

ChipQA: No-Reference Video Quality Prediction via Space-Time Chips

Joshua Peter Ebenezer¹, Zaixi Shang¹, *Student Member, IEEE*, Yongjun Wu, *Senior Member, IEEE*, Hai Wei, *Member, IEEE*, Sriram Sethuraman, and Alan C. Bovik, *Fellow, IEEE*

Abstract—We propose a new model for no-reference video quality assessment (VQA). Our approach uses a new idea of highly-localized space-time (ST) slices called Space-Time Chips (ST Chips). ST Chips are localized cuts of video data along directions that *implicitly* capture motion. We use perceptually-motivated bandpass and normalization models to first process the video data, and then select oriented ST Chips based on how closely they fit parametric models of natural video statistics. We show that the parameters that describe these statistics can be used to reliably predict the quality of videos, without the need for a reference video. The proposed method implicitly models ST video naturalness, and deviations from naturalness. We train and test our model on several large VQA databases, and show that our model achieves state-of-the-art performance at reduced cost, without requiring motion computation.

Index Terms—Video quality assessment, natural video statistics, human visual system.

I. INTRODUCTION

VIDEO content continues to proliferate, already accounting for more than 70% of internet traffic, and projected to exceed 82% of internet traffic by 2021. Live internet video will account for 13 percent of Internet video traffic by 2021, and is predicted to grow 15-fold from 2016 to 2021 [1]. Distortions can affect videos as they are captured, transmitted, and received. The task of assessing the quality of a video in the presence of distortions is thus an increasingly important open problem. In most instances in this process there is no reference against which to measure their eventual perceived quality. Nevertheless, it is of vital importance to providers of video content to be able to monitor and predict the perceptual quality of their videos, since this directly impacts customer satisfaction. Video quality tools can also help make well-informed design choices while creating systems for capturing, processing, transmitting, and displaying videos. Video quality assessment algorithms also have applications in video denoising, designing loss functions for deep learning, video compression, and many other high-impact areas.

Manuscript received December 14, 2020; revised July 5, 2021; accepted August 18, 2021. Date of publication September 17, 2021; date of current version September 24, 2021. This work was supported by Amazon Prime Video. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marta Mrak. (Joshua Peter Ebenezer and Zaixi Shang contributed equally to this work.) (Corresponding author: Joshua Peter Ebenezer.)

Joshua Peter Ebenezer, Zaixi Shang, and Alan C. Bovik are with the Laboratory for Image and Video Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: joshuaebenezer@utexas.edu).

Yongjun Wu, Hai Wei, and Sriram Sethuraman are with Amazon Prime Video, Seattle, WA 98109 USA.

Digital Object Identifier 10.1109/TIP.2021.3112055

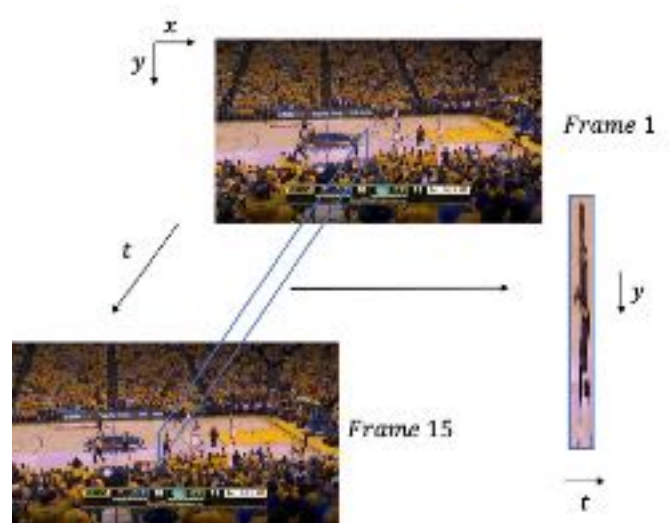


Fig. 1. Space-Time Chips capture elements in motion. On the left are frames 1 and 15 of a video of a basketball game. The ST-chip is marked in blue. The player near the edges of the chip on the xy plane is moving to his right as time progresses. The chip is a localized cut of all the frames between 1 and 15, perpendicular to the direction of his motion, which captures the player, as shown on the right.

Collecting a large number of human opinion scores on the quality of a video is the most reliable way to measure its quality. However, collecting subjective opinions of video quality is a cumbersome and expensive task. It is also time-consuming and cannot be deployed prior to or during transmission of a video, when they are being live-streamed or have other latency constraints. Subjective opinions are nevertheless useful as a gold standard when designing objective video quality assessment (VQA) algorithms. Objective VQA algorithms are designed to correlate well with these subjective human judgments, and deployed effectively and cheaply in video processing systems. VQA algorithms are typically evaluated on the basis of data gathered from studies on human judgments of video quality. Subjective judgments of video quality are first collected from a statistically significant number of human observers and normalized with respect to each observer's scores to form an opinion score for each observer and for each video. These opinion scores are then averaged across the observers yielding single mean opinion score (MOS) for each video. These mean opinion scores are the ground truth against which objective VQA algorithms are trained and tested.

Objective VQA algorithms fall into three categories: full-reference (FR), reduced-reference (RR), and no-reference

(NR). FR VQA algorithms require a reference video against which the distorted video is compared. RR VQA algorithms require only some information from the reference video, but not all, to predict the quality of a distorted video. NR VQA algorithms do not make use of a reference video, and the models we present here fall in this category. NR VQA algorithms rely on distortion-specific features or models of natural video statistics to predict video quality, and are of great interest because of their potentially broader applications.

In this work, we propose a NR VQA algorithm based on the natural video statistics of space-time (ST) chips. ST Chips are a new feature space that are defined as localized and oriented cuts of a video volume, and an illustration of the concept is shown in Fig. 1. We show that when a pristine video is processed using models derived from the human visual system, ST Chips extracted from the processed video that are along the direction of motion follow certain regular statistics, which is a breakthrough in our understanding of natural video statistics. We first proposed the idea of using ST Chips in [2], where we extracted ST Chips using optical flow in a prototype algorithm. In this work, we develop that idea further, introducing temporal processing of the video data based on models of the human visual system, and doing away with optical flow by using a simpler and more elegant approach based on regularities revealed by analysis of video statistics. Directions of motions are found in an implicit manner by using well-known models of natural image statistics and the smoothness of motion fields. The statistics of ST Chips extracted along these directions of motion can be modelled with parametrized distributions, and we show that these parameters can be used to reliably predict the quality of videos. We call our model ChipQA, which we designed to be able to handle different kinds of videos. We show that ChipQA achieves state-of-the-art (SOTA) performance on a large new high-motion VQA database. We also test ChipQA on several other VQA databases of professional and user generated content and show that it achieves high-correlations with human judgments of video quality, while also being very computationally efficient.

The paper is organized as follows. In the following section, we briefly review previous work in the area of NR VQA. In section III, we describe our algorithm and its perceptual underpinnings. In section IV we report and analyze results on several large VQA database, and we conclude in section V.

II. PREVIOUS WORK

V-BLIINDS [3] is a no-reference video quality algorithm that models the natural video statistics of the discrete cosine transform (DCT) of frame differences. V-BLIINDS also makes use of features that capture global and local motion coherency. VIIDEO [4] is a “completely blind” NR VQA algorithm, in that it is not trained on a database at all and can be deployed as-is. VIIDEO makes use of the high inter-subband correlations of statistical features that have been observed in pristine videos but not in distorted videos. VIIDEO predicts the quality of videos based on this observation and without any training. Manasa and Channappayya [5] proposed an NR

VQA algorithm based on the statistics of optical flow. The coefficient of variation of the standard deviation of optical flow at different spatial locations is used to quantify irregularities in motion. Dendi and Channappayya [6] proposed a statistical model for the distribution of spatio-temporal bandpass coefficients. The statistics of these coefficients are modelled as following an asymmetric generalized Gaussian distribution. The parameters from the statistical fits are used to predict quality. ChipQA-0 [2] introduced the idea of localized cuts in space-time, ST Chips, which may be viewed as highly localized variations of space-time slices, which are defined over the global range of spatial and temporal video coordinates, instead of locally. The ST Chips in ChipQA-0 were extracted using optical flow, making the algorithm expensive and impractical for use when low-latency is a requirement. The statistics of ST Chips are modelled based on the general observation that natural videos follow regular statistics, and that the regularity of these statistics is disturbed in the presence of distortions. Quantifying these deviations from natural statistics can thus be used to quantify the degree of distortion and the perceptual quality of the video, by learning mappings between these statistics and perception. Finding and describing these statistics is a challenge but many clues about these patterns can be gleaned from the human visual system. The human visual system has adapted to the regular statistics of videos, using them to reduce redundancies in visual signals. Mimicking the front-end visual processes involved in encoding the visual signal, it is possible to reveal departures from these regularities and use them to quantify video quality.

TLVQM [7] is a recent NR VQA algorithm that defines a number of distortion-specific and motion-related features that are relevant to video quality. It does this in two stages. In the first stage, a number of low-complexity features are computed on every frame. These features capture the intensity and spread of motion vectors and also include specific features responsive to blockiness, blur, and interlacing. In the second stage, high-complexity features are computed on one frame each second. These include features that are tailored to capture underexposure, overexposure, noise, blur, blockiness, low contrast, interlacing, low sharpness, low brightness, and low colorfulness. TLVQM has many distortion-specific features and can hence be used as a general purpose NR VQA algorithm in many settings, but it also has many parameters that must be tuned. It represents a different paradigm from natural video statistics-based models, since it does not attempt to model naturalness but instead explicitly models specific distortions. CNN-TLVQM [8] is a variant of this method where deep features are added to TLVQM features to obtain better performance.

MMSP-VQA [9] is a deep learning based approach for NR VQA. It was trained on a very large-scale dataset called FlickrVid-150k. Features are extracted on each frame from multiple layers of an Inception-Resnet-v2 [10] network pre-trained on ImageNet [11]. The features were then averaged and trained with a deep neural network. FlickrVid-150k and the source code for MMSP-VQA have not yet been released.

NR Image Quality Assessment (IQA) algorithms have been found to be quite competitive with NR VQA algorithms

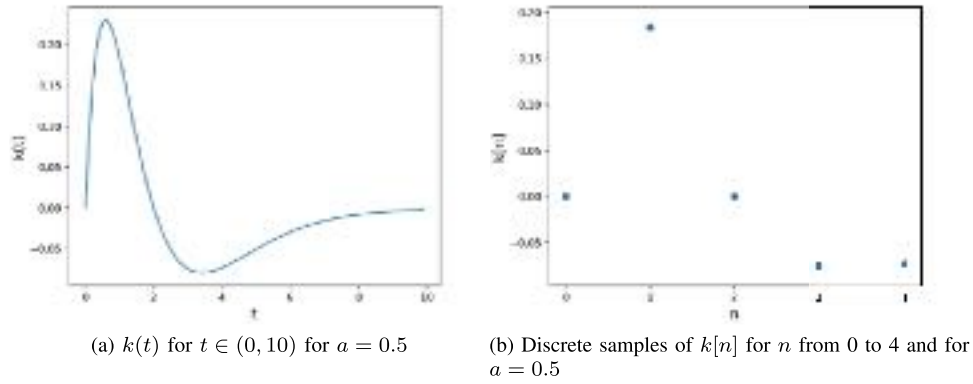


Fig. 2. Continuous and discrete versions of the temporal filter. We use 5 discrete coefficients for the filter in our implementation.

on user-generated content (UGC). This is because UGC is dominated by spatial distortions and does not usually present much temporal variation on quality. NR IQA algorithms such as FRIQUEE [12] and HIGRADE [13] have been found to outperform NR VQA algorithms on datasets such as LIVE VQC [14], Konvid-1k [15], and YouTube-UGC [15]. FRIQUEE uses a bag of perceptually motivated statistical features from different spaces, including luminance, color, and gradients. HIGRADE models the statistics of the log-derivative of gradients, and was designed for HDR content, but has been found to work well on UGC content as well. BRISQUE [16] is an earlier NR IQA algorithm that utilizes a spatial model of the statistics of distorted pictures. BRISQUE models the statistics of bandpass, divisively normalized coefficients of images, based on the observation that bandpass, divisively normalized pristine images follow a first order Gaussian distribution. Distorted images change these statistics, and the parameters of fits to the statistics of an image can be used to reliably predict the quality of an image. The statistical models of distorted pictures discovered in BRISQUE underpin subsequent advances in NSS-based IQA research. NIQE [17] also models spatially bandpassed coefficients, but does not require training. NIQE quantifies the deviation of the statistics of an image via a statistical fit to a small corpus of high-quality natural images. CORNIA [18] is an NR IQA algorithm that does not attempt to model the statistics of natural images, but instead uses a dictionary to effectively represent images for quality assessment. Li *et al.* [19] proposed a CNN based method for UGC quality assessment and trained it on a combination of three major UGC databases.

Space-time slices are cuts of a video through space and time along fixed, pre-determined directions and spanning an entire video. Space-time slices have been effectively used for FR VQA, but are only applicable to stored videos that are available to an algorithm in their entirety [20]–[23]. Space-time chips significantly modify this concept, since they are highly localized in space and time, are sensitive to local motion, and can be used in real-time applications. Space-time chips were first introduced in ChipQA-0 [2], but were found using optical flow in that method. Optical flow is generally expensive to compute, and their requirement can make algorithms computationally impractical, although motion is relevant to any study of video quality. Motion has also been used in several FR VQA

models [24]–[26] and NR VQA models [2], [3], [5], [7]. In our work, we use implicit motion to define the ST Chips used in quality prediction, based on a simple regularity maximizing concept and without using optical flow. Our new model is able to obtain better performance with much lower computational complexity than the original prototype ChipQA-0. ChipQA-0 also performed poorly on UGC databases, while the full model, ChipQA, incorporates temporal filtering, and utilizes chroma and gradient features, yielding a holistic algorithm that performs well on both professional and user generated content.

III. VIDEO QUALITY ASSESSMENT USING SPACE-TIME CHIPS

A. Space-Time Perception

When a video signal is incident on the retina, it is subjected to bandpass spatial filtering expressed at the outputs of the retinal ganglion cells. In a simple model of this process, local spatial averages of the signal are subtracted from the signal, and a form of adaptive gain control is applied on the difference [27]. The resultant signal has a greatly reduced entropy and is carried by the optical nerve at a reduced bandwidth to further stages along the visual pathway. This “contrast signal” is subsequently subjected to temporal entropy reduction filtering ([28]–[31]) which can also be modelled in a simple way as a temporal bandpass filter operation, with filter kernel given by

$$k(t) = t(1 - at) \exp(-2at)u(t), \quad (1)$$

where t denotes time, a is a constant parameter, and $u(t)$ is the unit step function. The function is plotted against t in Fig. 2a. These processes serve to spatially and temporally decorrelate the visual signal. These initial stages of the human visual system motivate the use of spatial and temporal decorrelating functions on videos before analyzing their statistics.

When the visual signal arrives at area V1 (the primary visual cortex) it is decomposed into orientation and scale-tuned spatial and temporal channels. Neurons in area V1 are also sensitive to specific local orientations of motion. From here, the visual signal is passed to area middle temporal (MT) in extrastriate cortex, where further motion processing occurs [24], [28]. MT contains neurons sensitive to motion

over larger spatial fields, and the neural representation of the space-time visual signal at this point makes efficient use of space-time regularity. Similarly, ST Chips are sensitive to local orientations of motions aggregated over large spatial fields, which we use to build spatiotemporal representations of video data.

B. Defining Space-Time Chips

The first step in our algorithm is to compute spatial mean-subtracted and contrast-normalized (MSCN) coefficients of each frame in a given video. Given a luminance image $I[i, j, n]$ at frame index (time) n , the MSCN coefficients $\hat{I}[i, j, n]$ are:

$$\hat{I}[i, j, n] = \frac{I[i, j, n] - \mu[i, j, n]}{\sigma[i, j, n] + C} \quad (2)$$

where $i \in 1, 2..M$, $j = 1, 2..N$ are the spatial indices, M and N are the height and width of the image respectively, C is a constant for numerical stability, and

$$\begin{aligned} \mu[i, j, n] &= \sum_{k=-K}^{k=K} \sum_{l=-K}^{l=K} w[k, l] I[i + k, j + l, n] \\ \sigma[i, j, n] &= \sqrt{\sum_{k=-K}^{k=K} \sum_{l=-K}^{l=K} w[k, l] (I[i + k, j + l, n] - \mu[i, j, n])^2} \end{aligned} \quad (3)$$

$$= \sqrt{\sum_{k=-K}^{k=K} \sum_{l=-K}^{l=K} w[k, l] (I[i + k, j + l, n] - \mu[i, j, n])^2} \quad (4)$$

are the local spatial mean and standard deviation of luminance, respectively. $w = \{w[k, l], k \in -K, \dots, K, l \in -K, \dots, K\}$ is a 2D circularly-symmetric Gaussian weighting function sampled out to 3 standard deviations and rescaled to unit volume. We use $K = 3$ in our implementation. Research on natural image statistics [12], [13], [16] has shown that, in the absence of distortion, the coefficients $\hat{I}[i, j, n]$ can be expected to reliably follow a first-order generalized Gaussian distribution (GGD). This is the basis of state-of-the-art NR IQA algorithms such as FRIQUEE, HIGRADE, and BRISQUE, and is used in VQA algorithms as well to model spatial statistics, e.g., in V-BLIINDS. Moreover, the MSCN operation supplies a reasonable approximation to bandpass processing and adaptive gain control (relevant to contrast masking) that occurs in the retina.

Following spatial MSCN processing, we apply the causal temporal filter (Eq.1) to groups of T' consecutive frames, with no overlap between adjacent groups of frames. This is cheaper than using overlapping blocks and we found that this does not affect performance. The discrete coefficients of the filter, $k[n]$, are shown in Fig. 2b for $a = 0.5$ and the length of the filter $P = 5$. We experimented with different values of a and discuss how they affect performance in the results section. We denote the result of this temporal operation as D .

$$D[i, j, n] = \hat{I}[i, j, n] * k[n] \quad (5)$$

We use reflective padding in the temporal dimension at the boundaries of each block of T' frames such that the output also has T' frames. We fix $P = T'$ to minimize the effect of boundary artifacts, as increasing P to

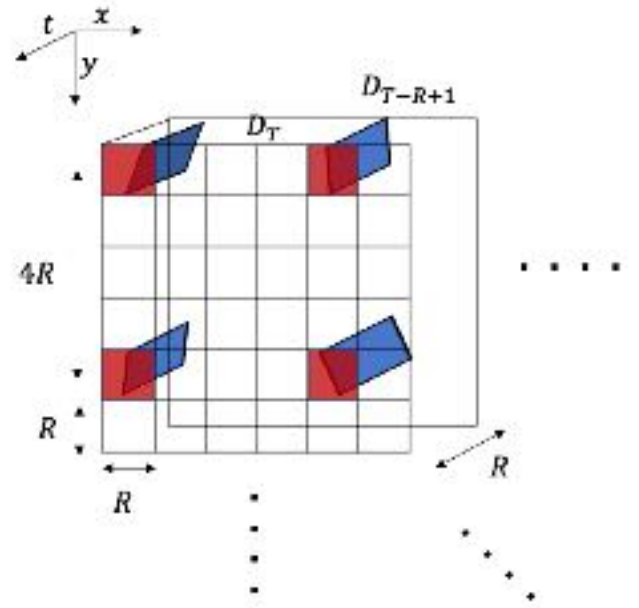


Fig. 3. Extracting ST Chips from a video volume of spatially and temporally decorrelated frames from D_{T-R+1} to D_T . One portion of the video is shown for illustration. ST Chips are extracted by cutting through the volume over $R \times R$ windows (that are spaced apart by $4R$ pixels) R frames back in time. ST Chips are the angled squares in blue, and the windows are shown in red.

be greater than T would result in a greater use of padded points. For ease of representation, denote the processed frame $D[i, j, n]$ for $i \in 1, 2..M$, $j = 1, 2..N$ at time instance n as D_n .

We are interested in finding important directions at different spatial locations along which ST Chips can be extracted. In our experimental model, we fixed $T' = R$ for simplicity, so that each ST chip is extracted from an $R \times R \times R$ volume. At a particular time instance T , consider the output of the previous operation D_n over indices $n = T - R + 1$ to $n = T$, which is a single block of R frames. Divide D_T into spatial windows of size $R \times R$. For each $R \times R$ window, we define chips that pass through the block of frames from D_T backwards in time to D_{T-R+1} , and that are constrained to pass through the center of the $R \times R$ window such that the normal vector to any chip lies on the xy plane. Some examples of chips are shown in Fig. 5 (in blue, with the $R \times R$ windows in red) and chips for a single $R \times R \times R$ volume are shown in Fig. 4. These chips can be oriented at diverse angles. Among these angles, one is assumed to best capture the local motion, and a chip that is oriented perpendicular to the motion vector at this location will capture objects in motion along the motion vector. This is illustrated in Fig. 1. Under these constraints on the chip, we are assuming that the motion is along a vector on the xy plane, which relies implicitly on the assumption that motion is linear and translational in small spatiotemporal volumes. This is a reasonable assumption that forms the basis of most modern motion estimation algorithms [32], [33].

In our earlier work [2], we found the directions of motion explicitly using optical flow. This is expensive and depends on the accuracy of the optical flow algorithm. Assuming that motion is smooth for a pristine video, we expect the chips that are perpendicular to the directions of motion to follow similar

statistics as natural images, since they contain projections of natural scenes as they move. The MSCNs of natural images are known to reliably obey a Gaussian law [3], [16], [17], [27], [34]. Assuming the veracity of these natural image statistics models, and smoothness and linearity of motion in local regions of pristine videos, we find the directions of motions implicitly by selecting a chip having a sample set that is closest to being Gaussian amongst all of the potential chips. This is done in a simple and direct way by computing the sample kurtosis [35] of each chip along Q different equally spaced angles, from 0 to π , and selecting the chip that has the kurtosis closest to 3, which is the kurtosis of a Gaussian random variable, as is illustrated in Fig. 4. There are many possible tests of the Gaussianity of a set of chip samples. These include the Shapiro-Wilk test [36], the Anderson-Darling test [37], the Martinez-Iglewicz test [38], and the D’agostino kurtosis test [39], which is similar to the kurtosis method that we apply. There are a number of reasons we use the simple sample kurtosis. First, we are not actually testing for Gaussianity, which these frequentist tests are designed for. Rather, we are instead ranking the chips by kurtosis and selecting which among them is most Gaussian in that sense. It is possible that all, none, or a subset of the chips may pass a given Gaussianity test, e.g., if there is little or no motion present, all may present as Gaussian. Ranking procedures on test statistics like those in [36]–[39] have not been shown to measure relative Gaussianity. Further, by using the sample kurtosis we are aligning with the powerful *a priori* and well-founded assumption that the bandpass chips will reliably obey a zero-mean generalized Gaussian law. The members of this distribution class may be viewed as differing *only* in kurtosis, hence we may view our use of kurtosis as a conditional measure. Given the small sample size of 25, this is a powerful constraint. Lastly, the computational efficiency of the kurtosis lends it to fast implementations. We chose $Q = 6$ in our implementation, and found that increasing Q improves performance, although computational cost increases as well. Variation in performance as Q is varied is discussed in section IV-E.

Having selected the chip that is most Gaussian at a window, we then aggregate them across windows. We do not collect chips from all $R \times R$ windows, but skip $D = 4$ windows in each of the x and y directions. We study how performance varies with D in Section IV-D. The centers of the windows from which ST Chips are extracted are thus separated from each other by a distance of $4R$ pixels in each dimension. This is shown in Fig. 3. We discuss how this spatial downsampling affects performance in Section IV-E. The aggregated chips form a single “frame” S for every group of T' frames. We discretize coordinates while searching for the best chip such that each chip is of dimension $R \times R$. The dimension of the aggregated frame S of ST Chips is $M' \times N'$, where $M' = \frac{R}{4} \lfloor \frac{M}{R} \rfloor$ and $N' = \frac{R}{4} \lfloor \frac{N}{R} \rfloor$. We chose $R = T' = 5$ in our implementation. Variation in performance as R is varied is discussed in section IV.

We repeated the process described above for the spatial gradient magnitude field of the video as well. Gradients contain important information descriptive of edges and contrast variations and have been found to be useful for image and video quality assessment. Gradient-based features find a place

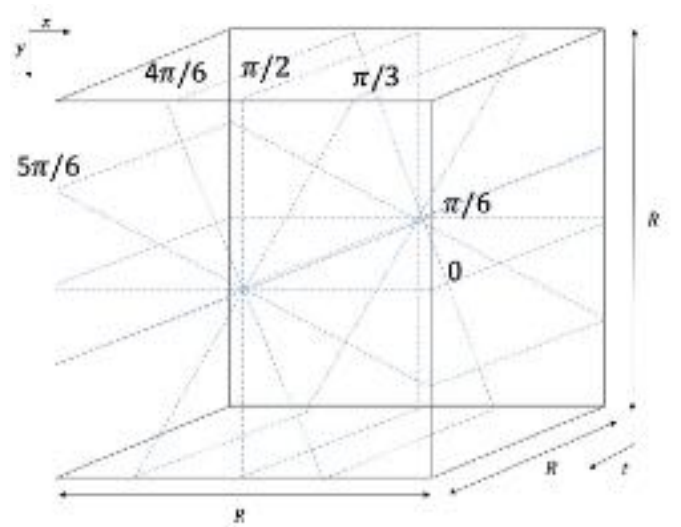


Fig. 4. Finding the best ST-Chip over a particular $R \times R$ spatial window. Chips are extracted from a $R \times R$ video volume along 6 angles that are equally spaced from 0 to π . The angles are shown next to their corresponding chips. The chip which has the minimum excess kurtosis is selected as the chip that best captures motion. This criterion is based on the Gaussianity of natural images and the smoothness of motion.

in most SOTA algorithms [7], [12], [13], [40]. ChipQA computes the gradient components in the vertical and horizontal directions using a Sobel kernel of size 3×3 . The Sobel filter eliminates low frequency information and has high-pass characteristics that detects edges. The statistics of these edges are useful for quality assessment since they are often heavily affected by distortions. We then find the MSCNs of the gradient magnitude, apply the temporal filter $k[n]$, and extract ST Chips along the directions with kurtosis closest to 3 at windows that are separated by a distance of $4R$ in each dimension. We refer to these as “ST Gradient Chips”.

C. Statistics of ST Chips

MSCNs of spatial frames and of gradient fields are known to follow regular statistics, and this is true of ST Chips of spatial frames and of gradient fields as well because of the smoothness of motion. ST Chips and ST Gradient Chips are found to follow a generalized Gaussian distribution (GGD) of the form:

$$f(x; \alpha; \beta) = \frac{\alpha}{2\beta\Gamma(\frac{1}{\alpha})} \exp\left(-\left(\frac{|x|}{\beta}\right)^\alpha\right) \quad (6)$$

where $\Gamma(\cdot)$ is the gamma function:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t) dt. \quad (7)$$

The shape parameter α of the GGD and the variance of the distribution are estimated using the moment-matching method described in [41]. Examples of the first-order distribution of ST Chips and ST Gradient Chips are shown in Fig. 5 and Fig. 6 respectively. Though the ST chip is chosen to be as Gaussian as possible, previous research has shown that Gaussianity breaks in the presence of distortions [13], [16], and for ST chips the statistics could also deviate from Gaussianity for large motion fields. These deviations from Gaussianity are useful for quantifying losses in quality, and we find that the GGD is able to model these deviations in the statistics well.

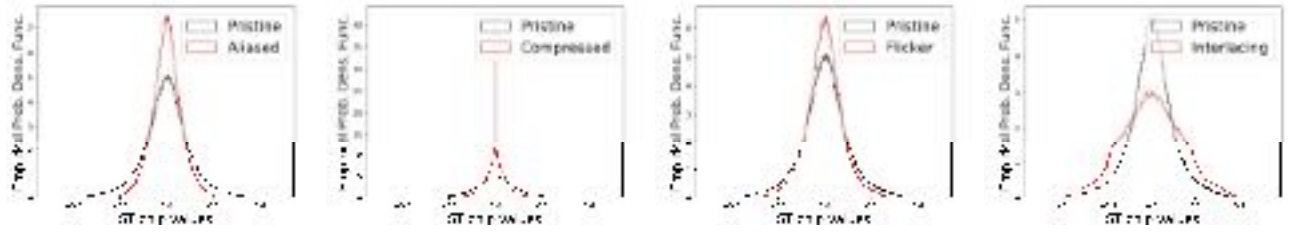


Fig. 5. Empirical distributions of ST Chips. Pristine (original) distributions are in black and distorted distributions are in red.

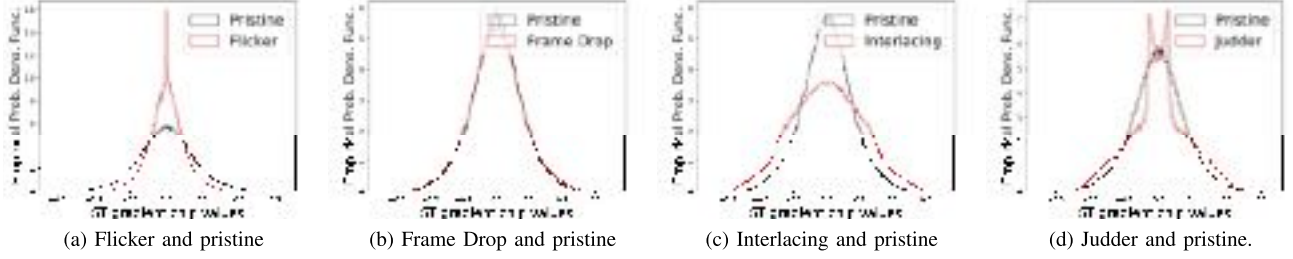


Fig. 6. Empirical distributions of ST Gradient Chips. Pristine (original) distributions are in black and distorted distributions are in red.

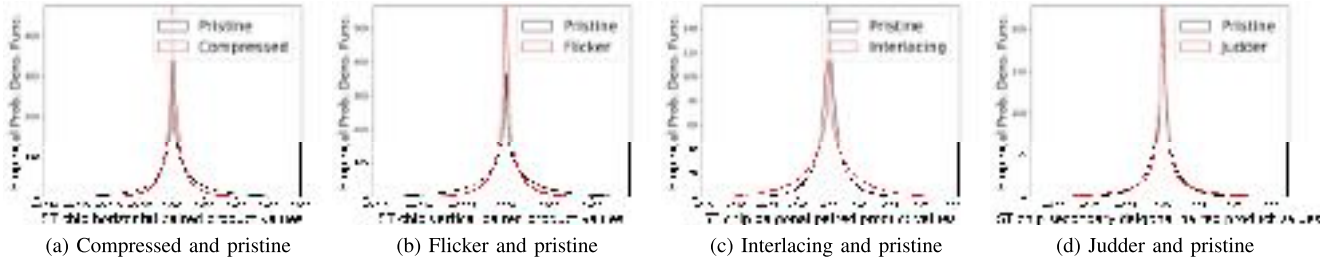


Fig. 7. Empirical distributions of paired products of ST-Chip. Pristine (original) distributions are in black and distorted distributions are in red.

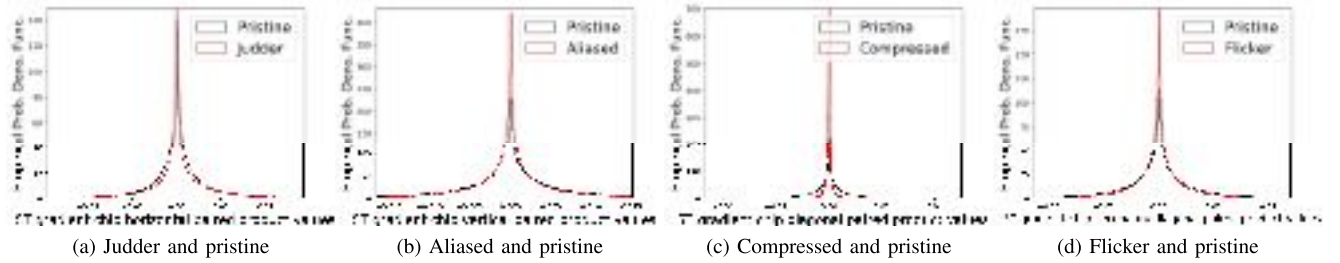


Fig. 8. Empirical distributions of paired products of ST Gradient Chips. Pristine (original) distributions are in black and distorted distributions are in red.

We also model the second-order statistics of ST Chips. Define the collection of ST Chips aggregated at each time instance T as S , and define the pairwise products

$$\begin{aligned} H[i, j, T] &= S[i, j, T]S[i, j + 1, T] \\ V[i, j, T] &= S[i, j, T]S[i + 1, j, T] \\ D_1[i, j, T] &= S[i, j, T]S[i + 1, j + 1, T] \\ D_2[i, j, T] &= S[i, j, T]S[i + 1, j - 1, T] \end{aligned} \quad (8)$$

H , V , D_1 , and D_2 are found to follow a asymmetric generalized Gaussian distribution (AGGD), which is given by:

$$f(x; \nu, \sigma_l^2, \sigma_r^2) = \begin{cases} \frac{\nu}{(\beta_l + \beta_r)\Gamma(\frac{1}{\nu})} \exp(-(\frac{x}{\beta_l})^\nu) & x < 0 \\ \frac{\nu}{(\beta_l + \beta_r)\Gamma(\frac{1}{\nu})} \exp(-(\frac{x}{\beta_r})^\nu) & x > 0 \end{cases} \quad (9)$$

where

$$\beta_l = \sigma_l \sqrt{\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}} \quad \text{and} \quad \beta_r = \sigma_r \sqrt{\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}} \quad (10)$$

where ν controls the shape of the distribution and σ_l and σ_r control the spread on each side of the mode. The parameters $(\eta, \nu, \sigma_l^2, \sigma_r^2)$ are extracted from the best AGGD fit to each pairwise product, where

$$\eta = (\beta_r - \beta_l) \frac{\Gamma(\frac{2}{\nu})}{\Gamma(\frac{1}{\nu})}. \quad (11)$$

Empirical histograms of the paired products of ST Chips and ST Gradient Chips are shown in Fig. 7 and Fig. 8, respectively. Images and videos are inherently multiscale, and distortions can manifest themselves differently at different scales. We compute ST Chips at two scales, and compute the above features at each scale. We first apply low-pass filtering

to the video and then downsample it to half the original size. ST Chips are computed as described previously, and statistical features are extracted from the lower scale.

D. Spatial Features

Spatial features are important for video quality assessment. Since user-generated content is often dominated by spatial distortions, NR IQA algorithms such as FRIQUEE and HIGRADE achieve very good performance on UGC datasets. Compression, aliasing, and interlacing are common spatial distortions that affect professional-grade content as well. In VIDEVAL [40], top-performing features were selected from a number of state-of-the-art algorithms. It was found that luma, color, and gradient information were vital to building a competitive VQA algorithm. Likewise, we incorporate features that describe the statistics of luma, color, and gradient magnitude in our algorithm.

1) Luma:

- **NIQE naturalness:** We compute luminance features and a naturalness score based on them every T' frames using the image naturalness index NIQE. They consist of parameters of GGD and AGGD fits to spatial MSCNs of selected patches of luma in each frame, and a naturalness score that measures the distance of the parameters from a statistical fit to a corpus of natural images. These features are averaged over the entire video over all non-overlapping groups of T' frames.
- **σ map:** We also model the statistics of the bandpass standard deviation σ in (4). The σ map is calculated using equation 4. The MSCNs of the σ map also follows a GGD, as shown in [34], we extract the shape, variance, skewness, and kurtosis of the distribution. These features are averaged over the entire video clip. Previous research [7], [40], [42] has also shown that standard deviation pooling of features over time is useful for video quality assessment. We therefore compute the standard deviation of these features over every non-overlapping five-frame interval, and average the standard deviation values over the entire video. Again, $T' = 5$ frames are used for ST chip computation, hence the pooling is copacetic with the computation of ST Chips.

2) *Color:* We also model the statistics of the chrominance of videos. We use the CIELAB [43] color space, which is designed to model human perception of color. CIELAB has a luminance channel (L^*) and two chrominance channels (a^* and b^*), where a^* denotes the position of the color along the red-green axis, and b^* denotes the position of the color on the yellow-blue axis. Chroma (C) captures the intensity of a color, and is defined as

$$C = \sqrt{a^{*2} + b^{*2}} \quad (12)$$

We compute the chroma map for each frame in the video, and find the MSCNs of the chroma map. The MSCNs are known to follow a first order GGD [12], and the shape, variance, skewness, and kurtosis are extracted from the empirical distribution of each frame.

We also compute the σ map of the chroma, using (4). We extract the shape, variance, skewness, and kurtosis of the distribution. These features are averaged over the entire video clip. We also find the average standard deviation of features over $T' = 5$ frame intervals, as described for the σ calculations in luma space.

E. Gradients

Gradients are known to capture important information about edges, and since some distortions modify (reduce or increase) gradients, they have been effectively used to predict video quality in [7], [12], [13], [40], [44]. We find the gradients of the luminance in the vertical and horizontal directions using a Sobel kernel of size 3. We then find the gradient magnitude at each pixel and compute the MSCNs of the gradient magnitude field. The second order statistics of the MSCNs of the gradient magnitude are particularly useful for predicting video quality, when combined with the features defined previously. The paired products of the MSCNs of the gradient magnitude are computed using (8), and modelled using AGGDs (9). Four parameters are extracted on each paired product of the gradient magnitude at two scales. Just as for chroma, we average these features across time and also find the average standard deviation over groups of 5 frames.

F. Quality Assessment

Table I gives a summary of all the features used in ChipQA. A total of 221 features are extracted on each video, starting from the $T' = 5^{\text{th}}$ frame. These are trained with a support vector regressor, as described in the following section.

IV. EXPERIMENTS AND RESULTS

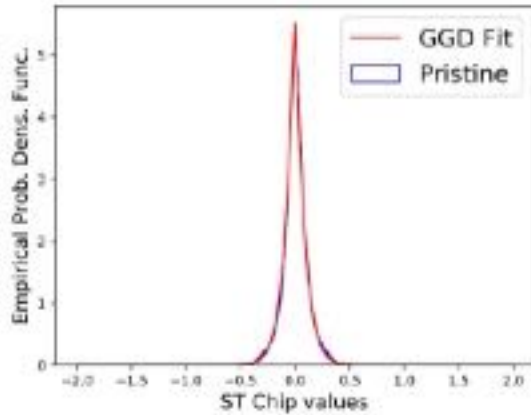
A. Validating ST Chips

We conducted a series of experiments to examine the effectiveness of our method of generating ST chips. First, we studied how well our method is able to predict motion directions. The goal of our design is not to conduct very accurate motion estimation, since it is not necessary for the end goal of perceptual video quality measurement; the perception of absolute motion by the human visual system is not particularly accurate. Instead, we are interested in obtaining motion data that is consistent with statistical information available to human perception, rather than from complex search or optimization processes. Nevertheless, it is of interest to evaluate the efficacy of our simple kurtosis-based motion orientation selector. However, several challenges arise when attempting to quantitatively evaluate ChipQA's implicit motion orientation estimation process:

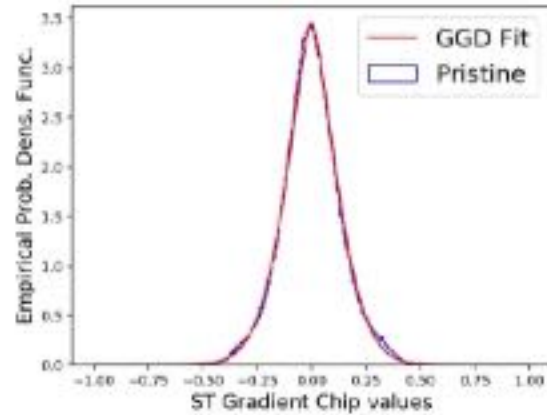
- Many popular optical flow databases [45]–[47] only have optical flow on pairs of images, while ChipQA requires at least 5 frames to select motion directions.
- ChipQA only finds the directions of motions, not the magnitudes. Therefore some standard metrics such as endpoint error (EPE) cannot be used to evaluate ChipQA.
- ChipQA computes the motion directions of patches, while most optical flow algorithms and ground truth data compute motion at the pixel level.

TABLE I
DESCRIPTORS OF FEATURES IN ChipQA

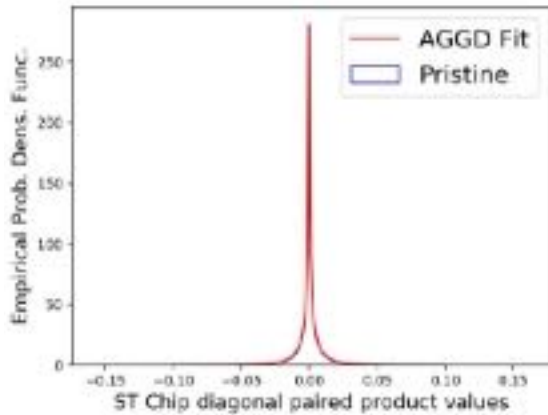
Domain	Description	Feature index
Chroma	Average across all frames of GGD shape, GGD scale, skewness, and kurtosis at two scales.	$f_1 - f_8$
Chroma σ map	Average across all frames of GGD shape, GGD scale, skewness, and kurtosis at two scales.	$f_9 - f_{16}$
Gradient	Average across all frames of four parameters from AGGD fitted to pairwise products at two scales.	$f_{17} - f_{48}$
Luma σ map	Average across all frames of GGD shape, GGD scale, skewness, and kurtosis at two scales.	$f_{49} - f_{56}$
Chroma	Standard deviation across 5 frames of GGD shape, GGD scale, skewness, and kurtosis at two scales, averaged across all non-overlapping groups of 5 frames.	$f_{57} - f_{64}$
Chroma σ map	Standard deviation across 5 frames of GGD shape, GGD scale, skewness, and kurtosis at two scales, averaged across all non-overlapping groups of 5 frames.	$f_{65} - f_{72}$
Gradient	Standard deviation across 5 frames of four parameters from AGGD fitted to pairwise products at two scales, averaged across all non-overlapping groups of 5 frames.	$f_{73} - f_{104}$
Luma σ map	Standard deviation across 5 frames of GGD shape, GGD scale, skewness, and kurtosis at two scales, averaged across all non-overlapping groups of 5 frames.	$f_{105} - f_{112}$
NIQE naturalness	Average over every 5 frames of features and scores of spatial naturalness index NIQE.	$f_{113} - f_{149}$
ST-Chip	Average over all chips of shape and scale parameters from GGD fits and four parameters from AGGD fits to pairwise products at two scales.	$f_{150} - f_{185}$
ST Gradient Chips	Average over all chips of shape and scale parameters from GGD fits and four parameters from AGGD fits to pairwise products at two scales.	$f_{186} - f_{221}$



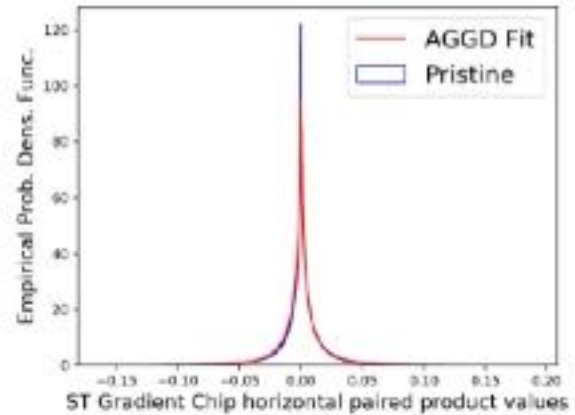
(a) GGD fitted to ST Chip distribution



(b) GGD fitted to ST Gradient Chip distribution



(c) AGGD fitted to ST Chip diagonal paired product distribution



(d) AGGD fitted to ST Gradient Chip horizontal paired product distribution

Fig. 9. Fits to empirical distributions. Fits are in red and histograms are in blue.

- ChipQA computes motion directions in quantized steps of $\pi/6$ from 0 to π , while the ground truth data in optical flow databases are generally more fine-grained.
- ChipQA does not differentiate between the angles θ and $\pi + \theta$ of any motion direction θ , since the chips along both directions are the same.

The Sintel optical flow database [48] is a viable way to examine the implicit motion estimation of ChipQA, since it has optical flow ground truth on entire videos. Sintel is not a database of natural videos but of 3D animated films, so the videos from this database do not necessarily obey natural video statistics models. However, since animations are often made

to be reasonably naturalistic, we have found that natural video statistics models appear to hold well on this database.

As discussed earlier, absolute motion metrics are not appropriate for evaluating ChipQA. We therefore used the mean absolute angular deviation (MAAD) between the ground truth angular directions of motion θ (in radians) and the angles of motion θ' predicted by ChipQA:

$$\text{MAAD}(\theta, \theta') = |\theta - \theta'|. \quad (13)$$

We compared ChipQA's performance against the classical New Three Step Search algorithm (NTSS) [49], which is a block motion algorithm, and the Farneback dense optical flow predictor [50]. NTSS is used in VBLINDS and some implementations of the MPEG and H.26x codec families. Farneback is used in ChipQA-0 and is the optical flow algorithm in the popular OpenCV library. Since ChipQA computes motion over patches, in order to compare the ground truth motion angles with those of ChipQA, the ground truth motion vectors were first averaged across 5×5 patches and across 5 frames in time. We then found those angles along which these vectors are oriented and mapped them to $[0, \pi)$ by replacing all angles $\theta > \pi$ with $\theta - \pi$. We did the same for the motion vectors predicted by NTSS and Farneback, so that they could be compared to ground truth in the same way as ChipQA. The results are shown in Table II, which shows that ChipQA was competitive with respect to MAAD, outperforming NTSS but was not as accurate than Farneback. The effectiveness of ChipQA is quite remarkable, given its extremely simple design. Indeed, we view these experiments as suggestive of the type of information that might be used by the visual brain to compute motion.

We also conducted a separate experiment whereby we studied the variation of kurtosis of ST chips on the Sintel database, as the angular difference between the chip's orientation and the ground truth was varied. We computed the kurtosis of all the chips on the Sintel database (on all videos), for fixed differences between the true motion direction and each chip's orientation, and plotted the results as the differences were increased by steps of $\pi/6$, in Fig. 10. When the angular difference was 0, i.e., when all the chips across all videos were exactly aligned with the ground truth motion vectors, the average kurtosis was closest to 3, the kurtosis of a Gaussian. As the angular difference was increased by steps of $\pi/6$, the average kurtosis diverged from 3 up to angular difference $\pi/2$, after which it again converged towards 3 as the angular cycle completed. This provides further validation of the simple angular motion estimator used in ChipQA, while providing insights into the relationship between natural video statistics and motion.

B. Databases

We evaluated our algorithm on four large databases, which are described below:

- 1) LIVE Livestream VQA database [51] - This is a new database containing 315 professional-grade videos with synthetic distortions. 45 videos are of pristine quality, and 7 different distortions were synthetically applied to

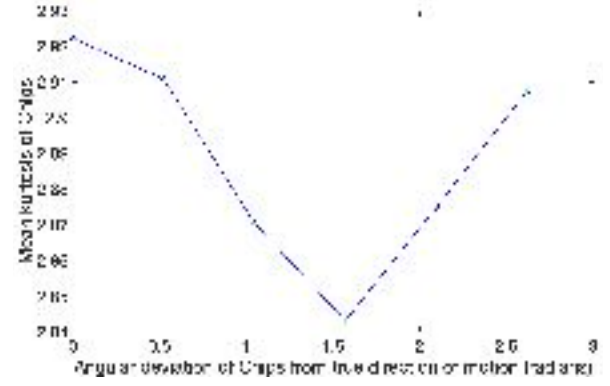


Fig. 10. Mean kurtosis across all chips in Sintel vs. the angular deviation from the true motion direction.

TABLE II
MEAN ABSOLUTE ANGULAR DIFFERENCE (MAAD) BETWEEN
PREDICTED AND ACTUAL DIRECTION OF MOTION θ FOR
SINTEL DATABASE

SEQUENCE	ChipQA	NTSS	FARNEBACK
alley_1	0.9133 (0.5975)	1.2202 (0.5934)	0.4295 (0.5410)
alley_2	1.0526 (0.7281)	1.1773 (0.7723)	0.7439 (0.8335)
ambush_2	1.0453 (0.7261)	0.9914 (0.6943)	0.8535 (0.6754)
ambush_4	1.0316 (0.7273)	0.9871 (0.6842)	0.9603 (0.7493)
ambush_5	0.9759 (0.6902)	1.0201 (0.7237)	0.9439 (0.7267)
ambush_6	1.0354 (0.7381)	0.8732 (0.5989)	0.7790 (0.5633)
ambush_7	1.0891 (0.7740)	1.0856 (0.7536)	0.9086 (0.8656)
bamboo_1	0.8173 (0.4868)	0.9146 (0.4904)	0.4114 (0.3262)
bamboo_2	1.3628 (0.8921)	1.3069 (0.5563)	1.5411 (1.3005)
bandage_1	0.8503 (0.6008)	0.7377 (0.5720)	0.6774 (0.5623)
bandage_2	0.9607 (0.6911)	0.8218 (0.6342)	0.8610 (0.7053)
cave_2	0.8766 (0.5860)	0.8615 (0.5611)	0.5518 (0.5526)
cave_4	0.9357 (0.6451)	0.9018 (0.6236)	0.6130 (0.6235)
market_2	0.8616 (0.5444)	0.8612 (0.6297)	0.3914 (0.4283)
market_5	1.0810 (0.7546)	1.0584 (0.7651)	0.8849 (0.7895)
market_6	1.1404 (0.7934)	1.1665 (0.7833)	1.0946 (0.8966)
mountain_1	1.0020 (0.6989)	1.0628 (0.7388)	0.9432 (0.6799)
shaman_2	0.8969 (0.6403)	0.8540 (0.6044)	0.6092 (0.6132)
shaman_3	1.1720 (0.8081)	1.2471 (0.8956)	1.0625 (0.7017)
sleeping_1	1.1975 (0.8345)	1.2405 (0.8591)	1.0812 (0.7469)
sleeping_2	1.0476 (0.7201)	1.0677 (0.7126)	0.8090 (0.8862)
temple_2	0.8724 (0.5483)	0.8519 (0.5546)	0.5240 (0.5110)
temple_3	0.8945 (0.6144)	0.9933 (0.6217)	0.7312 (0.6574)
ALL	1.0048 (0.1294)	1.0131 (0.1549)	0.8002 (0.2607)

each pristine video to create 315 synthetically distorted videos. The distortions that were applied are compression, aliasing, interlacing, judder, flicker, and frame-drop. All videos are of resolution 3840×2160 , and the study was conducted on a 4K TV. 40 people participated in the study.

- 2) LIVE ETRI database [52] - The LIVE ETRI database contains 437 videos generated by applying various levels of combined space-time subsampling and video compression on 15 diverse video contents. The bitrates in this database vary from around 500 kbps to around 50 Mbps, and the database has 30 fps, 60 fps, and 120 fps videos. A total of 34 subjects took part in the study.
- 3) YouTube UGC [53] - The YouTube UGC database consists of 1500 20 second clips sampled from millions of user-generated YouTube videos. Spatial and temporal features were used to ensure the sampled videos were diverse and representative of videos on YouTube. Each

video was rated by more than 100 subjects. These videos are of different resolutions and include content from categories such as animation and VR.

- 4) Konvid-1k [15] - This database consists of 1200 videos of authentically distorted user-generated content. All videos are of resolution 960×540 . Videos in this database are known to not have significant temporal variation in quality [40], [54]. Each video has 114 subject ratings on average.
- 5) LIVE Video Quality Challenge (VQC) [14] - LIVE VQC contains 585 videos of authentically distorted unique user-generated content. Each video was labeled by 40 human subjects on average. All videos are of user-generated content.

C. Evaluation Details

We used a support vector regressor (SVR) with a radial basis function as the learning engine for all our experiments. 20% of the data was randomly selected for testing and 80% was used for training and validation. 5-fold cross-validation was performed on the training set to find the best set of hyperparameters for the SVR. Content separation is performed for all experiments on the LIVE Livestream database and the LIVE ETRI database, hence videos containing the same content are always in the same fold of either training, validation, or testing. This prevents scores from being artificially boosted. We repeated 1000 random splits on LIVE Livestream and LIVE ETRI, which are databases containing synthetic distortions and reference videos. We conducted 100 splits for the Konvid-1K, VQC, and YT-UGC databases, which are UGC databases. Databases containing synthetic distortions generally result in larger standard deviations of the results, which is why we performed more train-validation-test splits on them. Grid search was performed over values of the kernel coefficient γ , and the regularization parameter C that controls the squared L2 penalty. γ was geometrically increased by 10 from 10^{-8} to 10. C was doubled from 2 to 1024. We report Spearman's Rank Order Correlation Coefficient (SROCC), Pearson's Linear Correlation Coefficient (LCC), and the root mean square error (RMSE) between the predicted scores and the mean opinion scores for each algorithm. SROCC measures the monotonicity of the relationship between the two quantities, while the LCC measures the linear correlation. Since the relationship between the predicted scores and the MOS may not necessarily be linear, the predicted scores s were first passed through a logistic non-linearity [55]

$$f(s) = \beta_1 \left(\frac{1}{2} - \frac{1}{(1 + \exp(\beta_2(s - \beta_3)))} \right) + \beta_4 s + \beta_5, \quad (14)$$

before computing the LCC. The parameters are found by fitting $f(s)$ to the MOS.

D. Implementation Details

We used the luma definition from ITU Recommendation BT.709 [56]. We use a frame-buffer of length $T' = 5$ frames to compute ST Chips, keeping in mind potential applications of our algorithm in scenarios requiring low-latency, where

TABLE III
MEDIAN SROCC FOR 1000 SPLITS ON THE LIVE LIVESTREAM VQA DATABASE FOR DIFFERENT VALUES OF K , WHERE WINDOW SIZE IS $2K + 1 \times 2K + 1$

K	SROCC
3	0.7967
5	0.7882
7	0.7963

TABLE IV
MEDIAN SROCC FOR 1000 SPLITS ON THE LIVE LIVESTREAM VQA DATABASE FOR DIFFERENT VALUES OF a AND $R = T'$

a	$R = T'$	SROCC
0.25	5	0.7847
0.25	9	0.7274
0.5	5	0.7967
0.75	5	0.7869
0.75	9	0.7864

using a larger buffer size might cause delays in the overall system. With $a = 0.5$, the discrete 5-tap causal filter $k[n]$ is representative of the continuous filter $k(t)$. After $t = 8$, there was a slight increase in the response which we rectified to 0 by choosing P (length of filter) = T' (length of frame buffer) = 5, since in any case the filter converges to 0 as t tends to infinity. Using a temporal filter of length greater than the frame buffer could cause serious boundary artifacts, hence we fixed the length of the filter (P) to be the length of the frame buffer (T'), which was fixed at 5. We also fixed $R = T'$, so that each ST chip is square and is chosen from a $5 \times 5 \times 5$ volume. The SROCC values obtained on the LIVE Livestream VQA database for different values of T' and a are shown in Table IV. The Table shows that taking $a = 0.5$ and $R = 5$ not only provides a discrete filter that is representative of $k(t)$, but also provides the best predictions.

We used a window size of $2K + 1 \times 2K + 1$, where $K = 3$, for MSCN computation. We also studied how performance varied as this parameter was changed. Consistent with previous studies [16] of spatial MSCNs, we found that the size of the window did not significantly affect performance. We report the median SROCC over 1000 splits of the LIVE Livestream VQA database for different values of K , in Table III. Since increasing K increases the computational cost, we used $K = 3$ in our final implementation.

A look-up table is used to implement the search for the best ST-chip direction. Coordinates along different directions from $\theta = 0$ to $\theta = \pi/Q$ are pre-computed and rounded to the nearest integer, since pixel coordinates are integer values. The look-up table coordinates are computed using the polar form, and are indexed by the value of θ and r , where r varies from $-(R+1)/2 - 1$ to $(R+1)/2$, where $R \times R$ is the dimension of each ST chip. Values are read from the look-up table during the search for the best direction. We fixed $Q = 6$ in our implementation, and thus search 6 directions at each $R \times R$ window to find the direction that best captures motion at that location. Results for different values of Q are shown in Table VI for ChipQA for $R = 5$. Increasing Q improves performance, but beyond $Q = 6$ performance seems to drop off.

We also studied the effects of the chip downsampling factor D . The performance of the algorithm on the LIVE Livestream database for $D = 1$, $D = 4$ and $D = 8$ are presented

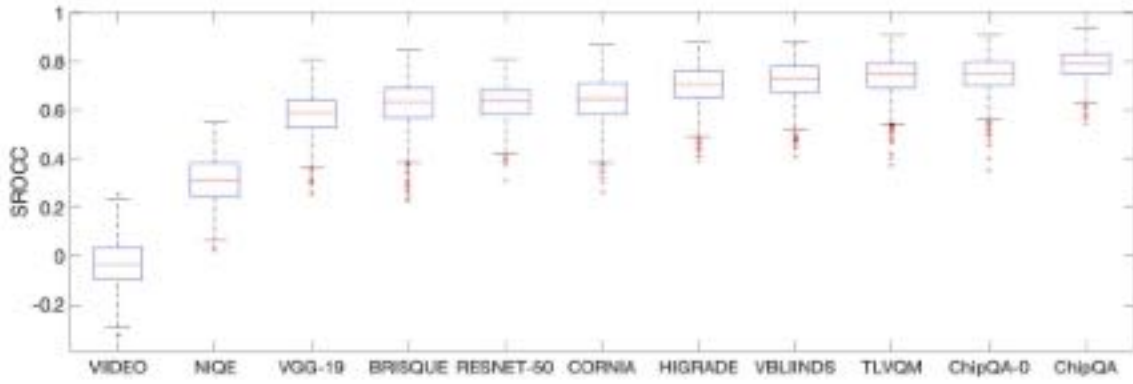


Fig. 11. Box-plot of distribution of SROCCs for different algorithms for 1000 splits of the LIVE Livestream database.

TABLE V
MEDIAN SROCC FOR 1000 SPLITS ON THE LIVE LIVESTREAM
VQA DATABASE FOR DIFFERENT VALUES OF CHIP
DOWNSAMPLING FACTOR D

D	SROCC
1	0.7877
4	0.7967
8	0.7920

TABLE VI
MEDIAN SROCC FOR 1000 SPLITS ON THE LIVE LIVESTREAM
VQA DATABASE FOR DIFFERENT VALUES OF Q

Q	SROCC
3	0.7852
4	0.7658
6	0.7967
9	0.7626

in Table V. Increasing D greatly reduces the cost of ChipQA (see Table XIX), and using all of the chips is not necessary to conduct effective video quality assessment. We used $D = 4$ in our final implementation.

E. Video Quality Assessment Results

Results on the LIVE Livestream VQA database, the LIVE ETRI database, Konvid-1k, LIVE VQC database and the YouTube UGC are shown in Tables VII, VIII, X, IX, and XI respectively. The scores of each top-performing algorithm are boldfaced in each table. We also give box-plots of the SROCCs over the 1000 splits on the LIVE Livestream database in Figure 11. ChipQA was able to achieve state-of-the-art performance on all databases. In particular, ChipQA significantly outperformed other compared algorithms on the LIVE Livestream database (Table VII) and the LIVE ETRI database (Table VIII), both of which contain videos captured by professional videographers subjected to distortions commonly encountered at different stages of professional video capturing pipelines. These distortions include judder, frame-drop, space-time subsampling, etc. ChipQA produced the least variation in SROCC across the splits of the LIVE Livestream database as can be seen from the box plot and the standard deviations of the SROCCs reported in Table VII. ChipQA also performed very competitively on the UGC databases, which are known to be dominated by spatial distortions [40], [54], hence image quality algorithms were among the top performing algorithms for UGC databases.

We also evaluated two deep networks (VGG-19 [57] and ResNet-50 [10]) on the LIVE Livestream database. Both networks were pre-trained on ImageNet and used to extract features from 25 random 227×227 crops from one frame per second on the LIVE Livestream database. The 4096 outputs of the fully connected layer of VGG-19, and the 2048 outputs of the average-pooled layer of ResNet-50 for each patch

were averaged across all patches, forming single 4096 and 2048 dimensional vectors, respectively, for each video. These video-level vectors were then used as features to train an SVR to predict video quality, using the same cross validation method as described earlier. These feature-extractors have been shown to perform quite well on full-reference tasks [58] as well as on UGC databases [40]. However, they did not perform as well as the other methods on the LIVE Livestream database, which presents a significant variety and amount of temporal distortions.

We ensured that videos of the same content did not appear in the same fold of training, validation, or testing. Reference videos and all of their distorted versions always appeared in the same fold. This ensured that the algorithm would not rely on learning content quality, and to prevent overfitting. We also made use of the regularization parameter C , chosen by cross-validation to prevent overfitting by the SVR. We also measured the difference between the training and test SROCC to verify that ChipQA was not overfitting to the database, and found that the median difference between the SROCC on the training set and the SROCC on the test set was 0.1598 (16%).

F. Performance by Distortion

We also evaluated each algorithm on each single distortion in the LIVE Livestream database. The reference video and the synthetically distorted video for each content were used to evaluate the performance of each algorithm on each distortion separately. The results of this experiment are shown in Table XII and Table XIII. ChipQA performed very competitively on each distortion as well. The LIVE Livestream database has a number of scenes with high motion and ChipQA appears to be able to predict video quality well even for complex, high-motion scenes. TLVQM uses “jerkiness” and “jerkiness consistency” features which measure the similarity of motion vectors in the forward and backward directions,

TABLE VII

MEDIAN SROCC, LCC, AND RMSE FOR LIVE LIVESTREAM DATABASE. STANDARD DEVIATIONS ARE IN PARENTHESES. BEST PERFORMING ALGORITHM IS BOLD-FACED

METHOD		SROCC↑	LCC↑	RMSE↓
IMAGE QUALITY METRICS	NIQE [17]	0.3104 (0.0963)	0.48674 (0.3323)	11.2799 (1.0362)
	VGG-19 [57]	0.5887 (0.0853)	0.6600 (0.0687)	9.3951 (0.6587)
	ResNet-50 [10]	0.6395 (0.0793)	0.7011 (0.0666)	8.9941 (0.6361)
	BRISQUE [16]	0.6381 (0.1035)	0.6841 (0.0835)	9.1330 (0.9514)
	HIGRADE [13]	0.7088 (0.0805)	0.7247 (0.0742)	9.0958 (1.1249)
	CORNIA [18]	0.6482 (0.0977)	0.6826 (0.0819)	9.6472 (1.0725)
VIDEO QUALITY METRICS	VIIDEO [4]	0.0005 (0.0965)	0.2643 (0.1670)	12.0508 (0.9087)
	V-BLIINDS [3]	0.7330 (0.0799)	0.7609 (0.0699)	8.2047 (1.0585)
	TLVQM [7]	0.7503 (0.0815)	0.7660 (0.0752)	8.0392 (1.2181)
	ChipQA-0 [2]	0.7513 (0.0762)	0.7667 (0.0702)	8.0195 (1.1147)
	ChipQA	0.7967 (0.0632)	0.8179 (0.0566)	7.6002 (1.0448)

TABLE VIII

MEDIAN SROCC, LCC, AND RMSE FOR LIVE ETRI. STANDARD DEVIATIONS ARE IN PARENTHESES. BEST PERFORMING ALGORITHM IS BOLD-FACED

METHOD		SROCC↑	LCC↑	RMSE↓
IMAGE QUALITY METRICS	NIQE [17]	0.3966 (0.1983)	0.4435 (0.3049)	12.2663 (1.6530)
	BRISQUE [16]	0.2656 (0.3042)	0.4315 (0.2461)	12.3339 (1.5116)
	TLVQM [7]	0.2343 (0.1934)	0.3018 (0.1957)	13.0663 (1.2398)
VIDEO QUALITY METRICS	V-BLIINDS [3]	0.4798 (0.1411)	0.5344 (0.1194)	11.7581 (1.4058)
	ChipQA-0 [2]	0.4012 (0.1591)	0.4634 (0.1356)	12.1476 (1.3464)
	ChipQA	0.6323 (0.1474)	0.6822 (0.1182)	10.0769 (1.4701)

TABLE IX

MEDIAN SROCC, LCC, AND RMSE FOR KONVID 1K DATABASE. STANDARD DEVIATIONS ARE IN PARENTHESES. BEST PERFORMING ALGORITHM IS BOLD-FACED

METHOD		SROCC↑	LCC↑	RMSE↓
IMAGE QUALITY METRICS	NIQE [17]	0.3559 (0.0544)	0.2533 (0.3338)	0.6318 (0.0299)
	BRISQUE [16] (1 fps)	0.5876 (0.0422)	0.5989 (0.0407)	0.5118 (0.0244)
	HIGRADE [13]	0.7310 (0.0311)	0.7390 (0.0296)	0.4306 (0.0219)
	FRIQUEE [12] (1 fps)	0.7452 (0.0274)	0.7506 (0.0260)	0.4212 (0.0208)
	CORNIA [18] (1 fps)	0.7685 (0.0253)	0.7671 (0.0240)	0.4093 (0.0196)
VIDEO QUALITY METHODS	VIIDEO [4]	0.3107 (0.0492)	0.3261 (0.0500)	0.6043 (0.0243)
	TLVQM [7]	0.7750 (0.0242)	0.7715 (0.0226)	0.4058 (0.0178)
	V-BLIINDS [3]	0.7127 (0.0355)	0.7085 (0.0330)	0.4523 (0.0210)
	ChipQA-0 [2]	0.6973 (0.0311)	0.6943 (0.0311)	0.4600 (0.0221)
	ChipQA	0.7629 (0.0260)	0.7625 (0.0256)	0.4105 (0.0206)

TABLE X

MEDIAN SROCC, LCC, AND RMSE FOR LIVE VQC. STANDARD DEVIATIONS ARE IN PARENTHESES. BEST PERFORMING ALGORITHM IS BOLD-FACED

METHOD		SROCC↑	LCC↑	RMSE↓
IMAGE QUALITY METRICS	NIQE [17]	0.4603 (0.0735)	0.3782 (0.4259)	16.7678 (1.3756)
	BRISQUE [16] (1 fps)	0.6192 (0.0529)	0.6519 (0.0470)	12.7489 (0.7664)
	HIGRADE [13]	0.6103 (0.068)	0.6332 (0.065)	13.027 (0.904)
	FRIQUEE [12] (1 fps)	0.6579 (0.053)	0.7000 (0.058)	12.198 (0.914)
	CORNIA [18] (1 fps)	0.6719 (0.047)	0.7183 (0.042)	11.832 (0.700)
VIDEO QUALITY METRICS	TLVQM [7]	0.8026 (0.0359)	0.7996 (0.0363)	10.1033 (0.7920)
	VIIDEO [4]	-0.0336 (0.0770)	-0.0064 (0.0954)	16.8163 (0.9094)
	V-BLIINDS [3]	0.7005 (0.0457)	0.7251 (0.0466)	11.4744 (0.7483)
	ChipQA-0 [2]	0.6692 (0.0532)	0.6965 (0.0477)	12.1574 (0.8065)
	ChipQA	0.7192 (0.0550)	0.7299 (0.0479)	11.6014 (0.8370)

TABLE XI

MEDIAN SROCC, LCC, AND RMSE FOR YouTube UGC. STANDARD DEVIATIONS ARE IN PARENTHESES. BEST PERFORMING ALGORITHM IS BOLD-FACED

METHOD		SROCC↑	LCC↑	RMSE↓
IMAGE QUALITY METRICS	NIQE [17]	0.1788 (0.050)	0.1788 (0.094)	0.6390 (0.026)
	BRISQUE [16] (1 fps)	0.3820 (0.051)	0.3952 (0.048)	0.5919 (0.021)
	HIGRADE [13]	0.7376 (0.033)	0.7216 (0.033)	0.4471 (0.024)
	FRIQUEE [12] (1 fps)	0.7652 (0.03)	0.7571 (0.032)	0.4169 (0.023)
	CORNIA [18] (1 fps)	0.5972 (0.041)	0.6057 (0.039)	0.5136 (0.024)
VIDEO QUALITY METRICS	TLVQM [7]	0.6693 (0.03)	0.6590 (0.03)	0.4849 (0.022)
	V-BLIINDS [3]	0.5590 (0.049)	0.5551 (0.046)	0.5356 (0.022)
	ChipQA-0 [2]	0.5595 (0.038)	0.5572 (0.040)	0.5386 (0.024)
	ChipQA	0.7014 (0.031)	0.6911 (0.033)	0.4729 (0.025)

and which are hand-crafted to capture frame drops, which is probably why it does so well on the frame drop category of distortions. CORNIA constructs codewords from image patches, which can capture simple patterns such as salt and

TABLE XII

SROCC OF THE COMPARED NR VQA MODELS ON DIFFERENT SUBSETS OF THE LIVE LIVESTREAM DATABASE. THE SCORES OF THE TOP PERFORMING ALGORITHM IN EACH CATEGORY ARE BOLDFACED

METHOD	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
NIQE(1 fps)	0.3232	0.2775	0.2860	0.2863	0.2832	0.2842	0.2780
BRISQUE(1 fps)	0.6381	0.5748	0.7564	0.8235	0.6574	0.3849	0.8689
VIIDEO	0.0044	0.0053	0.0073	0.0055	0.0013	0.0024	0.0064
CORNIA(1 fps)	0.6778	0.6894	0.7853	0.7657	0.7049	0.2776	0.8824
HIGRADE(1 fps)	0.6916	0.6244	0.7141	0.7441	0.6440	0.6130	0.8287
V-BLIINDS	0.7330	0.6450	0.7606	0.8679	0.6182	0.7131	0.8060
TLVQM	0.7503	0.5614	0.7420	0.8328	0.6202	0.8555	0.8173
ChipQA-0	0.7513	0.6594	0.7791	0.8513	0.6491	0.6780	0.8534
ChipQA	0.7575	0.7028	0.7792	0.8740	0.7647	0.7647	0.8513

TABLE XIII

PLCC OF THE COMPARED NR VQA MODELS ON THE LIVE LIVESTREAM DATABASE. THE SCORES OF THE THREE TOP PERFORMING ALGORITHMS ARE BOLDFACED

METHOD	OVERALL	COMPRESSION	ALIASING	JUDDER	FLICKER	FRAME DROP	INTERLACING
NIQE(1 fps)	0.4962	0.2805	0.2865	0.2860	0.2848	0.2849	0.2850
BRISQUE(1 fps)	0.6698	0.7345	0.9321	0.8726	0.7268	0.3902	0.9118
VIIDEO	0.2155	0.1222	0.1222	0.1235	0.1247	0.1256	0.1259
CORNIA(1 fps)	0.7257	0.8115	0.9508	0.7875	0.8063	0.2138	0.9142
HIGRADE(1 fps)	0.6990	0.7595	0.9353	0.7614	0.6889	0.6062	0.8800
V-BLIINDS	0.7477	0.8055	0.9202	0.9200	0.7086	0.7443	0.8873
TLVQM	0.7513	0.6788	0.9273	0.8914	0.7724	0.8738	0.8358
ChipQA-0	0.7565	0.7783	0.9490	0.9071	0.6609	0.6945	0.9075
ChipQA	0.7705	0.8719	0.9735	0.9472	0.8989	0.8730	0.9133

TABLE XIV

ABLATION STUDY WITH MEDIAN SROCC FOR 1000 SPLITS ON THE LIVE LIVESTREAM VQA DATABASE AS DIFFERENT FEATURE SPACES ARE REMOVED

RANK	FEATURE SPACES REMOVED	SROCC
1	ST chip+ST Gradient chip	0.7376
2	Luma	0.7805
3	ST Gradient chip	0.7865
4	Gradient	0.7876
5	ST chip	0.7963
6	Chroma	0.7967
7	Chroma+Gradient	0.8064

TABLE XV

ABLATION STUDY WITH MEDIAN SROCC FOR 100 SPLITS ON THE KONVID-1K DATABASE AS DIFFERENT FEATURE SPACES ARE REMOVED

RANK	FEATURE SPACES REMOVED	SROCC
1	Chroma+Gradient	0.7308
2	Luma	0.7448
3	Gradient	0.7460
4	ST chip + ST Gradient chip	0.7552
5	Chroma	0.7589
6	ST Gradient chip	0.7641
7	ST chip	0.7669

pepper noise, blockiness, and interlacing. Interlacing has a very specific pattern of even and odd fields which lends itself to codeword construction, which is probably why CORNIA does so well on that category of distortions.

G. Ablation Studies and Feature Ranking

We performed ablation studies of the studied feature spaces on the LIVE Livestream VQA database and the Konvid-1k database. The Konvid-1k database was chosen to

be representative of UGC datasets. The results are shown in Table XIV and Table XV. Each feature space contains important quality-aware information about the video, but removing some of the spatial features (chroma and gradient features) appears to improve performance on LIVE Livestream. On the other hand, removing chroma and gradient features severely affects performance on Konvid-1k, showing that spatial features are vital for UGC but can only bring about the curse of dimensionality for datasets that have a significant amount of temporal distortion. We also find that removing ST chip features and ST gradient chip features from the algorithm severely affects performance on LIVE Livestream, but does not significantly affect performance on Konvid-1k. This suggests that temporal information is not as important in current UGC databases, validating earlier work that came to the same conclusion [40], [54].

We also studied performance when just the ST Chips, ST Gradient Chips, and the NIQE features were used. This set of features corresponds to what was used in ChipQA-0, but with our different and novel method of finding the chips, and removing all the other feature spaces that are in ChipQA but not in ChipQA-0. The median SROCC for just this set of features is 0.7901, greater than the 0.7513 SROCC for ChipQA-0, showing that ChipQA is conclusively better than ChipQA-0, and that our novel method of finding chips contributes significantly to the boost in performance.

We also ranked the importance of individual features in the database using sequential forward selection. We started with an empty set and then added a single feature to the feature set that maximized the median SROCC over 100 content-separated splits of the LIVE Livestream database. Repeating

TABLE XVI
FEATURE RANKING FOR THE LIVE LIVESTREAM VQA DATABASE BASED ON SEQUENTIAL FORWARD SELECTION

RANK	FEATURE INDEX	FEATURE DESCRIPTION	SROCC*
1	130	Variance of right half of MSCN distribution of luma	0.6515
2	204	GGD shape of ST chips at half-scale	0.7133
3	149	NIQE score	0.7411
4	9	GGD shape of MSCNs of Chroma σ map	0.7696
5	25	AGGD shape of pairwise products (along the main diagonal) of MSCNs of the gradient magnitude	0.8066
6	150	GGD shape of ST chips at full-scale	0.8258
7	161	AGGD shape of pairwise products (along the main diagonal) of MSCNs of ST chips	0.8217
8	123	AGGD shape of pairwise products (along the main diagonal) of MSCNs of luma	0.8043
9	29	AGGD shape of pairwise products (along the horizontal) of the gradient magnitude	0.8062
10	98	Variance of left half of MSCN distribution of pairwise product (along the main diagonal) of MSCNs of the gradient magnitude at half-scale	0.8134

*Note: This is the median SROCC obtained when the feature in the row entry is added to the set of the features ranked above it.

TABLE XVII
CROSS DATABASE SROCC FOR METHODS TRAINED ON ROW ENTRY AND EVALUATED ON COLUMN ENTRY FOR DATABASES WITH PROFESSIONALLY-CAPTURED CONTENT

Method		Livestream	ETRI
CHIPQA	Livestream	-	0.1286
	ETRI	0.2201	-
TLVQM	Livestream	-	0.1478
	ETRI	0.2568	-
VBLIINDS	Livestream	-	0.3067
	ETRI	0.4331	-

TABLE XVIII
CROSS DATABASE SROCC FOR METHODS TRAINED ON ROW ENTRY AND EVALUATED ON COLUMN ENTRY FOR UGC DATABASES

Method		VQC	Konvid	YT-UGC
CHIPQA	VQC	-	0.5569	0.0757
	Konvid	0.5478	-	0.0817
	YT-UGC	0.5436	0.4030	-
TLVQM	VQC	-	0.6026	0.1768
	Konvid	0.6983	-	0.3241
	YT-UGC	0.3279	0.5588	-
VBLIINDS	VQC	-	0.6323	0.0969
	Konvid	0.6320	-	0.1677
	YT-UGC	0.1026	0.2072	-

this process, we obtained an ordering of the features. The top 10 features and their descriptions are presented in Table XVI. We also report the median SROCC obtained when each feature is added to the set of features ranked higher than it. The NIQE score is, not expectedly, one of the top 10 features. The GGD shape of the ST chips at both scales are also among the top 10 features. Gradients and chroma information also supply features that lie among the top 10.

H. Cross Database Performance

We also evaluated the generalizability of ChipQA, TLVQM, and VBLIINDS across the studied databases. LIVE Livestream and LIVE ETRI are databases of professionally captured content and are thus treated separately from Konvid, VQC, and YT-UGC, which are UGC databases. Results of the cross-database performance comparisons on LIVE Livestream and LIVE ETRI are shown in Table XVII, while results on the UGC databases are shown in Table XVIII. An SVR was trained on the features from the database in the row entry and the SROCC when the trained SVR was used to predict

scores using features from the database on the column entry are reported. All of the compared methods yielded poor cross-database performance, primarily because these databases have very different characteristics in terms of both distortion type and content type. On UGC videos, cross database performance between Konvid and UGC was much better than on other databases among all the compared methods, probably because their contents are similar. YT-UGC spans a very wide range of content, including animations, gaming, and VR, hence these videos do not share many properties with the videos of natural scenes found in VQC and Konvid, which is probably why the cross-database performances were so poor on YT-UGC. The distortions in the LIVE Livestream and LIVE ETRI databases are very different, involving high motions and temporal subsampling, respectively, though they both contain professionally-captured content. LIVE Livestream presents distortions such as interlacing, judder, flicker etc, while LIVE ETRI contains compression and temporal subsampling artifacts, which may explain the poor generalizability of all three VQA methods on these two databases. In summary, NR VQA algorithms trained on one type of database sometimes perform quite poorly on others. While this might suggest weaknesses of the predictive models, more likely it is because the databases contain very different contents.

I. Computational Complexity

We found the computational cost of each algorithm by computing the time each required to generate features on a single 4K video containing 210 frames. We also provide a crude estimate of the $O(n)$ computational complexity and the number of giga floating point operations (GFLOPS) for each algorithm. The results are shown in Table XIX. ChipQA was the most efficient NR VQA algorithm, much more efficient than applying SOTA NR IQA algorithms on each frame. The algorithms are not implemented in the same language and may not be optimized, hence the computational times and FLOPS cannot be compared directly, but since most practitioners use these algorithms off-the-shelf, this provides a rough estimate of relative complexity. All the algorithms run in linear time but differ in the coefficient of linearity. The compute times for the IQA algorithms were obtained assuming they would be applied on all frames of the video. The times taken for

TABLE XIX

COMPUTATIONAL ANALYSIS OF ALGORITHMS ON A SINGLE 3840×2160 VIDEO WITH 210 FRAMES FROM THE LIVE LIVESTREAM VQA DATABASE

METHOD	Number of features	Language	Computational Complexity	GFLOPS	Time (s)
BRISQUE [16]	36	MATLAB	$\mathcal{O}(d^2 NT)$ d : MSCN window size	352	301
ChipQA	221	Python	$\mathcal{O}((\frac{d^2}{D^2} + \frac{Q}{RD^2} + \frac{Q \log Q}{R^3 D^2}) NT)$ d : MSCN window size, D : down-sampling factor Q : chip search's quantization factor, R : size of each dimension of a chip	700	814
TLVQM [7]	75	MATLAB	$\mathcal{O}((h_1^2 N + k^2 K) T_1 + h_2^2 N T_2)$ h_1, h_2 : filter size, k : motion estimation block size, K : number of key points	477	1002
NIQE [17]	1	MATLAB	$\mathcal{O}(d^2 NT)$ d : MSCN window size	3094	1008
CORNIA [18]	10000	MATLAB	$\mathcal{O}(d^2 K NT)$, d : window size, K : codebook size	4480	2056
ChipQA-0 [2]	109	Python	$\mathcal{O}(d^2 + 1.5(1 - 0.5^{P+1}) I w^2) NT$, d : MSCN window size, P : number of pyramids for optical flow, I : number of iterations per pyramid, w : window size for optical flow	2605	2655
VBLINDS [3]	47	Python	$\mathcal{O}((d^N + \log(k)N + k^2 w^3) T)$, d : MSCN window size, k : DCT block size, w : motion vector tensor size	465	3086
VIIDEO [4]	1	Python	$\mathcal{O}(N \log(N) T)$	2094	6403
HIGRADE [13]	216	MATLAB	$\mathcal{O}(3(2d^2 + k) NT)$, d : MSCN window size, k : gradient kernel size	9604	16240
FRIQUEE [12]	560	MATLAB	$\mathcal{O}((f^2 dN + 4N(\log(N) + m^2)) T)$, h_1, h_2 : filter size, f : number of color spaces, k : motion estimation block size, K : number of key points	470,505	93240

*Note: N - number of pixels per frame. T - number of frames in video. T_1 - total number of frames divided by 2. T_2 - number of frames sampled at 1 frame per second.

FRIQUEE and HIGRADE are rough estimates obtained by finding the times taken on a single frame and multiplying by the number of frames. All the algorithms were run on an Intel Xeon E5-2620 CPU with a maximum frequency of 3 GHz. ChipQA is much faster than ChipQA-0, because it does not involve the use of optical flow, and also because it performs spatial and temporal downsampling.

V. CONCLUSION

We presented a new algorithm for no-reference video quality assessment that is highly competitive with the state-of-the-art and is computationally very efficient. ST Chips are a novel feature space and the fact that they follow regular statistics represents a significant advance in our understanding of natural videos. We showed how these statistics can be parametrized, and how the parameters of the statistical fits can be used to predict video quality without the use of any distortion-specific features. ChipQA achieves high correlations with human judgments of video quality, especially of high motion videos, and is also computationally very efficient.

ACKNOWLEDGMENT

The authors thank Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported in this paper. URL: <http://www.tacc.utexas.edu>.

REFERENCES

- [1] Cisco. (Accessed: Aug. 23, 2020). *Annual Internet Report (2018-2023)*. Accessed: 2020. [Online]. Available: <https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/>
- [2] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, and A. C. Bovik, "No-reference video quality assessment using space-time chips," 2020, *arXiv:2008.00031*. [Online]. Available: <http://arxiv.org/abs/2008.00031>
- [3] 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.
- [4] 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. 2015.
- [5] K. Manasa and S. S. Channappayya, "An optical flow-based no-reference video quality assessment algorithm," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2400–2404.
- [6] 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.
- [7] 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.
- [8] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3311–3319.
- [9] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild," 2019, *arXiv:1912.07966*. [Online]. Available: <http://arxiv.org/abs/1912.07966>
- [10] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [12] 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.
- [13] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped HDR pictures," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, Jun. 2017.
- [14] 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. 2018.
- [15] V. Hosu *et al.*, "The Konstanz natural video database (KoNViD-1k)," in *Proc. 9th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2017, pp. 1–6.
- [16] 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.
- [17] 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, Nov. 2012.
- [18] 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.

- [19] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-world videos with mixed datasets training," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1238–1257, Apr. 2021.
- [20] L. Liu, T. Wang, H. Huang, and A. C. Bovik, "Video quality assessment using space-time slice mappings," *Signal Process., Image Commun.*, vol. 82, Mar. 2020, Art. no. 115749.
- [21] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, pp. 1–25, 2014.
- [22] P. Yan, X. Mou, and W. Xue, "Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices," in *Mobile Devices and Multimedia: Enabling Technologies, Algorithms*, vol. 9411. Bellingham, WA, USA: SPIE, 2015, pp. 182–191.
- [23] P. Yan and X. Mou, "Video quality assessment based on motion structure partition similarity of spatiotemporal slice images," *J. Electron. Imag.*, vol. 27, no. 3, pp. 1–13, 2018.
- [24] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2009.
- [25] 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. (ICASSP)*, Apr. 2007, pp. I-869–I-872.
- [26] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.
- [27] M. Wainwright, O. Schwartz, and E. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons," in *Probabilistic Models of the Brain*, R. Rao, B. Olshausen, and M. Lewicki, Eds. Cambridge, MA, USA: MIT Press, 2002, pp. 203–222.
- [28] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vis. Res.*, vol. 38, no. 5, pp. 743–761, 1998.
- [29] D. W. Dong and J. J. Atick, "Temporal decorrelation: A theory of lagged and nonlagged responses in the lateral geniculate nucleus," *Netw., Comput. Neural Syst.*, vol. 6, no. 2, pp. 159–178, 1995.
- [30] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 2, pp. 284–299, 1985.
- [31] A. B. Watson and A. J. Ahumada, "Model of human visual-motion sensing," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 2, pp. 322–342, 1985.
- [32] B. K. P. Horn and B. G. Schunck, "Determining optical flow," in *Techniques and Applications of Image Understanding*, vol. 281. Bellingham, WA, USA: SPIE, 1981, pp. 319–331.
- [33] M. J. Black and P. Anandan, "Robust dynamic motion estimation over time," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1991, pp. 296–302.
- [34] D. L. Ruderman, "The statistics of natural images," *Netw., Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.
- [35] S. Kokoska and D. Zwillinger, *CRC Standard Probability and Statistics Tables and Formulae*. Boca Raton, FL, USA: CRC Press, 2000.
- [36] S. S. Shapiro and R. S. Francia, "An approximate analysis of variance test for normality," *J. Amer. Stat. Assoc.*, vol. 67, no. 337, pp. 215–216, Mar. 1972.
- [37] G. Marsaglia and J. C. W. Marsaglia, "Evaluating the Anderson-Darling distribution," *J. Stat. Softw.*, vol. 9, no. 2, pp. 1–5, Feb. 2004.
- [38] J. Martinez and B. Iglewicz, "A test for departure from normality based on a biweight estimator of scale," *Biometrika*, vol. 68, no. 1, pp. 331–333, 1981.
- [39] R. B. D'agostino, A. Belanger, and R. B. D'agostino, "A suggestion for using powerful and informative tests of normality," *Amer. Statistician*, vol. 44, no. 4, pp. 316–321, Nov. 1990.
- [40] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," 2020, *arXiv:2005.14354*. [Online]. Available: <http://arxiv.org/abs/2005.14354>
- [41] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, Feb. 1995.
- [42] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.
- [43] J. Schanda, *Colorimetry: Understanding the CIE System*. Hoboken, NJ, USA: Wiley, 2007.
- [44] 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," 2021, *arXiv:2101.10955*. [Online]. Available: <http://arxiv.org/abs/2101.10955>
- [45] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, 2011.
- [46] A. Dosovitskiy et al., "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
- [47] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.
- [48] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, vol. 7577, A. Fitzgibbon, Ed. Springer, Oct. 2012, pp. 611–625.
- [49] R. Li, B. Zeng, and M. L. Liou, "A new three-step search algorithm for block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, no. 4, pp. 438–442, Aug. 1994.
- [50] G. Farneback, "Very high accuracy velocity estimation using orientation tensors, parametric motion, and simultaneous segmentation of the motion field," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, Jul. 2001, pp. 171–177.
- [51] Z. Shang, J. P. Ebenezer, A. C. Bovik, Y. Wu, H. Wei, and S. Sethuraman, "Assessment of subjective and objective quality of live streaming sports videos," 2021, *arXiv:2106.08431*. [Online]. Available: <https://arxiv.org/abs/2106.08431>
- [52] D. Y. Lee et al., "A subjective and objective study of space-time subsampled video quality," 2021, *arXiv:2102.00088*. [Online]. Available: <http://arxiv.org/abs/2102.00088>
- [53] 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.
- [54] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," 2020, *arXiv:2002.10651*. [Online]. Available: <http://arxiv.org/abs/2002.10651>
- [55] 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, Nov. 2006.
- [56] ITU. (Accessed: May 15, 2021). *BT.709: Parameter Values for the HDTV Standards for Production and International Programme Exchange*. Accessed: 2015 [Online]. Available: <https://www.itu.int/rec/R-REC-BT.709-6-201506-I/en>
- [57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
- [58] 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.