CONVIQT: Contrastive Video Quality Estimator

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

Abstract—Perceptual video quality assessment (VQA) is an integral component of many streaming and video sharing platforms. Here we consider the problem of learning perceptually relevant video quality representations in a self-supervised manner. Distortion type identification and degradation level determination is employed as an auxiliary task to train a deep learning model containing a deep convolutional neural network (CNN) that extracts spatial features, as well as a recurrent unit that captures temporal information. The model is trained using a contrastive loss and we therefore refer to this training framework and resulting model as CONtrastive Video Quality Estimator (CONVIQT). During testing, the weights of the trained model are frozen, and a linear regressor maps the learned features to quality scores in a no-reference (NR) setting. We conduct comprehensive evaluations of the proposed model against leading algorithms on multiple VQA databases containing wide ranges of spatial and temporal distortions. We analyze the correlations between model predictions and ground-truth quality ratings, and show that CONVIQT achieves competitive performance when compared to state-of-the-art NR-VQA models, even though it is not trained on those databases. Our ablation experiments demonstrate that the learned representations are highly robust and generalize well across synthetic and realistic distortions. Our results indicate that compelling representations with perceptual bearing can be obtained using self-supervised learning.

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

I. INTRODUCTION

The smartphone revolution has led to an explosion of video consumption over the internet. A recent report by Cisco [1] estimates that more than 82% of internet traffic will be dominated by online videos. Popular video sharing and streaming platforms such as YouTube, Facebook, Netflix, and Amazon Prime Video are accessed by hundreds of millions of people worldwide, and more than one billion hours of video are watched every day. Given the central nature of videos as a primary medium of communication, it is important that these platforms be able to monitor and control perceptual video quality and to deliver better consumer experiences.

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 Prof. Joao M. Ascenso. (Corresponding author: Pavan C. Madhusudana.)

Pavan C. Madhusudana is with Samsung Research America, Plano, TX 75023 USA (e-mail: pavan.madhusudana@gmail.com).

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).

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

Data is available on-line at https://github.com/pavancm/CONVIQT

Digital Object Identifier 10.1109/TIP.2023.3310344

The problem of objectively quantifying and estimating perceptual video quality is referred to as Video Quality Assessment (VQA). Although subjective VQA, where quality judgments are obtained by human observers, provides the most accurate and reliable quality estimates, it is generally impractical, and does not scale with the number of videos being assessed. This necessitates the development of objective VQA models that can reliably predict the responses of human observers on the task of quality estimation. The goal of VQA models is to obtain predictions that correlate well with subjective opinions.

Due to the practical significance of NR-VQA, a diverse set of VQA databases and objective models have been created over the last decade. A VQA database contains a set of distorted videos, along with corresponding human opinion scores of them obtained either in a laboratory or a crowd-sourced environment. Earlier databases, such as LIVE-VQA [2], CSIQ-VQA [3], EPFLPoliMI [4], and MCL-V [5] datasets each contain a small set of pristine, high quality source content corrupted with compression and transmission artifacts. Each source video was synthetically degraded by only one or two types of distortions applied in a controlled manner. Although these databases were influential in improving video quality models, they do not represent the reality of pervasive UGC videos that are commonly corrupted by innumerable combinations of multiple distortions. Because of this, several large scale databases that contain authentic UGC distortions are now available such as LIVE-VQC [6], KoNViD [7], YouTube-UGC [8] and LSQV [9]. The UGC-VQA problem is particularly challenging due to the large variety of distortions that can occur, as the influence of video content on perceived quality. Advances in capture and display technologies has led to the widespread adoption of videos of 4K/ultra high definition (UHD) resolution, high dynamic range (HDR) and high frame rate (HFR). This has led to the creation of datasets like AVT-VQDB-UHD-I [10], ETRI-LIVE STSVQ [11], BVI-HFR [12] and LIVE-YT-HFR [13] which now make it possible to analyze the perceptual implications of these new developments. From an NR-VQA algorithm design standpoint, having a single reliable model that is able to accurately quantify both synthetic and authentic artifacts would be quite advantageous, since it could be used in wide ranges of scenario presenting highly diverse of the distortions and combinations of them.

Popular NR-VQA models employ a two stage design, where in the first stage quality relevant features are extracted followed by a regression stage that maps the features to perceptual quality scores. Earlier techniques relied heavily on the use of...
of Natural Scene Statistics (NSS) to derive features that are highly sensitive to visual quality. Although these models have been successful at predicting the perception of synthetically applied distortions, their performance has proven to be limited on the realistic distortions encountered in UGC videos. End-to-end trained deep learning based VQA models [9], [14], [15], [16] have achieved significant improvements over these traditional methods on UGC content. These performance improvements have mainly been arrived at using transfer learning techniques, whereby a deep model is first trained on a large labeled dataset such as ImageNet [17] or Kinetics-400 [18], then later fine-tuned on human opinion scores. Transfer learning is employed since currently available VQA datasets are generally too small to train deep models from scratch. Although fine-tuning can be an effective way to obtain good performance on both synthetic and realistic distortions, it requires careful choice of hyperparameters, which may change on different VQA databases.

A recent still picture (spatial) model called CONTRIQUE [19] employs a self-supervised training approach, where a deep model was trained to distinguish between different types of image artifacts and also to determine the degree of perceptual degradation present. The main advantage of this approach is that it does not require human opinion scores of visual quality for training, making model training feasible on unlabeled image databases. The performance of CONTRIQUE was shown to be highly competitive on multiple image quality datasets comprising both synthetic and authentic artifacts. Motivated by the successes of CONTRIQUE, here we extend the idea to video quality prediction, which is a much harder problem of higher dimensionality. We refer to the new model as the CONtrastive VIdeo Quality EstimaTor (CONVIQT). Notable attributes of CONVIQT are as follows:

1) Similar to CONTRIQUE, the CONVIQT model is first trained from scratch to solve the auxiliary tasks of distortion type and degradation level discrimination. Training is done on an unlabeled video dataset containing a mix of synthetic and realistic video distortions, using a contrastive loss function.

2) The CONVIQT architecture includes the existing CONTRIQUE model which is used to capture spatial attributes of video quality. The weights of the CONTRIQUE model are frozen and not modified during the training of CONVIQT.

3) In addition, recurrent unit processes the spatial features to extract temporal quality information. The recurrent unit is the only trainable component in the CONVIQT pipeline.

4) During evaluation of the CONVIQT representations, all of the weights are frozen and a linear regressor on each VQA dataset learns to map CONVIQT features to quality scores. The quality predictions produced by CONVIQT obtain competitive performance as compared to other state-of-the-art (SOTA) VQA models across multiple databases even though those models are trained on them. This is achieved without any additional fine-tuning of the backbone deep learning model.

5) The CONVIQT framework is simple but effective, and generalizes well across synthetic and realistic distortions.

The rest of the paper is organized as follows: In Section II we discuss related work on VQA and self-supervised learning. Section III provides a detailed description of the design of CONVIQT. Section IV analyzes and compares various experimental results obtained with CONVIQT, and Section V provides concluding remarks.

II. RELATED WORK

Next we discuss related work on NR-VQA and self-supervised learning.

A. NR-VQA Models

Objective VQA models can be broadly categorized based on the availability of reference information: Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR). FR/RR VQA [20], [21], [22], [23] models employ high quality pristine reference videos against which the perceptual fidelity of distorted videos are compared when determining quality, while NR models make judgments without any reference information. Here our primary focus will be on NR-VQA methods.

Blind assessment of perceptual video quality is challenging due to the extreme diversity of spatio-temporal artifacts that may be present. Moreover, the perception of these distortions is deeply affected by the video content, adding additional difficulty to the VQA problem. Although NR Image Quality Assessment (IQA) models can be used to conduct VQA in a simple way by computing quality scores at the frame-level, then temporally pooling them, this often neglects vital temporal quality information. Prior works [24], [25] have demonstrated the significance of motion in VQA. Thus, to obtain better performance, VQA models need to account for the spatio-temporal nature of video distortions and content. Early VQA models used extensive domain knowledge about video artifacts in their feature extraction frameworks, employing regressors to map features to quality scores. Using features derived from Natural Scene Statistics (NSS) and distorted image and video statistics models [26] has been a popular example employing this approach, whereby the videos are first transformed to a band-pass domain, where deviations from expected statistical regularities of the band-pass coefficients due to distortion are used to capture perceptual aspects of quality. Examples of such VQA models include VBLIINDS [27], Li et al. [28], which uses bandpass DCT coefficients, and BRISQUE [29], NIQE [30], VIDE [31] and Dendi et al. [32] which employ mean subtracted contrast normalized (MSCN) bandpass coefficients to obtain perceptually relevant quality features. V-CORNIA [33] employs codebook based unsupervised feature learning to conduct frame quality prediction, and hysteresis pooling [34] to obtain final video quality predictions. Korhonen proposed a two-level quality model TLVQM [35] containing a comprehensive set of low complexity features extracted from all frames present in the video, along with high complexity features such as spatial
activity, exposure etc. computed on a representative subset of the frames. VIDEVAL [36] uses an ensemble approach whereby features from multiple NR models were carefully selected, yielding a feature set that was shown to perform well on UGC videos.

Recently, end-to-end trained Deep Neural Networks (DNN) have achieved significant performance gains on NR-VQA problems. Since there is lack of large scale VQA datasets to pretrain DNNs from scratch, most of the DNN based models employ transfer learning, whereby a pretrained DNN (pre-trained on large datasets like ImageNet [17] or Kinetics-400 [18]) is fine-tuned in a supervised manner on human opinion scores. The RAPIQUE model proposed by Tu et al. [37] employs a hybrid set of features derived from NSS and an ImageNet pretrained CNN. Zhang et al. [38] used Full Reference (FR) VQA model predictions as weak labels to train a Convolution Neural Network (CNN), then used this weakly trained CNN to conduct NR-VQA. V-MEON [39] uses a multi-task training objective, where codec classification and video quality regression are jointly optimized. VSFA [14] employs ImageNet pretrained CNN to extract content dependent features, along with a Gated Recurrent Unit (GRU) to model temporal memory effects. Ying et al. [9] demonstrated that features combining local patch quality and global video quality can yield significant performance gains. GSTVQA [15] employs a pyramid temporal aggregation of short-term and long-term memory effects to achieve efficient quality prediction. Li et al. [16] employs transfer learning on IQA databases, along with a list-wise ranking loss objective to achieve competitive performance. In the Fast-VQA model proposed by Wu et al. [40], temporally aligned mini-patches sampled from videos are employed to train a transformer-based architecture to obtain effective VQA representations. Recently, a few methods that seek to obtain superior cross-dataset generalization have been proposed. These methods [41], [42], [43] primarily train a model using a mixed-training strategy, whereby a single model is trained on multiple IQA/VQA datasets. UNIQUE [42] deploys a unified, uncertainty-aware approach to the NR-IQA problem, whereby pairwise ranks of images are employed as supervisory signals, thereby providing better cross-distortion generalization. A continual learning approach was proposed in [43], where a model was learnt from a stream of datasets, building on previously seen data to obtain robust blind IQA models. All of the above methods employ supervised fine-tuning techniques to obtain good performances. By contrast, we focus on self-supervised feature learning with no requirement of human quality labels during training, and no additional fine-tuning during evaluation.
**B. Self-Supervised Learning**

Self-supervised techniques rely on generating supervisory signals from unlabeled data. This is achieved by using an auxiliary/proxy task from which labels can be easily obtained. The auxiliary task is usually closely related to the original task, so that model trained for the auxiliary problem can be easily applied to the original task. Example tasks that have been previously studied include rotation prediction [44], predicting the relative positions of image patches [45], image colorization [46], [47], and image inpainting [48]. Recent works employing contrastive learning [49], [50], [51], [52], [53] have used instance discrimination to distinguish a data samples and augmented versions of them from other input data samples, to achieve significant performance gains on multiple computer vision problems.

Several of the above ideas have been extended to learn video representations. Fernando et al. [54] employed rankings of frame level video features in chronological order as a proxy, to learn representations for action recognition. In [55], a temporal coherency measure is used as a supervisory signal to learn video representations. Han et al. [56] proposed Dense Predictive Coding (DPC) framework by recurrently predicting future representations. In [57], video playback rate perception is used as an auxiliary task for learning video embeddings. The successes of contrastive learning on pictures have been extended to videos in [58], [59], [60], and [61]. A recently proposed self-supervised VQA model called CSPT [62] employs a combination of contrastive and cross-entropy losses to obtain quality representative features. CSPT has three objectives: distortion-wise contrast prediction, content-wise contrast and distortion prediction. For the first two objectives, a contrastive loss is used to differentiate between distortions and content, respectively, while to optimize distortion predictions, a cross-entropy is employed. The contrastive loss is calculated between every pair of data present in the batch, while the cross-entropy is computed over all the samples present in a given batch of training samples. This combination of two different types of losses can make model training a bit involved. Our proposed method greatly differs from CSPT, since we employ a pure contrastive loss trained model (no other types of losses are used), with the main goal of discriminating different distortion types and perceptual degrees of degradation. Unlike CSPT, our training objective is based on supervised contrastive loss [63], enabling the inputs to contain multiple positive samples, which greatly simplifies the training process. Also, our objective does not contain the content-wise contrast objective used in CSPT, hence our objective only contains positive samples belonging to different contents but having same distortion and degradation factors. Moreover, our training data is much larger and encompasses a wider range of distortions compared to CSPT. This formulation allows us to achieve significantly better performance than CSPT.

**III. METHOD**

Our goal here is to learn generic video representations that may be used to predict video quality, with no requirement of having ground-truth quality labels during training. Given a video $x$ of dimensions $H \times W \times T$, where $H$, $W$ and $T$ correspond to height, width and number of frames respectively, we are interested in learning a transformation $p : \mathbb{R}^{3 \times H \times W \times T} \mapsto \mathbb{R}^r$ that maps $x$ to a $r$-dimensional representation $h$. Prior NSS inspired VQA models have employed band-pass transformations such as the DCT [27], [28], local mean-subtraction [29], [31], [32] and so on, to obtain
perceptually relevant quality features. Recent models employing deep learning based transforms [9], [14], [16], [62] have also been highly successful in capturing perceptual video distortions.

Our proposed training workflow is illustrated in Figs. 1 and 2. The pipeline can be divided into two parts: (a) Spatial component - spatial features are extracted from every video frame using CONTRIQUE [19] model as shown in Fig. 1. (b) Temporal component - a recurrent unit maps the observed temporal variations of spatial features to a r-dimensional feature vector \( h \) as illustrated in Fig. 2. In the following sections, each component present in the pipeline is described in detail.

### A. Proxy/Auxiliary Task

In self-supervised learning the original problem is modified into an alternate but closely related auxiliary task, for which ground-truth labels are easily available or can be obtained. A model is then trained to solve the auxiliary task. During the inference stage the trained model is evaluated on the original problem. Inspired by the success of CONTRIQUE in capturing still picture distortions, we follow the idea of learning representations that can discriminate different distortion types, as well as the strength of degradations. Thus, the auxiliary task is a classification problem, where each class contains videos with similar distortion type and degree of degradation.

#### 1) Generating Data for Auxiliary Task:

Since the auxiliary task requires a database of videos spanning different distortion types and degradation levels, the first step is to create a database of synthetically distorted videos. Here we mainly focus on three widely studied synthetic video distortions [2], [3], [10], [11], [12], [13]: compression, scaling, and temporal artifacts arising due to changes in frame rate. Of course, while these are synthetically applied, they are the same as those encountered in genuine practice. To obtain holistic learning of different artifact types, the created dataset should contain videos which reflect individual as well as the combined effects of the above distortion types. A pristine high quality source sequence \( x \) is provided as input to the distorted video generator module. A high level block diagram illustrating the workflow used to obtain synthetically distorted sequences is shown in Fig. 3. The description of each of the blocks shown in Fig. 3 is provided below.

- **Frame Rate Control**: This block modifies the frame rate of the source sequence, and generates videos having frame rates equal to or smaller than that of the source video. Similar to the LIVE-YT-HFR database [13], the frame rates were chosen from the set \( F = \{24, 30, 60, 82, 98, 120\} \). The values in the set \( F \) denote frames per second (fps). Note that we only choose values from \( F \) that are either equal to or smaller than the source frame rate. For example, for a source video having 120 fps, all of the values in \( F \) are chosen, while for a 60 fps source video only three values (60, 30 and 24 fps) are used. The low-frame rate versions were obtained by dropping frames by using suitable values in the arguments of the \( fps \) filter present in FFmpeg [64].

- **Compression**: The output videos from the frame rate control block are fed as input to the scaling module, which modifies the spatial resolutions of the videos. Scaling block consists of a downsampling module preceding a compression encoder, both followed by an upsampler, so that the resulting video has the same resolution as the original. This is done to replicate scenarios usually seen in streaming and social media platforms, where videos are encoded after scaling them to lower spatial resolutions before transmission. Encoding at lower resolutions might be preferable when either the content has low complexity leading to minimal loss when downsampled, or when there are limitations on the bandwidth available for transmission [65], [66], [67]. The Lanczos kernel was used when downsampling and upsampling the videos. To ensure that the aspect ratios of the videos remained unchanged after downsampling, the resizing factors were chosen from the set \( D = \{1, 2, 4, 8\} \). These values are similar to those used in the AVT-VQDB-UHD-1 [10] database. We also did not include downsampling factors that would have resulted in videos having resolutions less than 240p (320 × 240), since these are rarely used. For example, a 1080p source content will therefore only be subject to 5 levels of VP9 compression using single pass FFmpeg encoding in constant quality mode. Constant Rate Factor (CRF) values were varied to obtain different compression levels. The CRF values were chosen from

| Database          | Number of Videos | Number of Scenes | Resolution | Time Duration | Distortion Type                  |
|-------------------|------------------|------------------|------------|---------------|----------------------------------|
| KoNViD [7]        | 1200             | 1200             | 540p       | 8s            | Diverse distortions (authentic / UGC) |
| LIVE-VQC [6]      | 585              | 585              | 240p-1080p | 10s           |                                  |
| YouTube-UGC [8]   | 1380             | 1380             | 360p-4K    | 20s           |                                  |
| LSXQ [9]          | 39075            | 39075            | 99p-4K     | 5-12s         |                                  |
| CV2D1024 [68]     | 234              | 5                | 480p       | 10-25s        |                                  |
| LIVE-Qualcomm [69]| 208              | 54               | 1080p      | 15s           |                                  |
| LIVE-VQA [2]      | 160              | 10               | 768 × 432  | 10s           | Compression, scaling, transmission |
| CSIQ-VQA [3]      | 228              | 12               | 832 × 480  | 10s           |                                  |
| ETRI-LIVE STSVQ [11] | 437            | 15               | 540p-4K    | 4-7s          |                                  |
| AVT-VQDB-UHD-1 [10]| 432             | 10               | 360p-4K    | 8-10s         |                                  |
| LIVE-YT-HFR [13]  | 480              | 16               | 1080p-4K   | 8-10s         | Frame rate, compression          |
| BVI-HFR [12]      | 88               | 22               | 1080p      | 10s           |                                  |
the set $CF = \{\text{lossless, 24, 36, 48, 63}\}$, and the same CRF values were applied to all of the source contents. The values in the set $CF$ were chosen to include extreme cases (lossless and CRF=63, the latter corresponding to the highest possible compression level in VP9) as well as the wide range of compression levels between these. These CRF levels are similar to the ones used in the LIVE-YT-HFR [13] database. Here, we make the simplifying assumption that the CRF values can be used as proxies to measure the degree of compression.

Frame rate changes may introduce temporal artifacts such as judder and strobing, while spatial downsampling and upsampling procedures can introduce scaling distortions. Our approach to generating distortions, coupled with compression, aims to realize realistic video representation currently available over internet, social media and streaming platforms. These kind of videos are included so that the learned representations will be sensitive to the synthesized distortion types.

Let a pristine source sequence $x$ be subjected to the distortion pipeline shown in Fig. 3, resulting in a distorted sequence $x^{f,dc}$ where $f \in F$, $d \in D$, and $c \in CF$. We define the auxiliary task of identifying the values $f$, $d$ and $c$ on a given distorted video $x^{f,dc}$. This defines a classification problem having $|F| \times |D| \times |CF|$ classes, where $|.|$ is the set cardinality. In our experimental setting, $|F| = 6$, $|D| = 4$, and $|CF| = 5$, resulting in a total of 120 classes.

### B. Contrastive Loss

Motivated by the success of using contrastive loss for learning representations in CONTRIQUE [19], we follow a similar approach to capture space-time video distortions in CONVIQT. In the design of CONTRIQUE the primary focus was on capturing wide ranges of spatial artifacts commonly seen in still pictures, by contrasting different spatial distortions during training. The primary focus of CONVIQT is instead on characterizing the variations observed along the spatio-temporal dimensions, with more emphasis on the temporal axis. This is achieved by employing diverse sets of temporal distortions coupled with quality preserving transformations to allow CONVIQT to learn more robust video quality representations. To obtain embeddings, we defined a deep model consisting of three components: a spatial encoder, a recurrent unit and a projector. The spatial encoder extracts features on every frame of the input video, then these features are fed to the recurrent unit. The recurrent unit extracts temporal information from these features. The projector is a fully connected network (FCN) that reduces the dimensionality of the features produced by the recurrent unit. Since CONTRIQUE representations have been shown to achieve impressive performance in capturing spatial artifacts, we employ CONTRIQUE with frozen weights as a spatial encoder, and the CONTRIQUE weights are not modified during training. Thus, the recurrent unit and the projector network are the trainable components of CONVIQT. Let $a(.)$ and $b(.)$ denote the recurrent unit and the projector network, respectively. For a given video $x \in \mathbb{R}^{3 \times H \times W \times T}$

$$y^{(t)} = CONTRIQUE(x^{(t)}), \quad t = \{1, 2, \ldots, T\}$$

$$h(t) = a(y^{(t)}, h^{(t-1)}), \quad h = \frac{1}{T} \sum_{t=1}^{T} h(t),$$

$$z = b(h)h \in \mathbb{R}^r, \quad z \in \mathbb{R}^K$$

where $x^{(t)}$ is the $t^{th}$ frame of video $x$, and $h$ is the $r$-dimensional temporal pooled output from the recurrent unit. The representation $h$ is $L_2$ normalized [19], [52], [53] before being fed to the projector. The projector reduces the feature dimensions from $r$ to $K$ ($K$ is a hyperparameter in this setting). The objective is to obtain similar embeddings $z$ for videos belonging to the same class. The cosine distance between the representations $z_i$ and $z_j$, given by $\phi(z_i, z_j) = z_i^T z_j / ||z_i||z_j||z_j||$ is used as a similarity measure. This loss function is an application of the supervised contrastive loss proposed in [63], where it was proposed in the context of image classification against ground-truth class labels. Here we have applied this objective on the synthetically distorted videos and their corresponding distortion classes, which were obtained by the procedure detailed in Sec. III-A. Additionally we used task specific data transformations, different from those in [63], to obtain more effective video quality representations. Given a video $x_i$ modified by synthetic distortions, the loss $\mathcal{L}_i^{syn}$ is defined as

$$\mathcal{L}_i^{syn} = \frac{1}{|V(i)|} \sum_{j \in V(i)} - \log \frac{\exp(\phi(z_i, z_j) / \tau)}{\sum_{m=1}^{N} \mathbf{1}_{m \neq i} \exp(\phi(z_i, z_m) / \tau)},$$

where $N$ is the number of videos present in the batch, $I$ is the set indicator function, $\tau$ is a temperature parameter, $V(i)$ is a set containing video indices belonging to the same class as $x_i$ (but excluding the index $i$), and $|V(i)|$ is its cardinality. For example, if $x_i$ is a video which was encoded at 540p, 60fps and CRF=24, $V(i)$ will contain indices of all videos (excluding index $i$) which were encoded at 540p, 60fps and CRF=24 setting. There exists transformations of a video $x_i$ (video transformations are discussed in Sec. III-C) guaranteeing that $V(i)$ contains at least one sample. Note that a distorted video’s class is uniquely determined by its frame rate, spatial subsampling factor, and compression level, and is independent of the native resolution of the source video. For example, a 4K 120fps source encoded at 1080p, 60 fps, and CRF=48 will belong to the same class as that of a 1080p 120 fps source encoded at the 540p, 60 fps, and CRF=48, since both are spatially downsampled twice before encoding. This resolution independent class setting enables us to employ source videos spanning multiple spatial resolutions during training, without requiring any need to resize videos to predetermined sizes. The objective defined in (4) is a normalized temperature-scaled cross entropy (NT-Xent), and is calculated between every pair of videos present in the batch.

### C. Multiscale Learning and Temporal Transformations

It has been widely established that using spatial multi-scale representations of images leads to better quality representations [29], [30], [70], [71] since image quality is influenced by both local and global characteristics. By extension, videos
also exhibit multi-scale characteristics in time. Prior VQA models [27], [29], [37] have achieved significant gains in performance by employing multi-scale representations. In CONVIQT, we use two scales: native resolutions, and half-scale resolutions obtained by downscaling, twice along both spatial dimensions. An anti-aliasing filter is used before downscaling as shown in Fig. 1 to avoid aliasing artifacts. Note that this downscaling operation preserves the aspect ratio of the input video, which is desirable since changing this could modify the perceived quality of the input video.

The aim of the objective function in (4) is to obtain discriminative representations that can differentiate different video classes, and also demonstrate invariance to quality preserving transformations. Operations on videos which preserve video quality we will collectively refer to as quality preserving transforms. Depending on the coordinates on which video transforms are applied, they can be further categorized as spatial and temporal transforms. Spatial transforms are applied independently on each video frame, while temporal transforms are applied along the time axis. Since the CONTRIQUE model is only used to obtain spatial features, and during training of CONTRIQUE, spatial transforms such as horizontal flipping, color-space conversion were employed, we do not include any spatial transforms when training CONVIQT.

We employ temporal band-pass transforms inspired by similar successes [22], [23], [37], [72], [73] capturing temporal distortions. Temporal band-pass coefficients demonstrate reliable statistical regularities on very high quality videos, but these are predictably disturbed by the presence of distortions. We use them to access temporal quality information over multiple scales (subbands). Similar to [22], we employ three temporal band-pass filters: Haar, Daubechies-2 (db2), and Biorthogonal2.2 (bior2.2) with wavelet packet decomposition [74]. In each case, we use three levels of wavelet decomposition resulting in 7 subbands (ignoring the low pass response). During training, the wavelet filter and subbands were randomly chosen for each input video. The transformed videos at two spatial scales were then fed as input to the CONTRIQUE model to obtain spatial representations. Since the CONTRIQUE weights remain unchanged during CONVIQT training, the spatial representations of all the transformed videos present in the training data can be pre-computed and stored in the disk. This significantly reduces the overall training time of CONVIQT.

To facilitate batch training, the videos are randomly cropped along the temporal axis to length of \( T \) frames. This ensures that every video in the batch contained the same number of frames, which is necessary when training deep networks. We presume that this cropping operation did not alter the distortion class, and that the cropped portion inherits the class of the original video. Temporal cropping is performed on every video at both full-scale and at half-scale. We ensured that all of the videos in the training set contain at least \( T \) frames to avoid zero padding.

### D. Realistic Distortions

In the contrastive objective (4), prior knowledge of the distortion types and degradation degrees were used to learn video representations. However, this kind of information is usually not readily available for UGC videos, despite the massive volumes of streamed and shared on social media platforms like YouTube, Facebook, Instagram and TikTok. Given their consumer ubiquity, it is vital that the learned embeddings be sensitive to the realistic artifacts found on UGC sequences, which often include mixtures of multiple unknown distortion types. Additionally, the classes used in (4) are not even remotely exhaustive enough to span the set of all possible real-world distortion types. Indeed, synthetic distortion types can only be poorly regarded as relevant to UGC videos. Since identifying the distortion type and degree of degradation of UGC videos is an intractable task (even for human eyes), we instead simplify the problem by treating each UGC video as a unique class containing a distinctive blend of multiple distortions, separate and different from all the other UGC videos, as well as from synthetically distorted sequences. Thus, given a UGC video \( x_j \), only its scaled (and temporally transformed) version \( x_j \) belongs to the same class. The loss objective in (4) is modified to reflect this change as

\[
\mathcal{L}_{\text{UGC}}^{i} = -\log \frac{\exp(\phi(z_i, z_j)/\tau)}{\sum_{m=1}^{N} \exp(\phi(z_i, z_m)/\tau)}, \quad (5)
\]

This is essentially the instance discrimination objective used in prior work [52], [53]. The underlying presumption here is that the training data containing UGC videos contain a wide variety of realistic distortions, and can yield more efficient representations. Combining (4) and (5), the overall training objective is then

\[
\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{(x, y \neq \text{UGC})} \mathcal{L}_{\text{syn}}^{i} + \mathbb{I}_{(x, y \neq \text{UGC})} \mathcal{L}_{\text{UGC}}^{i}, \quad (6)
\]

where \( N \) is the total number of videos present in the batch, and \( \mathbb{I} \) is the indicator function determining whether the input video is UGC. To avoid any bias during training, at each iteration we sampled equal numbers of synthetic and authentically distorted videos.

### E. Evaluating Representations

The learned representations are evaluated with respect to their video quality prediction power by correlating their responses against human opinion scores. After the CONVIQT model is trained, the projector network \( g(.) \) is discarded and the outputs of the recurrent unit \( h = f(y^{(1)}, \ldots, y^{(T)}) \) are used as video quality embeddings. A regularized linear (ridge) regressor is trained on top of the frozen CONVIQT model in order to map embeddings to video quality scores. The regression expression is given by

\[
y_j = CONTRIQUES(x_j), \quad t = \{1, \ldots, T\}, j = \{1, \ldots, N\}
\]

\[
h_j = f(y_j^{(t)}, h_j^{(t-1)}), \quad h_j = \frac{1}{T} \sum_{t=1}^{T} h_j^{(t)},
\]

\[
p = Wh, \quad W^* = \arg \min_{W} \sum_{j=1}^{N} (GT_j - p_j)^2 + \lambda \sum_{i=1}^{d} W_i^2,
\]

where \( GT_j \) are ground-truth quality scores, \( p_j \) are predicted scores, \( W \) is a trainable vector having the same dimensions.
TABLE II
Performance Comparison of CONVIQT Against Different NR Models on VQA Databases Containing Authentic Distortions. In Each Column, the Three Best Models are Boldfaced. Entries Marked '-' Denote That the Results Are Not Available.

| Method       | Model Type       | KoNVd [7] | LIVE-VQC [6] | YouTube-UGC [8] | Test | Test-FHDp |
|--------------|------------------|-----------|---------------|------------------|------|-----------|
|              |                  | SROCC t   | PLCC t        | SROCC t          | PLCC t | SROCC t   | PLCC t   | SROCC t   | PLCC t   | SROCC t   | PLCC t   | SROCC t   | PLCC t   |
| BRSQUE [29]  |                  | 0.649     | 0.651         | 0.593            | 0.624 | 0.393     | 0.407     | 0.579     | 0.576     | 0.497     | 0.531     |
| VBIINDS [27] | Traditional      | 0.706     | 0.701         | 0.681            | 0.699 | 0.534     | 0.540     | -         | -         | -         | -         |
| TLQVM [35]   | Handcrafted      | 0.758     | 0.759         | 0.787            | 0.794 | 0.656     | 0.647     | 0.772     | 0.774     | 0.589     | 0.616     |
| VIDEVAL [36] | Features         | 0.770     | 0.770         | 0.743            | 0.747 | 0.776     | 0.771     | 0.794     | 0.783     | 0.545     | 0.554     |
| RAPIQUE [36] |                 | 0.784     | 0.805         | 0.741            | 0.761 | 0.747     | 0.756     | -         | -         | -         | -         |
| VSPA [14]    | Supervised       | 0.794     | 0.798         | 0.717            | 0.770 | 0.801     | 0.784     | 0.675     | 0.704     |           |           |
| PVQ [9]      | Pretraining      | 0.791     | 0.786         | 0.827            | 0.837 | -         | -         | 0.814     | 0.816     | 0.686     | 0.708     |
| GSTDQA [15]  |                 | 0.814     | 0.825         | 0.784            | 0.787 | -         | -         | -         | -         | -         | -         |
| Li et al. [16]|                 | 0.835     | 0.833         | 0.841            | 0.839 | 0.825     | 0.817     | 0.852     | 0.853     | 0.771     | 0.787     |
| ResNet-50 [75]|                | 0.820     | 0.823         | 0.770            | 0.804 | 0.805     | 0.811     | 0.820     | 0.817     | 0.706     | 0.735     |
| CSPT [62]    | Unsupervised     | 0.844     | 0.842         | 0.815            | 0.822 | 0.825     | 0.813     | 0.828     | 0.826     | 0.662     | 0.697     |
| CONTRIQUE [19]| Pretraining     | 0.851     | 0.849         | 0.808            | 0.817 | 0.832     | 0.822     | 0.821     | 0.820     | 0.668     | 0.702     |

TABLE III
Performance Comparison of CONVIQT Against Different NR Models on the CVD2014 and LIVE-Qualcomm Databases. In Each Column, the Three Best Models Are Boldfaced. Entries Marked '-' Denote That the Results Are Not Available.

| Method       | Model Type       | CVD2014 [68] | LIVE-Qualcomm [69] | SROCC t   | PLCC t   | SROCC t | PLCC t |
|--------------|------------------|--------------|---------------------|-----------|----------|---------|--------|
| BRSQUE [29]  |                  | 0.790        | 0.804               | 0.552     | 0.598    |         |        |
| VBIINDS [27] |                  | 0.795        | 0.806               | 0.570     | 0.626    |         |        |
| TLQVM [35]   |                  | 0.779        | 0.790               | 0.785     | 0.815    |         |        |
| VIDEVAL [36] |                  | 0.814        | 0.832               | 0.670     | 0.705    |         |        |
| RAPIQUE [37] |                  | 0.807        | 0.823               | 0.665     | 0.691    |         |        |
| VSPA [14]    |                  | 0.850        | 0.869               | 0.708     | 0.774    |         |        |
| GSTDQA [15]  |                  | 0.831        | 0.844               | 0.801     | 0.825    |         |        |
| Li et al. [16]|                | 0.862        | 0.882               | 0.833     | 0.837    |         |        |
| ResNet-50 [75]|                | 0.835        | 0.838               | 0.650     | 0.691    |         |        |
| CSPT [62]    |                  | 0.683        | 0.691               |           |          |         |        |
| CONTRIQUE [19]| Pretraining     | 0.826        | 0.836               | 0.765     | 0.777    |         |        |
| CONVIQT      |                  | 0.858        | 0.837               | 0.797     | 0.802    |         |        |

We also report the performance of CONVIQT against SOTA VQA models on multiple VQA datasets. To investigate the generalizability of the CONVIQT representations, we further performed cross-dataset evaluations. We also performed a collection of ablation experiments to analyze the significance of distortion types present in the pretraining data, and the impact of the various temporal transforms.

A. Experimental Settings

1) Pretraining Data: The pretraining data contains a mix of videos afflicted with synthetic and realistic artifacts.
   - Synthetic distortions: To learn synthetic artifacts, we first generated data as detailed in Sec. III-A. The high quality pristine contents required to generate the distorted sequences were obtained from 5 different sources as described below.
     1) Waterloo GRD database [76]: Contains 1000 videos of 1080p resolution captured at 24/30 fps and 10s in duration. We randomly sampled 500 source sequences from this database.
     2) Dareful [77]: The Dareful website contains open source high quality 4K stock video footages shot at 30 fps. We used 46 sequences from this platform.
     3) REDS [78]: The REDS dataset contains videos of 720p resolution captured at 120 fps. This dataset was originally created to benchmark video deblurring and super-resolution algorithms. We used a total of 270 (240 from training set and 30 from validation set) videos from this database.
     4) MCML [79]: MCML is a VQA dataset containing 240 distorted videos obtained from 10 reference sequences of 4K resolution captured at 30 fps.
     5) UVG [80]: Ultra Video Group (UVG) dataset contains 16 diverse 4K contents at 30/120 fps. To avoid overlap of contents between the VQA databases used for evaluation, we used only three contents of 4K resolution at 120 fps: ‘Bosphorus’, ‘Lips’ and ‘ShakeNDry’.

In this section we evaluate the performance of CONVIQT by carrying out a series of experiments. First we will describe the experiment settings used when pretraining CONVIQT, the evaluation protocol, and the VQA models used for comparison.

IV. EXPERIMENTS AND RESULTS

As $h$, $\lambda$ is a regularization parameter, $d$ is the number of dimensions of $h$, and $N$ is the number of videos present in the VQA database training set. Similar to training, multiscale processing is conducted during evaluation, and features are computed at two scales: full-scale and half-scale, with the final embedding being a concatenation of embeddings from both scales. During the evaluation phase, all of the features were calculated without performing any additional temporal transforms or cropping. Note that when the ridge regressor was trained, the weights of the recurrent unit (as well as those of CONTRIQUE) were frozen without any additional fine-tuning. Although fine-tuning could lead to better correlations, it modifies the encoder weights, and would not be a true measure of the effectiveness of the unsupervised learning process. However, we show in Sec. IV-B that even without fine-tuning, CONVIQT obtains competitive performance against SOTA VQA models.
Thus, 828 pristine sequences spanning multiple resolutions and frame rates were used to generate a total of 30855 synthetically distorted sequences.

- Authentic distortions: To capture realistic distortions we used videos from the training set of Kinetics-400 dataset [18]. Kinetics-400 is a human action video dataset containing 400 human action classes where each video is of duration around 10s. The training set of Kinetics-400 contains 240K videos of which we randomly sampled 30K videos, discarding all labels before training.

Thus, a total of 60K videos were used to pretrain CONVIQT. Additionally, we ensured that there was no content overlap between the pretraining sequences and the sequences present in VQA datasets used for evaluation.

2) Pretraining Details: We employed a Gated Recurrent Unit (GRU) [81] as the recurrent unit \( f(.) \) for modeling temporal information, and 2 layers of MLP as the projector \( g(.) \). The GRU contains a single layer with a hidden size of 1024. The output \( f(.) \) of the GRU is temporal average pooled (as shown in (2)) before being fed to the MLP \( g(.) \). The dimension of the final output \( z \) was fixed to \( K = 128 \). The batch size for pretraining was \( N = 1024 \), with equal numbers of videos randomly sampled from synthetic and authentically distorted sequences. The spatial features from the CONTRIQUE engine were extracted at the native resolution of the input videos, while temporal crops of length \( T = 16 \) frames were fed to the GRU to learn temporal information. During training, the videos were loaded in RGB format, and the same temporal transform was applied to all 3 channels of the input video. The transformed video was normalized to lie in the range \([0, 1] \) before being fed as input. The temperature parameter present in expressions (4) and (5) was fixed at \( \tau = 0.1 \). The GRU and the projector network were trained from scratch for 10 epochs using stochastic gradient descent (SGD) at a learning rate of 1.2 using a batch size \( N = 1024 \). A linear warmup over the first two epochs was applied to the learning rate, followed by a cosine decay schedule without restarts [82]. The values for hyperparameters: temperature parameter, hidden size and learning rate were chosen based on generally recommended values in the literature [19]. Batch size and length of temporal crops influence contrastive loss convergence with larger batch size and crop length yielding better performance. Based on the computational resources available we empirically observed the chosen values to obtain the best performance. All the experiments were done in Python using the PyTorch\(^1\) deep learning modules.

3) Evaluation Datasets: We evaluated CONVIQT on 12 large VQA datasets spanning both synthetic and realistic distortions. The characteristics of the VQA datasets used for evaluation are summarized in Table I. When using the LSVQ [9] database, which also has human-labeled space-time video patches, we only used the full videos (and corresponding human opinion scores), leaving out the video patches.

4) Compared Methods: We compared the performance of CONVIQT against a variety of SOTA NR-VQA models. The compared models can be classified into three categories based on the feature design methodology: (a) Traditional/hand-crafted features - BRISQUE [29], VBLINDS [27], TLVQM [35], VIDEVAL [36], and RAPIQUE [37]. These models employ support vector regressors (SVRs) to map features to quality scores. (b) Supervised deep learning based models - VSFA [14], PVQ [9], GSTVQA [15], and Li et al. [16]. Note that RAPIQUE is a hybrid model employing a combination of NSS and deep CNN features. (c) Self-supervised deep learning based models - CSPT [62] and CONTRIQUE [19]. To numerically compare the above models, we copied the values reported by the respective authors or in the literature. For PVQ, which is capable of inferencing on video patches, we only report the performance of the model trained on videos, since video patch quality scores are not employed in our evaluation. To compare the results obtained by the supervised and unsupervised pretraining techniques, we also included a 2D Resnet-50 [75] model pretrained on Imagenet [17]. Since BRISQUE, Resnet-50 and CONTRIQUE are IQA models, the respective features were computed on each video frame then average pooled along the temporal dimension to obtain the final video embedding. These embeddings were then mapped to quality scores using an SVR on the BRISQUE features, and a linear regressor on the Resnet-50 and CONTRIQUE features. For fair comparison of CSPT with CONVIQT, we report the CSPT performance under linear evaluation.

5) Evaluation Protocol: Spearman’s rank order correlation coefficient (SROCC) and Pearson’s linear correlation coefficient (PLCC) were used as evaluation metrics to compare the VQA models. A four-parameter logistic non-linearity [83] was applied to the quality predictions before PLCC was computed.

---

\(^1\)https://pytorch.org/

| Method       | Model Type       | LIVE-VQA [2] | CSIQ-VQA [3] | ETRI-LIVE STSVQ [11] | AVT-VQDB-UHD-1 [10] |
|--------------|------------------|--------------|--------------|----------------------|---------------------|
|              |                  | SROCC †     | FLCC †      | SROCC †              | PLCC †              |
| BRISQUE [29] | Traditional/     | 0.367        | 0.371        | 0.654                | 0.672               |
| VBLINDS [27] | Handcrafted      | 0.724        | 0.748        | 0.794                | 0.771               |
| TLVQM [35]   | Features         | 0.733        | 0.751        | 0.794                | 0.774               |
| VIDEVAL [36] | Supervised       | 0.475        | 0.463        | 0.368                | 0.300               |
| VSF [14]     | Pretraining      | 0.720        | 0.727        | 0.798                | 0.781               |
| Resnet-50 [75]|                 | 0.442        | 0.385        | 0.703                | 0.686               |
| CSPT [62]    | Unsupervised     | 0.713        | 0.726        | 0.751                | 0.749               |
| CONTRIQUE [19]| Pretraining      | 0.636        | 0.623        | 0.795                | 0.775               |
| CONVIQT      |                 | 0.622        | 0.595        | 0.766                | 0.749               |

**TABLE IV**

**Performance Comparison of CONVIQT Against Different NR Models on VQA Databases Containing Synthetic Distortions**. In Each Column, the Three Best Models Are Boldfaced. Entries Marked ’†’ Denote That the Results Are Not Available.
During evaluation, the videos were processed at their native resolution, and were partitioned into non-overlapping clips of \( T \) continuous frames. From each clip CONVIQT produced a single feature vector as output, and the features from all clips were average pooled to obtain the final video representations. These representations, along with the corresponding human opinion scores were used to learn the regressor weights. When learning regressor weights, each VQA dataset was randomly split into 70%, 10%, and 20% partitions corresponding to training, validation, and testing sets, respectively. The validation set was used to calculate the regularization coefficient of the regressor via grid search. For VQA datasets containing multiple distorted videos belonging to the same content, the random splits were conducted in a way that ensured no content overlap between the training and testing sets. The training protocol was repeated 100 times to avoid biases towards the choice of videos present in the training set, and the median performance was reported. Due to the large size of the LSVQ dataset, the single split reported by the authors [9] was used.

### B. Correlation Against Human Judgments

We compared the performance of CONVIQT against SOTA VQA models on authentically distorted videos, as shown in Table II. It may be observed from the Table that CONVIQT achieved competitive correlation values compared to the other VQA models. The compared VQA models are categorized based on the type of feature extraction techniques employed. The performance differences between CONTRIQUE and CONVIQT is indicative of the significance of the recurrent unit. This difference depends on the VQA dataset being used, but in most cases employing the recurrent unit yielded better performance. Notably, the performances of CONTRIQUE and CONVIQT were achieved without any supervised fine-tuning, unlike the other models.

Table III shows performance comparisons on the CVD2014 [68] and LIVE-Qualcomm [69] datasets. These datasets contain multiple videos of the same scene impaired by authentic distortions. Thus, the train/test splits were done ensuring no content overlap between them. Again, CONVIQT obtained excellent correlations when compared to other VQA models.

The performance of CONVIQT on videos modified by synthetic distortions is analyzed in Table IV. CONVIQT obtains superior correlations on recent VQA datasets such as ETRI-LIVE STSVQ [11] and AVT-VQDB-UHD-1 [10], while its performance is slightly lower on legacy datasets like LIVE-VQA [2] and CSIQ-VQA [3]. This may be attributed to the type of distortions present. ETRI-LIVE STSVQ and AVT-VQDB-UHD-1 mainly contain scaling and compression artifacts which were well modeled in the data used to pretrain CONVIQT. The legacy datasets contain distortions such as transmission errors arising in wireless and IP networks, and packet loss errors, which were not included in the training data and could be a contributing factor in obtaining lower correlation values.

### C. Performance Comparison on Variable Frame Rate Videos

Table V compares CONVIQT against other VQA models on variable frame rate (VFR) videos. The LIVE-YT-HFR [13] and BVI-HFR [12] VQA databases contain videos of same content but sampled at multiple combinations of frame rates and compression factors, along with corresponding human opinion scores. VFR-VQA is a challenging problem, since quality predictors must account for subtle perceptual quality changes occurring along the temporal dimension due to frame rate changes. As comparisons we also included FAVER [73], which is a current SOTA no-reference VFR-VQA model. From Table V, it may be observed that CONVIQT achieved competitive performance when compared against other NR-VQA models, indicating the frame rate discrimination capability of CONVIQT representations. We may also observe significant performance gaps between CONTRIQUE and CONVIQT on same datasets, highlighting the importance of the recurrent unit.

### D. Cross Dataset Evaluation

We conducted a cross dataset analysis in Table VII. The regressors were trained and tested on different VQA datasets. Cross data evaluations are beneficial when analyzing the generalizability of compared models. We compared CONTRIQUE and CONVIQT to study whether the inclusion of the recurrent unit yields better generalization. We chose a subset of VQA datasets from Table I, two containing synthetic distortions, two containing UGC videos and one containing VFR videos. From the Table, it may be observed that in most cases, CONVIQT delivered much better performance than CONTRIQUE, highlighting the importance of the recurrent unit in obtaining more generalized representations. Note that even for cross dataset evaluation, only the regressor weights were modified based on the training data, while the weights of the CONVIQT model were kept intact.

### E. Significance of Training Data

We noted in Sec. III that the pretraining data for CONVIQT contains a combination of videos afflicted by synthetic and authentic distortions. In this experiment we investigate the generalization of the learned representation when CONVIQT is pretrained with either only synthetic or only authentic artifacts. The results are reported in Table VI following similar
TABLE VI
SROCC PERFORMANCE COMPARISON TO ANALYZE DIFFERENT CONVIQT TRAINING PROTOCOLS. syn AND UGC INDICATE MODELS TRAINED ON DATA CONTAINING ONLY SYNTHETIC AND AUTHENTIC DISTORTIONS, RESPECTIVELY. IN EACH ROW THE TOP PERFORMING MODEL IS BOLDFACED.

| Training       | Testing       | CONVIQT-syn | CONVIQT-UGC | CONVIQT |
|----------------|---------------|-------------|-------------|---------|
| YouTube-UGC    | KoNVID        | 0.728       | 0.750       | 0.761   |
| KoNVID         | YouTube-UGC   | 0.580       | 0.601       | 0.731   |
| ETRI-LIVE STSVQ| AVT-VQDB UHD-1| 0.759       | 0.759       | 0.774   |
| ETRI-LIVE STSVQ| ETRI-LIVE STSVQ| 0.697       | 0.607       | 0.651   |
| LIVE-YT-HFR    | AVT-VQDB UHD-1| 0.754       | 0.730       | 0.767   |
| AVT-VQDB UHD-1 | LIVE-YT-HFR   | 0.593       | 0.448       | 0.556   |

Frames Resnet-50 (ImageNet pretrained) CONVIQT

(a) Frame sampled from the 4744073127.mp4 sequence in the KoNVID [7] dataset.

(b) Frame sampled from the tr16_23fps sequence in the LIVE-VQA [2] dataset.

Fig. 4. Visualization of activation maps from the (a) KoNVID [7] and (b) LIVE-VQA [2] datasets. The attention maps were obtained by summing feature channels in the conv4 layer of the learned Resnet-50 architecture [84].

TABLE VII
CROSS DATABASE SROCC COMPARISON OF VQA MODELS. IN EACH ROW THE TOP PERFORMING MODEL IS BOLDFACED.

| Training       | Testing       | CONTRIQUB | CONVIQT |
|----------------|---------------|-----------|---------|
| YouTube-UGC    | KoNVID        | 0.718     | 0.761   |
| KoNVID         | YouTube-UGC   | 0.498     | 0.731   |
| ETRI-LIVE STSVQ| AVT-VQDB UHD-1| 0.785     | 0.774   |
| AVT-VQDB UHD-1 | ETRI-LIVE STSVQ| 0.550     | 0.651   |
| LIVE-YT-HFR    | AVT-VQDB UHD-1| 0.740     | 0.767   |
| AVT-VQDB UHD-1 | LIVE-YT-HFR   | 0.378     | 0.556   |

F. Importance of Temporal Transforms

To obtain more robust representations, the CONVIQT pipeline contains a temporal transform module as illustrated in Fig. 1. In this experiment we analyzed the importance of the temporal transform by removing the module during CONVIQT training. The resulting model performance is compared in Table VIII where it may be observed that the absence of a temporal transform leads to less generalizable representations.

G. Visualization of Learned Representations

In Fig. 4 we visualize the activation maps of selected frames from the KoNVID and LIVE-VQA datasets. For evaluation purposes we compared two models: Resnet-50 (pretrained on ImageNet) and CONVIQT. Attention maps were generated...


by summarizing feature channels in the conv4 layer of the Resnet-50 architecture as discussed in [84]. In attention maps, the brightness of each region represents the importance the model assigns to it during inference. Thus, in Fig. 4 the bright red regions in the attention maps denote regions of high importance during inference. For each frame we focus on two regions (red and blue boxes) containing significant distortions, and enlarge these regions for better visualization and analysis. The attention maps of CONVIQT in both Fig.4-a and Fig. 4-b clearly represent higher sensitivity to severely distorted regions relative to that of Resnet-50. The Resnet-50 was pretrained for classification task where learned features are used to discriminate different categories based on semantic information such as structures. By contrast, CONVIQT focuses on regions that could affect perceptual quality, thus yielding relevant representations pertaining to the VQA task.

### H. Qualitative Analysis

In this subsection we qualitatively compare successful and failure examples of CONVIQT sampled from different VQA datasets in Fig. 5 and Fig. 6. Fig. 5 contains cases with authentic distortions sampled from the KoNViD [7] dataset. Top-row: Successful cases sampled from the 7136906983.mp4 and 5668795950.mp4 sequences. Bottom-row: Failure examples sampled from the 5621374631.mp4 and 9524637688.mp4 sequences. GT refers to ground truth quality scores.

![Fig. 5. Illustration of successful and failure examples sampled from the KoNViD [7] dataset. Top-row: Successful cases sampled from the 7136906983.mp4 and 5668795950.mp4 sequences. Bottom-row: Failure examples sampled from the 5621374631.mp4 and 9524637688.mp4 sequences. GT refers to ground truth quality scores.](image)

![Fig. 6. Illustration of successful and failure examples sampled from the LIVE-VQA [2] dataset. Top-row: Successful cases sampled from the pr14_25fps and rh13_25fps sequences. Bottom-row: Failure examples sampled from the rb6_25fps and mc2_50fps sequences. GT refers to ground truth quality scores.](image)



**TABLE VIII**

| Training       | Testing       | w/ Temporal Transform | w/ Temporal Transform |
|----------------|---------------|-----------------------|-----------------------|
| YouTube-UGC    | KoNViD        | 0.739                 | 0.761                 |
| KoNViD         | YouTube-UGC   | 0.722                 | 0.731                 |
| ETRI-LIVE STSVQ| AVT-VQDB UHD-1| 0.759                 | 0.774                 |
| AVT-VQDB UHD-1 | ETRI-LIVE STSVQ| 0.722                 | 0.651                 |
| LIVE-YT-HPR    | AVT-VQDB UHD-1| 0.716                 | 0.767                 |
| AVT-VQDB UHD-1 | LIVE-YT-HPR   | **0.599**             | 0.556                 |

Fig. 5. Illustration of successful and failure examples sampled from the KoNViD [7] dataset. Top-row: Successful cases sampled from the 7136906983.mp4 and 5668795950.mp4 sequences. Bottom-row: Failure examples sampled from the 5621374631.mp4 and 9524637688.mp4 sequences. GT refers to ground truth quality scores.

**TABLE IX**

| Distortion Type | SROCC† | PLCC† |
|-----------------|--------|-------|
| Wireless        | 0.595  | 0.629 |
| IP              | 0.486  | 0.510 |
| H264            | 0.738  | 0.719 |
| HIP             | 0.810  | 0.800 |
| Overall         | **0.622** | **0.595** |

(a) Predicted = 45.1, GT = 46.8  
(b) Predicted = 47.2, GT = 48
(c) Predicted = 55, GT = 68  
(d) Predicted = 59.76, GT = 78.34

![Fig. 6. Illustration of successful and failure examples sampled from the LIVE-VQA [2] dataset. Top-row: Successful cases sampled from the pr14_25fps and rh13_25fps sequences. Bottom-row: Failure examples sampled from the rb6_25fps and mc2_50fps sequences. GT refers to ground truth quality scores.](image)

**TABLE X**

| Distortion Type | SROCC† | PLCC† |
|-----------------|--------|-------|
| H264            | 0.817  | 0.819 |
| Packet Loss     | 0.533  | 0.417 |
| MPEG            | 0.800  | 0.806 |
| Wavelet         | 0.867  | 0.867 |
| White Noise     | 0.800  | 0.796 |
| HEVC            | 0.717  | 0.735 |
| Overall         | **0.766** | **0.749** |

I. Limitations

From Table IV it may be observed that CONVIQT underperformed on the legacy datasets LIVE-VQA and CSIQ-VQA. In this subsection we investigate this behavior by analyzing the performance of CONVIQT on each individual distortion types present in these datasets.

These artifacts were not included in the pretraining data since they are less prevalent in current video sharing and streaming scenarios. Table IX and X contain the computed correlations of CONVIQT for each type of distortion present in LIVE-VQA and CSIQ-VQA respectively. It maybe observed from the Tables that CONVIQT obtained lower correlations on the wireless, IP and packet loss distortions than on the others. The lack of training examples corresponding to these distortion types could be a factor contributing to this lower performance.

**TABLE IX**

| Distortion Type | SROCC† | PLCC† |
|-----------------|--------|-------|
| Wireless        | 0.595  | 0.629 |
| IP              | 0.486  | 0.510 |
| H264            | 0.738  | 0.719 |
| MPEG            | 0.810  | 0.800 |
| Overall         | **0.622** | **0.595** |

(a) Predicted = 45.1, GT = 46.8  
(b) Predicted = 47.2, GT = 48
(c) Predicted = 55, GT = 68  
(d) Predicted = 59.76, GT = 78.34

![Fig. 6. Illustration of successful and failure examples sampled from the LIVE-VQA [2] dataset. Top-row: Successful cases sampled from the pr14_25fps and rh13_25fps sequences. Bottom-row: Failure examples sampled from the rb6_25fps and mc2_50fps sequences. GT refers to ground truth quality scores.](image)

**TABLE X**

| Distortion Type | SROCC† | PLCC† |
|-----------------|--------|-------|
| H264            | 0.817  | 0.819 |
| Packet Loss     | 0.533  | 0.417 |
| MPEG            | 0.800  | 0.806 |
| Wavelet         | 0.867  | 0.867 |
| White Noise     | 0.800  | 0.796 |
| HEVC            | 0.717  | 0.735 |
| Overall         | **0.766** | **0.749** |

I. Limitations

From Table IV it may be observed that CONVIQT underperformed on the legacy datasets LIVE-VQA and CSIQ-VQA. In this subsection we investigate this behavior by analyzing the performance of CONVIQT on each individual distortion types present in these datasets.

These artifacts were not included in the pretraining data since they are less prevalent in current video sharing and streaming scenarios. Table IX and X contain the computed correlations of CONVIQT for each type of distortion present in LIVE-VQA and CSIQ-VQA respectively. It maybe observed from the Tables that CONVIQT obtained lower correlations on the wireless, IP and packet loss distortions than on the others. The lack of training examples corresponding to these distortion types could be a factor contributing to this lower performance.
In the design of CONVIQT we employed a frozen CONTRIQUE model to conduct spatial feature extraction. This restricts the spatial artifacts to be similar to the ones used during the training of CONTRIQUE. This could also limit the performance of CONVIQT when applied on spatial distortions unique to videos and not observed on images.

V. CONCLUSION AND FUTURE WORK

We proposed a self-supervised training procedure to obtain perceptually relevant video quality representations. The model was trained with the goal of discriminating different distortion types and degradation degrees using a large unlabeled video database, containing videos afflicted by both synthetic and realistic distortions. We performed comprehensive evaluations of our proposed design, and found that the performance of CONVIQT is comparable to many supervised deep VQA models. Notably, CONVIQT achieves this performance even without fine-tuning, underscoring the effectiveness of the auxiliary task. We conducted ablation experiments to understand the significance of the temporal transformations, and inferred that these transformations lead to better generalization performance. We also studied the impact of the distortion types present in training data, and deduced that a training database containing a combination of synthetic and authentic distortions yielded more robust features. A software release of CONVIQT has been released online\(^2\) to promote reproducible research.

Though CONVIQT achieves competitive performance on VFR videos, there is significant room for improvement, viz., the representations can be made more sensitive to reflect perceptual quality changes occurring due to frame rate variations. As part of future work, we would like to explore auxiliary tasks which promote feature learning with high sensitivity to frame rate artifacts.

ACKNOWLEDGMENT

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

REFERENCES

[1] CIVNI. (2019). Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper. San Jose, CA, USA. Accessed: Sep. 1, 2021 [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[2] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” IEEE Trans. Image Process., vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[3] P. V. Vu and D. M. Chandler, “VI$^3$: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices,” J. Electron. Imag., vol. 23, no. 1, Feb. 2014, Art. no. 013016.

[4] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, “A H.264/AVC video database for the evaluation of quality metrics,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Mar. 2010, pp. 2430–2433.

[5] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-J. Kuo, “MCL-V: A streaming video quality assessment database,” J. Vis. Commun. Image Represent., vol. 30, pp. 1–9, Jul. 2015.

[6] Z. Sinno and A. C. Bovik, “Large-scale study of perceptual video quality,” IEEE Trans. Image Process., vol. 28, no. 2, pp. 612–627, Feb. 2019.

[7] V. Hosu et al., “The Konstanz natural video database (KoNViD-1k),” in Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX), May 2017, pp. 1–6.

[8] Y. Wang, S. Inguva, and B. Adsumilli, “YouTube UGC dataset for video compression research,” in Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP), Sep. 2019, pp. 1–5.

[9] Z. Ying, M. Mandal, D. Ghiadhim, and A. Bovik, “Patch-VQ: ‘Patch up’ the video quality problem,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14014–14024.

[10] R. R. Ramachandra Rao, S. Göring, W. Robitzka, B. Feiten, and A. Raake, “AVT-VQDB-UHD-1: A large scale video quality database for UHD-1,” in Proc. IEEE Int. Symp. Multimedia (ISM), Dec. 2019, pp. 17–17.

[11] D. Y. Lee et al., “A subjective and objective study of space-time subsampled video quality,” IEEE Trans. Image Process., vol. 31, pp. 934–948, 2022.

[12] A. Mackin, F. Zhang, and D. R. Bull, “A study of high frame rate video formats,” IEEE Trans. Multimedia, vol. 21, no. 6, pp. 1499–1512, Jun. 2019.

[13] P. C. Madhusudana, X. Yu, N. Birbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “Subjective and objective quality assessment of high frame rate videos,” IEEE Access, vol. 9, pp. 108069–108082, 2021.

[14] D. Li, T. Jiang, and M. Jiang, “Quality assessment of in-the-wild videos,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 2351–2359.

[15] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, “Learning generalized spatial–temporal deep feature representation for no-reference video quality assessment,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 4, pp. 1903–1916, Apr. 2022.

[16] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, “Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 9, pp. 5944–5958, Sep. 2022.

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

[18] W. Kay et al., “The kinetics human action video dataset,” 2017, arXiv:1705.06950.

[19] P. C. Madhusudana, N. Birbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “Image quality assessment using contrastive learning,” IEEE Trans. Image Process., vol. 31, pp. 4149–4161, 2022.

[20] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, Toward a Practical Perceptual Video Quality Metric. Accessed: Nov. 2019. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video-quality.html

[21] 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. 2019.

[22] P. C. Madhusudana, N. Birbeck, 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.

[23] P. C. Madhusudana, N. Birbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “Making video quality assessment models sensitive to frame rate distortions,” IEEE Signal Process. Lett., vol. 29, pp. 897–901, 2022.

[24] K. Seshadrinathan and A. C. Bovik, “A structural similarity metric for video based on motion models,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Apr. 2007, p. 869.

[25] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birbeck, B. Adsumilli, and A. C. Bovik, “A comparative evaluation of temporal pooling methods for blind video quality assessment,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2020, pp. 141–145.

[26] A. K. Moorthy and A. C. Bovik, “Statistics of natural image distortions,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Mar. 2010, pp. 962–965.

[27] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind prediction of natural video quality,” IEEE Trans. Image Process., vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[28] X. Li, Q. Guo, and X. Lu, “Spatiotemporal statistics for video quality assessment,” IEEE Trans. Image Process., vol. 25, no. 7, pp. 3329–3342, Jul. 2016.

\(^2\)https://github.com/pavancm/CONVIQT

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
[29] 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.

[30] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘completely blind’ image quality analyzer,” IEEE Sig. Process. Lett., vol. 20, no. 3, pp. 209–212, Mar. 2013.

[31] A. Mittal, M. A. Saad, and A. C. Bovik, “A completely blind video integrity Oracle,” IEEE Trans. Image Process., vol. 25, no. 1, pp. 289–300, Jan. 2016.

[32] S. V. R. Dendi and S. S. Channappayya, “No-reference video quality assessment using natural spatiotemporal scene statistics,” IEEE Trans. Image Process., vol. 29, pp. 5612–5624, 2020.

[33] J. Xu, P. Ye, Y. Liu, and D. Doermann, “No-reference video quality assessment via feature learning,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2014, pp. 491–495.

[34] K. Seshadrinathan and A. C. Bovik, “Temporal hysteresis model of time varying subjective video quality,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2011, pp. 1153–1156.

[35] J. Korhonen, “Two-level approach for no-reference consumer video quality assessment,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 5923–5938, Dec. 2019.

[36] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, “UGC-VQA: Benchmarking blind video quality assessment for user generated content,” IEEE Trans. Image Process., vol. 30, pp. 4449–4464, 2021.

[37] 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.

[38] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, “Blind video quality assessment with weakly supervised learning and resampling strategy,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 8, pp. 2244–2255, Aug. 2019.

[39] W. Liu, Z. Duanmu, and Z. Wang, “End-to-end blind quality assessment of compressed videos using deep neural networks,” in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 546–554.

[40] H. Wu et al., “Fast-VQA: Efficient end-to-end video quality assessment with fragment sampling,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2022, pp. 538–554.

[41] D. Li, T. Jiang, and M. Jiang, “Unified quality assessment of in-the-Wild videos with mixed datasets training,” Int. J. Comput. Vis., vol. 129, no. 4, pp. 1238–1257, Apr. 2021.

[42] W. Zhang, K. Ma, G. Zhai, and X. Yang, “Uncertainty-aware blind image quality assessment in the laboratory and wild,” IEEE Trans. Image Process., vol. 30, pp. 3474–3486, 2021.

[43] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, “Continual learning for blind image quality assessment,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 2864–2878, Mar. 2023.

[44] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in Proc. Int. Conf. Learn. Mach. Intell., vol. 45, no. 3, pp. 2864–2878, Mar. 2023.

[45] G. Larsson, M. Mair, and G. Shakhnarovich, “Colorization as a proxy for best basis selection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 667–676.

[46] T. Han, W. Xie, and A. Zisserman, “Video representation learning by dense predictive coding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 1483–1492.

[47] Y. Yao, C. Liu, D. Luo, Y. Zhou, and Q. Ye, “Video playback rate perception for self-supervised spatio-temporal representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6547–6556.

[48] L. Tao, X. Wang, and T. Yamashita, “Self-supervised video representation learning using inter-intra contrastive framework,” in Proc. 28th ACM Int. Conf. Multimedia, Oct. 2020, pp. 2193–2202.

[49] R. Qian et al., “Spatiotemporal contrastive video representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 6960–6970.

[50] T. Pan, Y. Song, T. Yang, W. Jiang, and W. Liu, “VidemoCo: Contrastive video representation learning with temporally adversarial examples,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 11200–11209.

[51] C. Feichtenhofer, H. Fan, B. Xiong, R. Girshick, and K. He, “A large-scale study on unsupervised spatiotemporal representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 3298–3308.

[52] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, “Contrastive self-supervised pre-training for video quality assessment,” IEEE Trans. Image Process., vol. 31, pp. 458–471, 2022.

[53] P. Khosla et al., “Supervised contrastive learning,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 18661–18673.

[54] FFmpeg. Encoding for Streaming Sites. Accessed: Oct. 1, 2021. [Online]. Available: https://ffmpeg.org/wiki

[55] H. Knoche, J. D. McCarthy, and M. A. Sasse, “Can small be beautiful? Assessing image resolution requirements for mobile TV,” in Proc. 13th Annu. ACM Int. Conf. Multimedia, Nov. 2005, pp. 829–838.

[56] G. Cermak, M. Pinson, and S. Wolf, “The relationship among video quality, screen resolution, and bit rate,” IEEE Trans. Broadcast., vol. 57, no. 2, pp. 258–262, Jun. 2011.

[57] G. Georgis, G. Lentaris, and D. Reisis, “Reduced complexity superresolution for low-bitrate video compression,” IEEE Trans. Circuits Syst. Video Technol., vol. 26, no. 7, pp. 1332–1345, Feb. 2016.

[58] M. Nutuinen, T. Virtanen, M. Vahteranoksa, T. Vuori, P. Ottinnen, and J. Hakkinen, “CVD2014—A database for evaluating no-reference video quality assessment algorithms,” IEEE Trans. Image Process., vol. 25, no. 7, pp. 3073–3086, 2016.

[59] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, “In-capture mobile video distortions: A study of subjective behavior and objective algorithms,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 9, pp. 2061–2077, Sep. 2018.

[60] 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.

[61] 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. 5128–5137.

[62] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, “High frame rate video quality assessment using VMAF and entropic differences,” in Proc. Picture Coding Syst. Work (PCS), Jan. 2021, pp. 1–5.

[63] Q. Zheng, Z. Tu, P. C. Madhusudana, X. Zeng, A. C. Bovik, and Y. Fan, “FAVER: Blind quality prediction of variable frame rate videos,” arXiv:2201.01492.

[64] R. R. Coifman and M. V. Wickerhouser, “Entropy-based algorithms for best basis selection,” IEEE Trans. Inf. Theory, vol. 38, no. 2, pp. 713–718, Mar. 1992.

[65] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
A. Mercat, M. Viitanen, and J. Vanne, “UVG dataset: 50/120 fps 4 K resolution: Dataset and study,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1996–2005.

M. Cheon and J.-S. Lee, “Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 7, pp. 1467–1480, Jul. 2018.

A. Mercat, M. Viitanen, and J. Vanne, “UVG dataset: 50/120 fps 4 K sequences for video codec analysis and development,” in *Proc. 11th ACM Multimedia Syst. Conf.*, May 2020, pp. 297–302.

K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” 2014, arXiv:1409.1259.

L. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–16.

Final Report From the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment, *Video Qual. Experts Group (VQEG)*, USA, 2000.

N. Komodakis and S. Zagoruyko, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.

Dareful. *Stock Video Footage in 4 K*. Accessed: Sep. 1, 2021. [Online]. Available: https://dareful.com

Z. Duanmu, W. Liu, Z. Li, K. Ma, and Z. Wang, “Characterizing generalized rate-distortion performance of video coding: An Eigen analysis approach.” *IEEE Trans. Image Process.*, vol. 29, pp. 6180–6193, 2020.

S. Nah et al., “NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1996–2005.

Yilin Wang 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 interests 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, working on watermark-based error resilience in video communications. He currently manages and leads the Media Algorithms Group, YouTube/Google. From 2005 to 2011, he was a Senior Research Scientist with Citrix Online and from 2011 to 2016, he was a Senior Manager Advanced Technology with GoPro, at both places developing algorithms for images/video quality enhancement, compression, capture, and streaming. He is an active member of IEEE (and MMSP TC), ACM, SPIE, and VES. He has coauthored more than 120 papers and patents. His research interests include image/video processing, machine vision, video compression, spherical capture, VR/AR, visual effects, and related areas.

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 with Siemens Corporate Research working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Software Engineer with the Media Algorithms Team, Youtube/Google. His research interests include perceptual video processing, video coding, and video quality assessment.

Alan C. Bovik (Life Fellow, IEEE) is currently the Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. He is 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. His books include *The Essential Guides to Image and Video Processing*. He co-founded and was the longest-serving Editor-in-Chief for the IEEE TRANSACTIONS ON IMAGE PROCESSING and also created/chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, in 1994.

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, the M.Tech. (Research) degree in electrical and communication engineering from the Indian Institute of Science (IISc), Bengaluru, India, in 2018, and the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin, USA, in 2022. He is currently a Research Engineer with Siemens Corporate Research working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Senior Research Scientist with Citrix Online and from 2011 to 2016, he was a Senior Manager Advanced Technology with GoPro, at both places developing algorithms for images/video quality enhancement, compression, capture, and streaming. He is an active member of IEEE (and MMSP TC), ACM, SPIE, and VES. He has coauthored more than 120 papers and patents. His research interests include image/video processing, machine vision, video compression, spherical capture, VR/AR, visual effects, and related areas.

Yilin Wang 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 interests 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, working on watermark-based error resilience in video communications. He currently manages and leads the Media Algorithms Group, YouTube/Google. From 2005 to 2011, he was a Senior Research Scientist with Citrix Online and from 2011 to 2016, he was a Senior Manager Advanced Technology with GoPro, at both places developing algorithms for images/video quality enhancement, compression, capture, and streaming. He is an active member of IEEE (and MMSP TC), ACM, SPIE, and VES. He has coauthored more than 120 papers and patents. His research interests include image/video processing, machine vision, video compression, spherical capture, VR/AR, visual effects, and related areas.

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 with Siemens Corporate Research working on automatic detection and segmentation of anatomical structures in full body medical images. He is currently a Software Engineer with the Media Algorithms Team, Youtube/Google. His research interests include perceptual video processing, video coding, and video quality assessment.

Alan C. Bovik (Life Fellow, IEEE) is currently the Cockrell Family Regents Endowed Chair Professor with The University of Texas at Austin. He is 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. His books include *The Essential Guides to Image and Video Processing*. He co-founded and was the longest-serving Editor-in-Chief for the IEEE TRANSACTIONS ON IMAGE PROCESSING and also created/chaired the IEEE International Conference on Image Processing which was first held in Austin, TX, USA, in 1994.