 2006.

[6] 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. 1, 2010, Art. no. 011006.

[7] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jun. 2015.

[8] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.

[9] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3575–3585.

[10] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3677–3686.

[11] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[12] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[13] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[14] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[2]https://github.com/pavancm/CONTRIQUE

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.

[17] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[18] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2018.

[19] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2016.

[20] H. Zeng, L. Zhang, and A. C. Bovik, "A probabilistic quality representation approach to deep blind image quality prediction," 2017, *arXiv:1708.08190*.

[21] S. Su *et al.*, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3667–3676.

[22] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[23] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 766–774.

[24] P. Bojanowski and A. Joulin, "Unsupervised learning in noise," in *Proc. Int. Joint Conf. Neural Netw.*, 1989, pp. 517–526.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[27] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

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

[29] 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.

[30] 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 Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2017.

[31] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[32] H. Talebi and P. Milanfar, "Learning to resize images for computer vision tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 497–506.

[33] H. Talebi, E. Amid, P. Milanfar, and M. K. Warmuth, "Rank-smoothed pairwise learning in perceptual quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3413–3417.

[34] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14143–14152.

[35] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[37] H. Chen *et al.*, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.

[38] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[39] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5148–5157.

[40] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Represent.*, 2018.

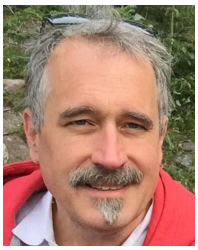
[41] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 649–666.
[42] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6874–6883.
[43] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.
[44] X. Liu, J. V. D. Weijer, and A. D. Bagdanov, "Exploiting unlabeled data in CNNs by self-supervised learning to rank," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1862–1878, Aug. 2019.
[45] W. Kim, A.-D. Nguyen, S. Lee, and A. C. Bovik, "Dynamic receptive field generation for full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4219–4231, 2020.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[47] P. Khosla *et al.*, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.
[48] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Jul. 2003, pp. 1398–1402.
[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[50] U. Rajashekar, Z. Wang, and E. P. Simoncelli, "Perceptual quality assessment of color images using adaptive signal representation," in *Proc. 15th Hum. Vis. Electron. Imag.*, vol. 7527, 2010, pp. 467–475.
[51] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, 2016.
[52] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," *IEEE Open J. Signal Process.*, vol. 2, pp. 425–440, 2021.
[53] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.
[54] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2018.
[55] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 7446–7457, 2021.
[56] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018.
[57] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.
[58] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1733–1740.
[59] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2017.
[60] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.
[61] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.
[62] H. Lin, V. Hosu, and D. Saupe, "DeepFL-IQA: Weak supervision for deep IQA feature learning," 2020, *arXiv:2001.08113*.
[63] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2408–2415.
[64] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[65] E. Mavridaki and V. Mezaris, "No-reference blur assessment in natural images using Fourier transform and spatial pyramids," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 566–570.
[66] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2010.
[67] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.
[68] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.
[69] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2017.
[70] B. Thomee *et al.*, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.
[71] *Final Report From the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment*, VQEG, Glasgow, U.K., 2000.
[72] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.
[73] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[74] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[75] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Aug. 2014.
[76] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1808–1817.
[77] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2020.

Pavan C. Madhusudana received the B.Tech. degree in electronics and communication engineering from the National Institute of Technology Karnataka (NITK), Surathkal, India, in 2016, and the M.Tech. (research) degree in electrical and communication engineering from the Indian Institute of Science (IISc), Bengaluru, India, in 2018. He is currently pursuing the Ph.D. degree in electrical and computer engineering with The University of Texas at Austin, USA. His research interests include image and video signal processing, computer vision, and machine learning.

Neil Birkbeck received the Ph.D. degree from the University of Alberta in 2011, working on topics in computer vision, graphics, and robotics, with a specific focus on image-based modeling and rendering. He went on to become a Research Scientist at Siemens Corporate Research working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Software Engineer at the Media Algorithms Team, YouTube/Google, with research interests in perceptual video processing, video coding, and video quality assessment.

Yilin Wang (Member, IEEE) received the B.S. and M.S. degrees in computer science from Nanjing University, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from The University of North Carolina at Chapel Hill in 2014, working on topics in computer vision and image processing. After graduation, he joined the Media Algorithm Team, Youtube/Google. His research fields include video processing infrastructure, video quality assessment, and video compression.

Balu Adsumilli received the master's degree from the University of Wisconsin–Madison in 2002 and the Ph.D. degree from the University of California at Santa Barbara in 2005, with a focus on watermark-based error resilience in video communications. From 2005 to 2011, he was a Senior Research Scientist at Citrix Online and from 2011 to 2016, he was a Senior Manager of advanced technology at GoPro, at both places developing algorithms for images/video quality enhancement, compression, capture, and streaming. He currently manages and leads the Media Algorithms Group, YouTube/Google. He has coauthored more than 120 papers and patents. His fields of research include image/video processing, machine vision, video compression, spherical capture, VR/AR, visual effects, and related areas. He is an Active Member of IEEE (and MMSP TC), ACM, SPIE, and VES.



Alan C. Bovik (Fellow, IEEE) is currently the Cockrell Family Regents Endowed Chair Professor at The University of Texas at Austin. His books include *The Essential Guides to Image and Video Processing*. As an Elected Member of the United States National Academy of Engineering, his research interests include image processing, digital photography, digital television, digital streaming video, social media, and visual perception. For his work in these areas, he was a recipient of the 2022 IEEE Edison Medal, the 2019 Progress Medal from The Royal Photographic Society, the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from The Optical Society, the 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, the 2020 Technology and Engineering Emmy Award from the National Academy for Television Arts and Sciences, and the Norbert Wiener Society Award and Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. He has also received about ten best journal paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. He has created/chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, in 1994. He has co-founded and was the longest-serving Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING.