

Exploring the Effectiveness of Video Perceptual Representation in Blind Video Quality Assessment

Liang Liao
S-lab, NTU, Singapore
liang.liao@ntu.edu.sg

Kangmin Xu
School of Computer Science, WHU,
China
xukangmin@whu.edu.cn

Haoning Wu
S-lab, NTU, Singapore
haoning001@e.ntu.edu.sg

Chaofeng Chen
S-lab, NTU, Singapore
chaofeng.chen@ntu.edu.sg

Wenxiu Sun
Qiong Yan
SenseTime Research and Tetras AI
irene.wenxiu.sun@gmail.com
sophie.yanqiong@gmail.com

Weisi Lin
S-lab, NTU, Singapore
wslin@ntu.edu.sg

ABSTRACT

With the rapid growth of in-the-wild videos taken by non-specialists, blind video quality assessment (VQA) has become a challenging and demanding problem. Although lots of efforts have been made to solve this problem, it remains unclear how the human visual system (HVS) relates to the temporal quality of videos. Meanwhile, recent work has found that the frames of natural video transformed into the perceptual domain of the HVS tend to form a straight trajectory of the representations. With the obtained insight that distortion impairs the perceived video quality and results in a curved trajectory of the perceptual representation, we propose a temporal perceptual quality index (TPQI) to measure the temporal distortion by describing the graphic morphology of the representation. Specifically, we first extract the video perceptual representations from the lateral geniculate nucleus (LGN) and primary visual area (V1) of the HVS, and then measure the straightness and compactness of their trajectories to quantify the degradation in naturalness and content continuity of video. Experiments show that the perceptual representation in the HVS is an effective way of predicting subjective temporal quality, and thus TPQI can, for the first time, achieve comparable performance to the spatial quality metric and be even more effective in assessing videos with large temporal variations. We further demonstrate that by combining with NIQE, a spatial quality metric, TPQI can achieve top performance over popular in-the-wild video datasets. More importantly, TPQI does not require any additional information beyond the video being evaluated and thus can be applied to any datasets without parameter tuning. Source code is available at <https://github.com/UoLMM/TPQI-VQA>.

CCS CONCEPTS

• Computing methodologies → Computer vision tasks.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/XXXXXXX.XXXXXXX>

KEYWORDS

Perceptual trajectories, primary visual cortex, temporal modeling, blind video quality assessment

ACM Reference Format:

Liang Liao, Kangmin Xu, Haoning Wu, Chaofeng Chen, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2000. Exploring the Effectiveness of Video Perceptual Representation in Blind Video Quality Assessment. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Recently, video streams have exploded on social media platforms, and most of them are captured by users in the wild with portable mobile devices [1, 34, 41]. Compared to videos from professionals, in-the-wild videos usually suffer from complicated distortion issues such as out of focus, over/under-exposure, and camera shake. Therefore, it is highly desirable to have an automatic quality assessment to eliminate low-quality videos or improve them during the acquisition and enhancement process. As these videos do not have pristine counterparts, a blind video quality assessment (VQA) is required. Although significant algorithms for blind images quality assessment (IQA) have been proposed [17, 27, 30, 33, 60], for videos, the temporal-domain quality is another integral aspect of blind VQA, since video perception is highly correlated with motion and temporal variations, leading VQA a more challenging problem.

Up till now, some efforts have been made in modeling the quality in the temporal domain for VQA. One direct way is to compute frame-level quality scores and then express their relative importance over time by applying temporal pooling to these frame-level quality scores [43, 44, 57]. However, the temporal pooling of spatial quality ignores the motion among frames. The most popular approach to temporal quality modeling is to deploy regular parametric bandpass models of natural scene statistics (NSS), such as 3D discrete cosine transform (3D-DCT) [20] and 3D mean-subtracted contrast-normalized (3D-MSCN) coefficients [32, 37], which characterize the perceived quality degradation by predicting the deviation of the distribution of frame-difference coefficients in the presence of distortion. However, in-the-wild videos contain authentic and commonly intermixed distortions, making NSS-based models designed for one specific distortion in each video unsuitable. Inspired

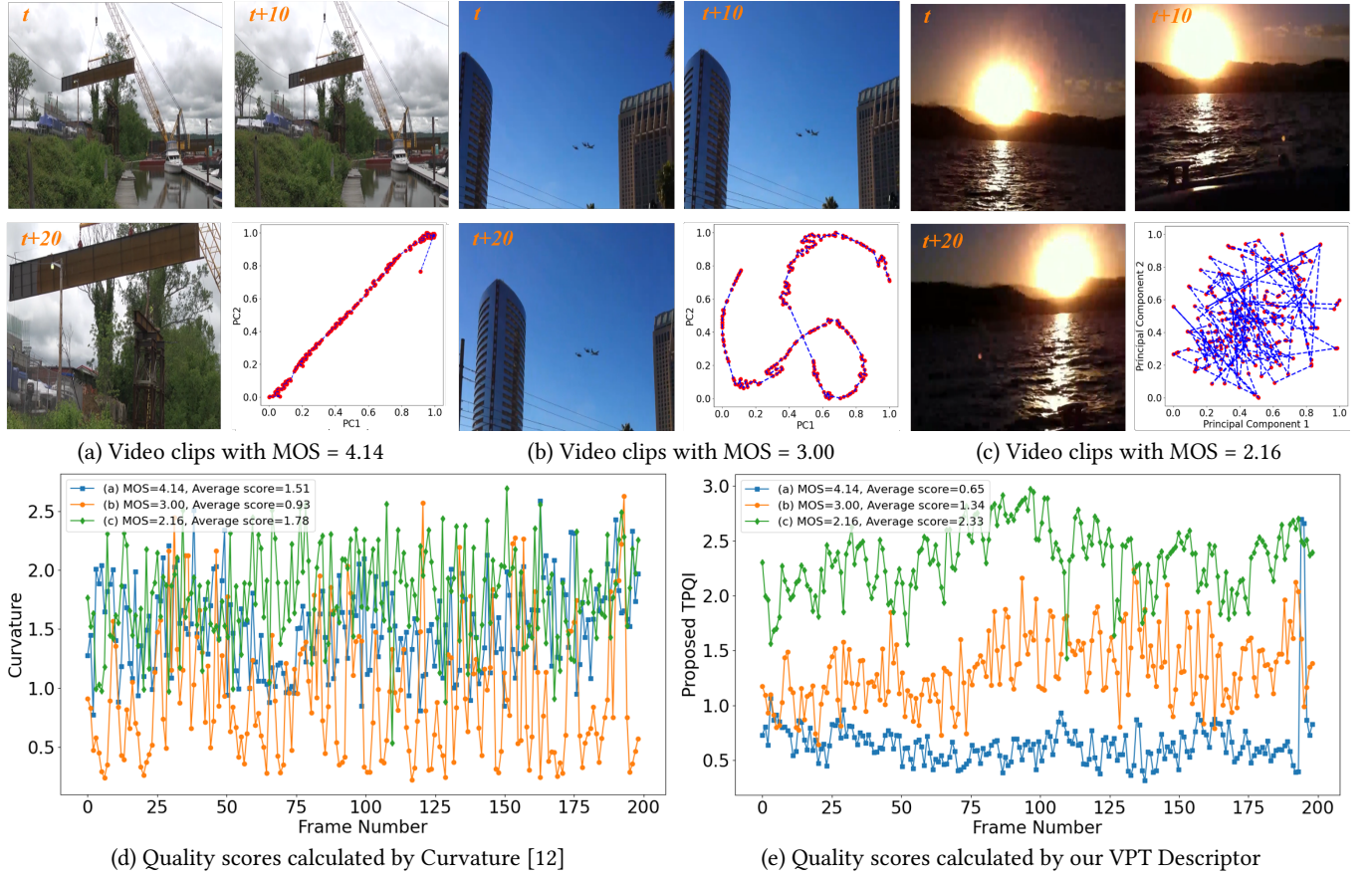


Figure 1: Examples of three video clips from KoNViD-1k dataset. The chart next to the frames is visualized temporal trajectory (the two principal components) of all frames in the LGN domain. From (a) to (c), the temporal trajectory changes from nearly straight for high quality to curves for medium quality and to fragments for low quality. (d) and (e) show the distribution of quality scores measured by the curvature [12] and the proposed TPQI. In (d), the curves are mixed together and the average scores of the curvature do not have clearly relation with the MOS values; whilst in (e), the proposed TPQI scores are better separated for different qualities and inversely related to the MOS values.

by standout performance on a wide variety of computer vision tasks, deep learning-based VQA models [3, 4, 19, 24, 50, 57] have been proposed to extract content-aware and distortion-sensitive features to predict the quality of in-the-wild videos. However, most of them still integrate the frame-wise deep features by some pooling modules such as gated recurrent unit (GRU) [19] and long short-term memory (LSTM) [57], making their performance constrained without effective extraction of motion information for video quality perception. Although some work has been done on temporal quality modeling of video, research in this area is still in its infancy. The great challenge is that it is still unclear how humans perceive temporal distortion, especially for the in-the-wild videos.

Recent researches have reported the discovery on the straightness of the perceptual representation of natural videos [11, 12]. It demonstrated that HVS transforms the incoming natural video signals into more regular representations in the perceptual domains, which are aligned along straighter trajectories in time. Examples of temporal trajectories of video clips, which track the perceptual representation of each frame in HVS along time, are visualized in Fig. 1(a)-(c) (red point is the two principal components of the

representation and dashed blue line is the temporal trajectory). The temporal trajectory is close to a straight line for the video with a high subjective quality score, but degenerates to haphazard curves for videos with low subjective quality scores. However, although the discovery has been successfully used to discriminate the natural and unnatural videos, *i.e.* artificial and naturalistic sequences, the quality scores, predicted based on the curvature index induced in the temporal trajectories [12], do not have clear relation with the human subjective scores for VQA (Fig. 1(d)). We suspect that this may be due to only measuring the extent of straightness of the temporal trajectory is not enough for VQA.

In this paper, we make the hypothesis that video distortions that harm the perceived quality of the videos will result in curved representations in HVS and propose a generalized and completely blind **Temporal Perceptual Quality Index (TPQI)**, through measuring the graphical morphology of the temporal representations of the videos in HVS (Fig. 1(e)). Specifically, we first employ two computational models of HVS simulating the neural activity in the lateral geniculate nucleus (LGN) and primary visual area (V1) to transform the videos into their neural temporal trajectory representations.

Then, we design a video perceptual trajectory (VPT) descriptor to quantify the temporal distortions. Considering that the distortions affect both the orientation and the fragmentation of the temporal trajectory, the VPT descriptor integrates two morphology elements into the measurement, *i.e.* change of the direction and distance of the change. Experiments on various combinations of the elements are conducted to determine the final VPT descriptor.

We have evaluated the proposed TPQI on four popular in-the-wild video datasets and the results demonstrate that the proposed TPQI achieves almost the same performance as the well-established spatial quality index, NIQE [33], indicating that the temporal quality index can be the same effective as the spatial quality index for VQA. Moreover, the representations of HVS in TPQI outperform the deep features from convolutional neural networks (CNNs), including Alexnet [16], VGG [39], and Resnet [9], which are considered as candidate models for biological vision, showing that the VPT descriptor can well represent the neural perception of the temporal distortions. By integrating with NIQE, we can achieve state-of-the-art performance in the area of completely blind VQA, and even better than some of the opinion-aware blind VQA methods. The main contributions of our paper can be summarized in threefold:

- We exploit to characterize the complex temporal distortions of in-the-wild videos in the perceptual domain and demonstrate that the graphical morphology of the temporal trajectory representation can be treated as the indication of temporal distortion.
- We propose a generalized and completely blind Temporal Perceptual Quality Index (TPQI) to measure the perceived temporal quality of video data by quantifying the loss of straightness and compactness of the temporal trajectory with a newly designed video perceptual trajectory descriptor.
- We show for the first time that only the proposed TPQI can achieve VQA performance comparable to that of the spatial quality index and combining the spatial quality index and TPQI can achieve top performance over popular in-the-wild video datasets.

2 RELATED WORK

2.1 Video Quality Assessment

Early VQA methods were specifically designed for synthetic distortions (*e.g.*, Gaussian blur, compression, and transmission artifacts) based on statistical characteristics of the video, such as frame difference [37, 38], gradient [26], and optical flow [28, 51]. The most popular algorithms deploy perceptually relevant low-level features captured from natural bandpass statistical models. Typical approaches include Video-BLIINDS [38], which uses a combination of temporal features from block-based motion estimation and DCT coefficients computed from frame differences, and spatial features from NIQE [33], GM-LOG [54], which computes the joint statistics of gradient magnitude and Laplacian of Gaussian responses in the spatial domain, DESIQUE [61] in log-derivative and log-Gabor domains, and HIGRADE [17] in LAB color-transformed gradient domain. These methods estimate the deviations in the statistical distribution as a perceptual quality metric and have achieved good performance in quality assessment of synthetic distorted videos.

However, their performance decreases significantly when applied to in-the-wild videos containing multiple unknown distortions.

Quality assessment of in-the-wild videos has attracted great attention given its potential broader practical utility. Attempting to capture the unknown and highly diverse distortions as possible as they can, recently proposed models used dozens of such perceptually relevant features and achieved state-of-the-art performance on existing in-the-wild datasets. For example, VIDEVAL [44] is a bag of features-based blind VQA model on **KoNViD-1k** and **YouTubUGC**, which uses a feature ensemble and selection procedure on top of existing efficient blind VQA models. RAPIQUE [45] combines and exploits the advantages of both quality-aware scene statistics features and semantics-aware deep convolutional features, designing a general and efficient spatial and temporal bandpass statistical model for VQA. Instead of extracting handcraft features, deep VQA methods [4, 19, 48, 49, 56, 58] use CNNs to extract rich semantic features and run regression on the extracted features to predict video quality. For example, MLSP-FF [8] extracts frame-wise features with Inception-ResNetv2 model [42] and some works [48, 56, 58] introduce 3D-CNN instead of 2D-CNN to extract more efficient temporal features. Along with more powerful feature extraction, deep VQA methods attempt to achieve the temporal-memory effect by some temporal regression modules, such as recurrent neural network (RNN) [4], GRU [19] and LSTM [48, 58].

2.2 Completely Blind Quality Assessment

Most IQA/VQA methods require a large number of distorted images or videos with human subjective scores to learn the quality regression model, which leads to a massive workload in collecting these annotations. More importantly, it is difficult to collect enough training samples to cover the numerous distortion types, resulting in a weak generalization of opinion-aware quality assessment.

A few works have been carried out on completely blind IQA, *i.e.* assessing quality without any additional information. Mittal *et al.* [31] first proposed a probabilistic latent semantic analysis of the statistical features of the pristine and distorted image patches and discovered latent quality factors that can infer a quality score from the test image patches. Later, they [33] proposed the Natural Image Quality Evaluator (NIQE), which infers the quality of a test image by measuring the distance between its Multivariate Gaussian (MVG) model learned from a set of local features from the image and the MVG model learned from the pristine natural images. Inspired by NIQE, Zhang *et al.* [59] replaced the local features with various NSS features and computed them from a collection of pristine image patches instead of the whole image. Liu *et al.* [23] proposed an unsupervised NR-IQA model based on the free energy principle to quantify the image quality in terms of structure, naturalness, and perceived quality changes during the degradation of test image.

As for the completely blind VQA, Mittal *et al.* [32] proposed Video Intrinsic Integrity and Distortion Evaluation Oracle (VIDEIO), which was built on an NSS model of consecutive frame differences and measured departures from the statistical regularities in natural videos. Kancharla *et al.* [13] assumed that the increase in the straightness of perceptual domain representation was positively related to MOS values and performed a linear prediction in the LGN domain to model temporal distortions by the prediction errors.

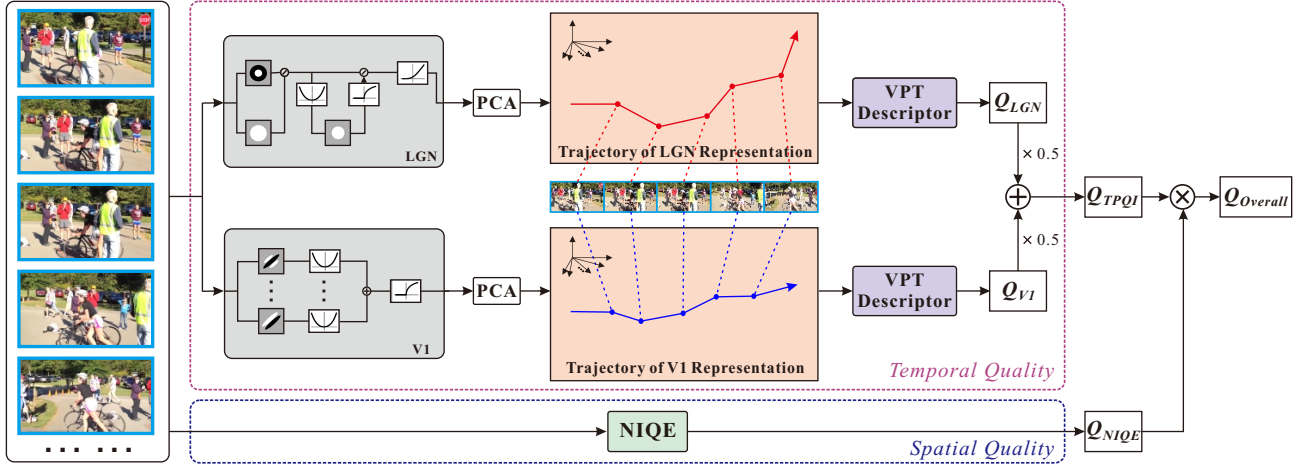


Figure 2: Framework of the proposed completely blind VQA method. Each individual video frame is first transformed and obtained the temporal trajectory in perceptual domain, i.e. LGN and V1 domains. A video perceptual trajectory descriptor is used to quantify temporal quality from both perceptual domains. The fusion of the temporal and spatial quality is used to predict the overall quality estimation.

3 THE PROPOSED METHOD

We propose a video quality assessment algorithm based on the perceptual domain representation of the video extracted using band-pass models of the visual system, aiming to explore the mapping from the temporal trajectory of a video to the subjective perception of temporal quality. The framework of the proposed method is shown in Fig. 2. We first transform each individual video frame into the perceptual domain, i.e. LGN and V1 responses in this work, and obtain the temporal trajectory of the video by arranging the perceptual features along time. Then, we propose a video perceptual trajectory descriptor (VPT) to jointly measure the straightness and compactness of the trajectory, with the former modeling the distortions from naturalness of video and the latter modeling the content continuity between video frames. The temporal perceptual quality index (TPQI) is the average score from the VPT descriptors on both the LGN and V1 features. Finally, we combine the proposed TPQI with a spatial quality metric to predict the overall video quality.

3.1 Perceptual Domain Representation of HVS

We first extract the perceptual domain representations simulating the responses of LGN and V1, which are vital regions of the HVS for visual information processing. Specifically, LGN performs luminance and contrast gain control, while V1 is known to be tuned to different orientations, scales, and frequencies. As stated in [12], the LGN representation likely straightens natural videos by providing robustness to local fluctuations in luminance and contrast, whereas the V1 representation provides straightening by its position- and phase-invariant properties. We use both representations to mimic the nonlinear functional properties of the early visual system.

3.1.1 Extracting LGN representation. The LGN model consists of center-surround filtering followed by local luminance and contrast gain control operations to simulate the primary nonlinear transformations performed by the retina and lateral geniculate nucleus [29]. In this work, we employ the LGN blocks as proposed by Laparra *et al.* [18]. In this model, the linear components for luminance subtraction are implemented using difference of Gaussian (DoG) filters

and a Laplacian pyramid. The non-linearity components perform contrast gain control to capture the local gain control property of the LGN neurons, which is achieved by performing a contrast normalization operation on the output of the linear bandpass filters. To relieve the tuning of hyper-parameters of the proposed model, we keep the settings of the LGN model as they are in [13].

3.1.2 Extracting V1 representation. The V1 model aims to transform the visual signal using a set of oriented filters whose responses are squared and combined over phases to capture the nonlinear behavior of complex cells in the primary visual cortex (V1) [2]. In this work, we adopt Gabor filters as the simulation model of V1, motivated by the HVS hypothesis that Gabor filters have a good approximation of the response of V1 [6]. Specifically, in the spatial domain, the 2-D Gabor filter is a Gaussian kernel function modulated by a complex sinusoidal plane wave, defined as:

$$g_{\theta}(x, y) = \frac{f^2}{\pi\gamma\eta} \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp(j2\pi f x' + \phi), \quad (1)$$

$$x' = x \cos \theta + y \sin \theta,$$

$$y' = -x \sin \theta + y \cos \theta,$$

where f is the frequency of the sine wave, θ represents the orientation of the normal to the parallel stripes of the Gabor function, ϕ is phase offset, σ is the standard deviation of the Gaussian envelope, and γ is the spatial aspect ratio specifying the ellipticity of the support of the Gabor function. The Gabor filters are then used to convolve with each video frame, and the features from all Gabor filters are concatenated as the V1 representation of this frame.

3.1.3 Dimensionality reduction of the perceptual domain representations. To better understand the perceptual activities, many recent studies [7, 25, 35] have used dimensionality reduction techniques to transform the high-dimensional neural data into low-dimensional subspaces where the underlying manifolds are topologically simple. In this work, we apply principal component analysis (PCA) to reduce the representations of LGN and V1 to low-dimensional features of dimension d .

3.2 Temporal Perceptual Quality Index

3.2.1 Motivation. Taking the temporal trajectories from LGN and V1 representations, we attempt to quantify their straightness and compactness losses to predict the temporal quality. A recent computational neuroscience model [12] defines the curvature of the trajectory as the average of the unsigned angles between difference vectors of successive frames to represent the straightness loss (Fig. 3(a)). Although it is able to distinguish between natural and unnatural videos, it fails to map the curvature to the subjective quality scores (Fig. 1(d)), probably because it does not quantify how far away a new frame deviates from the straight line, especially in the case of a large gap between two frames. Another alternative proposed in [13] attempts to measure temporal distortion by calculating the distance between the predicted values from a linear model fit of the perceptual representations and the true representation (Fig. 3(b)). It has taken the deviation distance from the trajectory into account, but ignores the variation between successive frames.

In this paper, we propose to measure both the straightness and the compactness of the trajectory. The former measures the angular change between the difference vectors of successive frames, while the latter measures the degree of deviation between these frames (Fig. 3(c)). In this way, if there is no directional change between two vectors, the magnitude of the vectors does not matter, but if there is, the distance will increase the penalty for the angular changes.

3.2.2 Video perceptual trajectory descriptor. In this work, we adopt three nearby frames as a temporal trajectory unit as in [12] and integrate two types of morphology change elements, including change of the direction and distance of the change to measure temporal distortion degree and thereby predict the temporal quality.

Direction change: The curvature is used to measure the direction of the change, which is defined as the angle between the nearby difference vectors. Specifically, let x_{i-1} , x_i and x_{i+1} be the perceptual representations of three consecutive frames, which are located in a d -dimensional space after dimensionality reduction, the curvature can be calculated as:

$$\overrightarrow{x_{i-1}x_i} = x_i - x_{i-1}, \quad (2)$$

$$\overrightarrow{x_i x_{i+1}} = x_{i+1} - x_i, \quad (3)$$

$$\theta_i = \arccos\left(\frac{\overrightarrow{x_{i-1}x_i} \cdot \overrightarrow{x_i x_{i+1}}}{\|\overrightarrow{x_{i-1}x_i}\| \|\overrightarrow{x_i x_{i+1}}\|}\right), \quad (4)$$

where $\overrightarrow{x_{i-1}x_i}$ and $\overrightarrow{x_i x_{i+1}}$ denote the two difference vectors of the three consecutive frames. \cdot denotes to vector dot product and θ_i is the curvature in radians.

Distance change: We use the magnitude of sum of the two difference vectors to measure the distance of the change. Specifically, the distance can be calculated as:

$$S_i = \|\overrightarrow{x_{i-1}x_i} + \overrightarrow{x_i x_{i+1}}\| = \|x_{i+1} - x_{i-1}\|. \quad (5)$$

VPT descriptor: Assuming both the extent of the change direction and distance are negatively correlated to the temporal quality, *i.e.* the larger the change in direction or the longer the change in distance, the greater temporal distortion and thus the lower temporal quality, we define a generic representation of these two morphology change elements to measure the temporal distortion at each time instant i , given by:

$$Q_i = f(\theta_i, S_i), \quad (6)$$

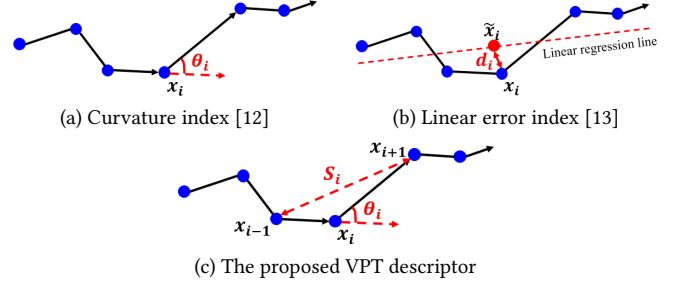


Figure 3: Comparison of three kinds of trajectory descriptor for measuring temporal distortion.

where $f(\cdot)$ denotes a pairwise function of θ_i and S_i . In this work, we use $f(\theta_i, S_i) = \theta_i \times \sqrt{S_i}$ after experimenting with various choices.

3.2.3 Calculating the score of TPQI. To estimate the video-level temporal quality, the temporal averages of Q_i in the LGN and V1 domains are calculated and followed by logarithmic compression as in [13] separately, and then the temporal quality score of the whole video is obtained from the average of the two domains.

$$Q_{TPQI} = \frac{\log(\frac{1}{N-2} \sum_{i=2}^{N-1} Q_i^{LGN}) + \log(\frac{1}{N-2} \sum_{i=2}^{N-1} Q_i^{V1})}{2}, \quad (7)$$

where N is the total number of video frames, and Q_{TPQI} is calculated from frame 2 to frame $N - 1$.

3.3 Natural Video Quality Evaluator

3.3.1 Spatial quality estimation. It is commonly acknowledged that frame-level spatial quality plays a very important role in estimating the overall video quality in the VQA problem. In our work, we employ the well-established blind NR-IQA algorithm, NIQE [33], for spatial quality estimation. The overall spatial quality of a video is the average of the frame-level spatial quality scores, given by:

$$Q_{NIQE} = \frac{1}{N} \sum_{i=1}^N q_i, \quad (8)$$

where q_i denotes to the spatial quality score of i -th frame and N is the total frame number of the video.

3.3.2 Overall video quality estimation. The overall video quality estimate is the fusion of spatial and temporal quality given by:

$$Q_{overall} = \text{fusion}(Q_{NIQE}, Q_{TPQI}), \quad (9)$$

where $\text{fusion}(\cdot)$ denotes the fusion strategy. There are two fusion strategies of defining the overall quality index for a video, *i.e.* average or product of spatial and temporal quality, which both cause the index to respond to percentage changes in either spatial or temporal indices. We study the effectiveness of the two fusion strategies in the experimental section.

4 EXPERIMENTAL RESULTS

4.1 Experimental Settings

4.1.1 Datasets. We evaluate the effectiveness of the proposed method on four popular in-the-wild VQA datasets, including: **KoNViD-1k**, **LIVE-VQC**, **CVD2014**, and **YouTube-UGC**.

KoNViD-1k [10]: The dataset contains 1,200 videos and the resolution of all videos is 960×540 . The videos are 8 seconds in duration with frame rates of 24/25/30 frames per second (fps).

Table 1: Performance comparison on the four VQA datasets. TPQI (LGN), TPQI (V1) and TPQI (LGN+V1) are the three TPQI variants adopting temporal representations from different perceptual domains. Overall (Sum) and Overall (Product) denote two different strategies of fusing spatial and temporal quality scores.

Category	Method	KoNViD-1k			LIVE_VQC			CVD2014			YouTube-UGC		
		SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓
Opinion-aware blind VQA	V-BLIINDS [38]	0.710	0.704	0.460	0.694	0.718	11.765	0.700	0.710	15.222	0.559	0.555	0.536
	TL-VQM [15]	0.780	0.770	0.406	0.799	0.803	10.145	0.830	0.850	11.330	0.669	0.659	0.485
	NSTSS [37]	0.625	0.639	-	-	-	-	0.615	0.653	-	-	-	-
	VIDEVAL [44]	0.783	0.780	0.403	0.752	0.751	11.100	-	-	-	0.779	0.773	0.405
Completely blind VQA	NIQE [33]	0.543	0.548	0.536	0.598	0.622	13.356	0.492	0.612	16.950	0.266	0.290	0.640
	VIIDEO [32]	-0.015	0.013	0.639	0.029	0.137	16.882	0.145	0.119	23.644	0.160	0.146	0.637
	STEM [13]	0.629	0.629	0.497	0.656	0.670	12.649	0.532	0.619	16.813	0.284	0.318	0.623
Completely blind VQA (Ours)	TPQI (LGN)	0.453	0.439	0.576	0.519	0.516	14.071	0.229	0.356	20.243	0.047	0.064	0.635
	TPQI (V1)	0.531	0.527	0.545	0.596	0.618	13.276	0.408	0.469	19.057	0.269	0.311	0.617
	TPQI (LGN+V1)	0.556	0.549	0.541	0.636	0.645	12.907	0.413	0.464	19.107	0.111	0.218	0.626
	Overall (Sum)	0.660	0.659	0.482	0.708	0.721	11.714	0.553	0.637	16.693	0.268	0.297	0.613
	Overall (Product)	0.693	0.693	0.462	0.718	0.730	11.550	0.524	0.597	17.368	0.230	0.288	0.615

LIVE-VQC [40]: The dataset contains 585 videos with more temporal variations than the **KoNViD-1k** dataset. The resolution of the videos ranges from 240P to 1080P, and the videos are 10 seconds in duration with frame rates ranging from 19 to 30 fps.

CVD2014 [36]: The dataset contains 234 videos and the resolutions of the videos are 480P and 720P. The duration and frame rates range from 10 to 25 seconds and 11 to 31 fps, respectively.

YouTube-UGC [47]: The dataset has 1,131 videos with authentic distortions of 15 categories. The resolution of the videos varies from 360P to 4K, and all videos are 20 seconds in duration.

4.1.2 Baseline Methods. For comparison, we select three completely blind methods, *i.e.* NIQE [33], VIIDEO [32], and STEM [13]. We also compare our proposed method with four representative opinion-aware blind VQA algorithms, including V-BLIINDS [38], TLVQM [15], NSTSS [37], and VIDEVAL[44]. Unlike our proposed method that does not require any annotation, these opinion-aware algorithms need a training procedure that regresses various features extracted to the annotated MOS values. The numerical results of these baselines are presented from the literature [13, 37, 44].

4.1.3 Evaluation Metrics. We report three widely used metrics to evaluate the VQA performance, including Spearman’s rank correlation coefficient (SRCC), Pearson’s linear correlation coefficient (PLCC), and root mean square error (RMSE). SRCC and PLCC measure the correlation between predicted quality scores and labeled MOS values, and RMSE indicates the relative error. A better VQA method would result in higher SRCC and PLCC, but lower RMSE.

Considering the inconsistency of the scale between the predicted quality scores and the subjective scores, we perform the nonlinear mapping with a 4-parameter logistic function as suggested by VQEG [46]. The function is formulated as follows.

$$Q_{fit} = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp\left(\frac{-(Q_{pre} - \beta_3)}{\beta_4}\right)} \quad (10)$$

where Q_{pre} and Q_{fit} denote the predicted score and mapped score, respectively. β_1 , β_2 , β_3 and β_4 are the four fitting parameters of the logistic function.

Table 2: Numerical comparison on the performances of trajectory descriptors. Linear denotes the linear prediction error in [13]. The features used for comparison are from both LGN and V1.

	Descriptor	KoNViD-1k		LIVE_VQC		CVD2014	
		SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
Temporal	Linear [13]	0.504	0.497	0.567	0.611	0.238	0.410
	Curvature	0.425	0.423	0.539	0.548	0.380	0.430
	Distance	0.411	0.403	0.460	0.505	0.310	0.315
	Ours	0.556	0.549	0.636	0.645	0.413	0.464
Overall	Linear [13]	0.642	0.644	0.662	0.686	0.394	0.450
	Curvature	0.401	0.429	0.660	0.677	0.479	0.515
	Distance	0.632	0.641	0.666	0.684	0.354	0.359
	Ours	0.693	0.693	0.718	0.730	0.524	0.597

4.1.4 Implementation details. The PCA dimension is set to 10 and a spatial resolution of 480×270 for all videos to extract the V1 representations, which were both analyzed in the experiments. We employ 48 Gabor filters with 6 scales and 8 orientations and the size of the Gabor filters was set to 39×39 . The raw spatial resolution was used for calculating NIQE.

4.2 Performance Evaluation

We first compare the performance with the baselines on four datasets. The results are shown in Table 1 and analyzed in detail as follows.

Overall performance. The proposed TPQI delivers competitive performance over all completely blind baselines, and the overall performance of the combined TPQI and spatial metric NIQE achieves the best performance over three datasets. The overall performance is even better than some opinion-aware VQA baselines, *e.g.*, better than V-BLIINDS on **LIVE_VQC** dataset and better than NSTSS on **KoNViD-1k** and **CVD2014** datasets. The results show that the proposed TPQI does not require any dataset-specific information, and can be generalized to any video with natural settings.

Notice that the proposed method does not reach the best performance on **YouTube-UGC** dataset, and the performance degradation is in accordance with the results of NIQE, which is constructed based on the statistical regularities of natural images. The reason may be that this dataset contains many unnatural video categories,

Table 3: Numerical comparison on the performances of different choices of change distance in the trajectory unit. ($x_{i+1} \rightarrow \overrightarrow{x_{i-1}x_i}$ means the distance from the representation x_{i+1} to vector $\overrightarrow{x_{i-1}x_i}$).

Dataset	Domain	$\ \overrightarrow{x_{i-1}x_i}\ $		$\ \overrightarrow{x_ix_{i+1}}\ $		$\ \overrightarrow{x_{i-1}x_i}\ + \ \overrightarrow{x_ix_{i+1}}\ $		$\ \overrightarrow{x_{i-1}x_i} + \overrightarrow{x_ix_{i+1}}\ $		$x_{i+1} \rightarrow \overrightarrow{x_{i-1}x_i}$	
		SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow
KoNViD-1k	Temporal	0.544	0.535	0.546	0.537	0.550	0.541	0.556	0.549	0.554	0.545
	Overall	0.678	0.680	0.680	0.681	0.690	0.690	0.693	0.693	0.673	0.677
LIVE-VQC	Temporal	0.632	0.638	0.632	0.639	0.635	0.641	0.638	0.647	0.636	0.645
	Overall	0.710	0.722	0.711	0.723	0.717	0.730	0.718	0.730	0.697	0.714
CVD2014	Temporal	0.388	0.442	0.386	0.441	0.396	0.452	0.413	0.464	0.421	0.470
	Overall	0.522	0.584	0.522	0.584	0.544	0.609	0.524	0.597	0.533	0.592
YouTube-UGC	Temporal	0.114	0.216	0.115	0.217	0.115	0.218	0.111	0.218	0.117	0.216
	Overall	0.211	0.291	0.212	0.292	0.236	0.298	0.230	0.288	0.211	0.265

Table 4: Numerical comparison of different video resolutions for V1 feature on KoNViD-1k. The time unit is second/frame on CPU.

Video Resolution (Downsam. rate)	TPQI (V1)			Overall		
	SRCC \uparrow	PLCC \uparrow	Time	SRCC \uparrow	PLCC \uparrow	Time
960 × 540 (1)	0.522	0.518	0.928	0.695	0.694	1.234
480 × 270 (1/2)	0.531	0.527	0.254	0.693	0.693	0.334
240 × 135 (1/4)	0.521	0.520	0.067	0.680	0.682	0.095
120 × 67 (1/8)	0.501	0.501	0.033	0.662	0.665	0.044

such as *Animation*, *Gaming*, and *Lyric Video*, which are subjected to artificial processing and do not conform to the straightness hypothesis in the perceptual domain.

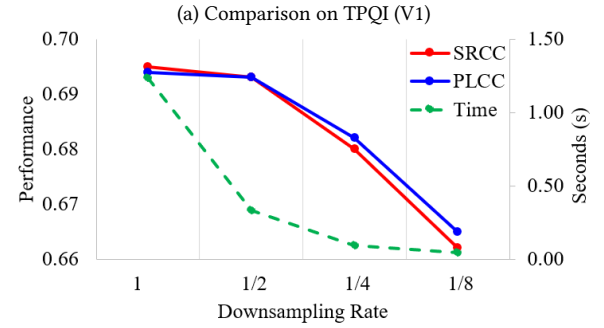
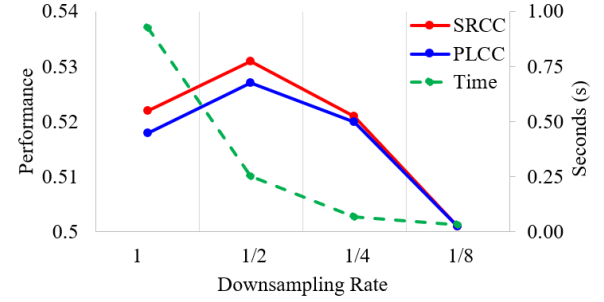
Employing features from different perceptual domains. The comparison among TPQI (LGN), TPQI (V1) and TPQI (LGN+V1) studies the effectiveness of the perceptual representations from different domains on measuring the temporal quality. The performance increases from the LGN domain to the V1 domain with the depth of the visual system, and the linear combination of LGN and V1 features can further boost the performance. The results show that both features play important roles in predicting subjective scores, and these features can also compensate for each other. They also show that our proposed TPQI, which measures only the temporal quality, can achieve better VQA performance than NIQE, especially on **LIVE_VQC**, which includes large camera motions.

Fusion of TPQI with spatial quality metric. We conduct comparative experiments on the fusion strategies, *i.e.* summation or multiplication, of the scores from the proposed temporal index (TPQI) and the spatial quality metric (NIQE). It can be observed that the product of the spatial and temporal scores leads to higher accuracy for **KoNViD-1k** and **LIVE_VQC**, which is probably contributed by the relative insensitivity of our indices to the range of values occupied by the spatial and temporal indices. Thus we adopt the product strategy in the ablation study to obtain the overall quality predictions.

4.3 Ablation Study

4.3.1 Design of the VPT descriptor. We propose to describe the loss in the straightness and compactness of perceptual domain representation by two components, namely the curvature representing how the representation deviates from the straight line and the distance representing how fast it deviates over a certain time interval.

Effectiveness of curvature and distance components. To investigate the validity of these two components, we conduct an ablation study on three variants: a) curvature only; b) distance

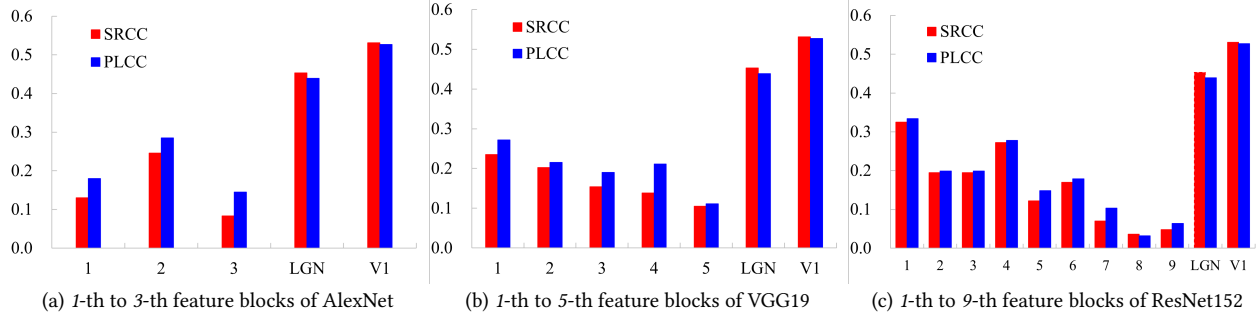
**Figure 4: Performance comparison of different video resolutions for V1 feature extraction on KoNViD-1k.**

only; c) a combination of curvature and distance (ours). To make the comparison more comprehensive, we also include d) the linear prediction error for temporal modeling in [13]. The numerical comparisons are shown in Table. 2. As stated in [12] that the curvature of the perceptual representation is vital to discriminating between natural and artificial videos, and it is also more effective than the distance for assessing temporal quality. But taking both curvature and distance results in better performance than just utilizing a single component, showing that the distance may show the intensity of temporal distortion, which also accounts for the low subjective score. The proposed descriptor also achieves better performance than the linear model, indicating that the proposed descriptor can better measure the temporal distortions in the perceptual domain.

Options for distance components. To make a better descriptor, we have tested the possible distance measurement options from the perceptual representation. Table. 3 shows all the tested options and their performance. In general, the performance of option measuring the compactness of the trajectory by the magnitude of sum of the two difference vectors S_i , *i.e.* $\|\overrightarrow{x_{i-1}x_i} + \overrightarrow{x_ix_{i+1}}\|$, is better

Table 5: Numerical comparison on the performances of different dimensions d of the representation in V1 domain.

Dataset	Domain	$d=5$		$d=10$		$d=30$		$d=50$		$d=80$	
		SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow	SRCC \uparrow	PLCC \uparrow
KoNViD-1k	Temporal	0.530	0.527	0.531	0.527	0.524	0.519	0.506	0.503	0.473	0.475
	Overall	0.692	0.694	0.693	0.693	0.685	0.685	0.677	0.678	0.668	0.667
LIVE-VQC	Temporal	0.595	0.617	0.596	0.618	0.596	0.609	0.583	0.597	0.544	0.565
	Overall	0.714	0.726	0.718	0.730	0.710	0.726	0.699	0.717	0.686	0.705
CVD2014	Temporal	0.400	0.467	0.408	0.469	0.450	0.504	0.470	0.534	0.452	0.505
	Overall	0.502	0.569	0.524	0.597	0.521	0.585	0.511	0.582	0.492	0.577
YouTube-UGC	Temporal	0.281	0.326	0.269	0.311	0.277	0.319	0.261	0.312	0.248	0.311
	Overall	0.241	0.305	0.228	0.287	0.208	0.258	0.198	0.235	0.198	0.235

**Figure 5: Performance comparison between bio-inspired features, i.e. features extracted by computational models of LGN and V1, and deep features from different layers of three classical CNN models, which have been well-trained on ImageNet [5] for classification tasks.**

than the other options. According to these results, we use it for the distance in the TPQI algorithm.

4.3.2 The impact of various settings on V1 feature. The representation in V1 domain has important contribution on the performance of the proposed TPQI, so that we test various settings including the resolutions for feature extraction and feature dimensionality reduction of the presentation in V1 domain. The experiments are conducted on two models: 1) TPQI (V1) model to eliminate effects from the LGN feature and the spatial quality, and 2) Overall model to check its effects on the final proposed VQA algorithm.

Resolutions for extracting V1 feature. We first test various downsampled video resolutions for V1 feature extraction since temporal modeling may not require a high spatial resolution. We conduct the test on the **KoNViD-1k** dataset with a unified raw resolution of 960×540 , and apply the conclusion to other datasets. The results are presented in Table. 4 and Fig. 4, respectively. For TPQI (V1) model, the best performance is reached at downsampling rate of $1/2$ (270P); while for the Overall model, the performance also almost reaches saturation at this resolution. Considering that the computational complexity decreases exponentially with the resolution downsampling, we chose to use 270P for the resolution of the input videos to extract the representation in V1 domain.

Dimension of V1 feature. The representation in V1 domain is the feature after dimensionality reduction of the original V1 feature map. We perform an extensive study on the parameter d for the feature dimensionality reduction, and the results are reported in Table 5. The best results are mostly achieved at $d = 10$, and lower or higher dimensions will cause the degradation of the performance. Therefore, we set $d = 10$ for the dimension of the representation in V1 domain in the proposed VQA algorithm.

4.3.3 Bio-inspired handcrafted feature v.s. deep feature. As being stated that the convolutional neural networks (CNNs) have shown impressive ability in object recognition and been proposed as candidate models for biological vision [14, 55], we compare the straightening capabilities of their features with the proposed biological LGN and V1 features for VQA. The experiment is conducted by replacing the LGN and V1 features with the features extracted in each stage of the classical CNN models, including Alexnet [16], VGG19 [39], and Resnet152 [9]. The results in Fig. 5 show that the proposed biological features can outperform those deep features extracted from the CNN models in temporal quality perception, which motivate us to address the related vision problems, such as image restoration [21, 22], action recognition [62, 63], and video compression [52, 53] considering the characteristics of the HVS.

5 CONCLUSIONS

In summary, we have applied the perceptual straightening hypothesis of the HVS to design a blind temporal quality prediction algorithm called TPQI. We demonstrate the efficacy of TPQI by its superior performance over a number of in-the-wild datasets. The performance of TPQI is noteworthy since it even surpasses supervised VQA algorithms on related datasets. Importantly, temporal consistency checks introduced by this hypothesis play a key role in achieving performance gains in video quality prediction. The proposed TPQI algorithm is explainable and generalizes well over a variety of in-the-wild datasets.

ACKNOWLEDGMENTS

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

REFERENCES

- [1] 99Content. 2021. *Facebook Video Statistics*. Retrieved March 23, 2022 from <https://99firms.com/blog/facebook-video-statistics>
- [2] Edward H Adelson and James R Bergen. 1985. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* 2, 2 (1985), 284–299.
- [3] Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. 2021. Learning Generalized Spatial-Temporal Deep Feature Representation for No-Reference Video Quality Assessment. *IEEE Trans. Circuits Syst. Video Technol.* (2021), 1–1. <https://doi.org/10.1109/TCSVT.2021.3088505>
- [4] Pengfei Chen, Leida Li, Lei Ma, Jinjian Wu, and Guangming Shi. 2020. RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment. In *Proc. ACM Int. Conf. Multimedia*. 834–842.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 248–255.
- [6] David J. Field. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* 4, 12 (Dec 1987), 2379–2394.
- [7] Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. 2015. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences* 112, 44 (2015), 13455–13460.
- [8] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. 2021. KonVid-150k: A Dataset for No-Reference Video Quality Assessment of Videos in-the-Wild. *IEEE Access* 9 (2021), 72139–72160.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 770–778.
- [10] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz natural video database (KoNViD-1k). In *QoMEX*. 1–6.
- [11] Olivier J. Hénaff, Yoon Bai, Julie A. Charlton, Ian Nauhaus, Eero P. Simoncelli, and Robbe L. T. Goris. 2021. Primary visual cortex straightens natural video trajectories. *Nature Communications* 12 (2021), 5982.
- [12] Olivier J. Hénaff, Robbe L. T. Goris, and Eero P. Simoncelli. 2019. Perceptual straightening of natural videos. *Nature Neuroscience* 22 (2019), 984–991.
- [13] Parimala Kancharla and Sumohana S. Channappayya. 2022. Completely Blind Quality Assessment of User Generated Video Content. *IEEE Trans. Image Process.* 31 (2022), 263–274. <https://doi.org/10.1109/TIP.2021.3130541>
- [14] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology* 10 (11 2014), 1–29.
- [15] Jari Korhonen. 2019. Two-Level Approach for No-Reference Consumer Video Quality Assessment. *IEEE Trans. Image Process.* 28, 12 (2019), 5923–5938.
- [16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. Adv. Neural Inf. Process.* 1106–1114.
- [17] Debarati Kundu, Deepti Ghadiyaram, Alan C. Bovik, and Brian L. Evans. 2017. No-Reference Quality Assessment of Tone-Mapped HDR Pictures. *IEEE Trans. Image Process.* 26, 6 (2017), 2957–2971.
- [18] Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P Simoncelli. 2016. Perceptual image quality assessment using a normalized Laplacian pyramid. *Electronic Imaging* 6, 0 (2016), 1–6.
- [19] Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality Assessment of In-the-Wild Videos. In *Proc. ACM Int. Conf. Multimedia*. 2351–2359.
- [20] Xuelong Li, Qun Guo, and Xiaoqiang Lu. 2016. Spatiotemporal Statistics for Video Quality Assessment. *IEEE Trans. Image Process.* 25, 7 (2016), 3329–3342.
- [21] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. 2020. Guidance and Evaluation: Semantic-Aware Image Inpainting for Mixed Scenes. In *Proc. Eur. Conf. Comput. Vis.* 683–700.
- [22] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. 2021. Image Inpainting Guided by Coherence Priors of Semantics and Textures. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 6539–6548.
- [23] Yutao Liu, Ke Gu, Yongbing Zhang, Xiu Li, Guangtao Zhai, Debin Zhao, and Wen Gao. 2020. Unsupervised Blind Image Quality Evaluation via Statistical Measurements of Structure, Naturalness, and Perception. *IEEE Trans. Circuits Syst. Video Technol.* 30, 4 (2020), 929–943.
- [24] Yongxu Liu, Jinjian Wu, Leida Li, Weisheng Dong, Jinpeng Zhang, and Guangming Shi. 2021. Spatiotemporal Representation Learning for Blind Video Quality Assessment. *IEEE Trans. Circuits Syst. Video Technol.* (2021), 1–1. <https://doi.org/10.1109/TCSVT.2021.3114509>
- [25] Ryan J. Low, Sam Lewallen, Dmitriy Aronov, Rhino Nevers, and David W. Tank. 2018. Probing variability in a cognitive map using manifold inference from neural dynamics. *bioRxiv* (2018).
- [26] Wen Lu, Ran He, Jiachen Yang, Changcheng Jia, and Xinbo Gao. 2019. A spatiotemporal model of video quality assessment via 3D gradient differencing. *Inf. Sci.* 478 (2019), 141–151.
- [27] Jupo Ma, Jinjian Wu, Leida Li, Weisheng Dong, Xuemei Xie, Guangming Shi, and Weisi Lin. 2021. Blind Image Quality Assessment With Active Inference. *IEEE Trans. Image Process.* 30 (2021), 3650–3663.
- [28] K. Manasa and Sumohana S. Channappayya. 2016. An Optical Flow-Based Full Reference Video Quality Assessment Algorithm. *IEEE Trans. Image Process.* 25, 6 (2016), 2480–2492.
- [29] Valerio Mante, Vincent Bonin, and Matteo Carandini. 2008. Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron* 58, 4 (2008), 625–638.
- [30] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* 21, 12 (2012), 4695–4708.
- [31] Anish Mittal, Gautam S. Muralidhar, Joydeep Ghosh, and Alan C. Bovik. 2012. Blind Image Quality Assessment Without Human Training Using Latent Quality Factors. *IEEE Signal Process. Lett.* 19, 2 (2012), 75–78.
- [32] Anish Mittal, Michele A. Saad, and Alan C. Bovik. 2016. A Completely Blind Video Integrity Oracle. *IEEE Trans. Image Process.* 25, 1 (2016), 289–300.
- [33] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.* 20, 3 (2013), 209–212.
- [34] Maryam Mohsin. 2021. *10 Tiktok Statistics That You Need to Know in 2021*. Retrieved March 23, 2022 from <https://www.oberlo.com/blog/tiktok-statistics>
- [35] Edward H. Nieh, Manuel Schottdorf, Nicolas W. Freeman, Ryan J. Low, Sam Lewallen, Sue Ann Koay, Lucas Pinto, Jeffrey L. Gauthier, Carlos D. Brody, and David W. Tank. 2021. Geometry of abstract learned knowledge in the hippocampus. *Nature* 595 (2021), 80–84.
- [36] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oitinen, and Jukka Häkkinen. 2016. CVD2014—A Database for Evaluating No-Reference Video Quality Assessment Algorithms. *IEEE Trans. Image Process.* 25, 7 (2016), 3073–3086.
- [37] Sathya Veera Reddy Dendi and Sumohana S. Channappayya. 2020. No-Reference Video Quality Assessment Using Natural Spatiotemporal Scene Statistics. *IEEE Trans. Image Process.* 29 (2020), 5612–5624.
- [38] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. 2014. Blind Prediction of Natural Video Quality. *IEEE Trans. Image Process.* 23, 3 (2014), 1352–1365.
- [39] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.
- [40] Zeina Sinno and Alan Conrad Bovik. 2019. Large-Scale Study of Perceptual Video Quality. *IEEE Trans. Image Process.* 28, 2 (2019), 612–627.
- [41] Kit Smith. 2020. *57 Fascinating and Incredible YouTube Statistics*. Retrieved March 23, 2022 from <https://www.brandwatch.com/blog/youtube-stats/>
- [42] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proc. AAAI Conf. Artif. Intell.* 4278–4284.
- [43] Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2020. A Comparative Evaluation Of Temporal Pooling Methods For Blind Video Quality Assessment. In *Proc. IEEE Conf. Image Process.* 141–145.
- [44] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content. *IEEE Trans. Image Process.* 30 (2021), 4449–4464.
- [45] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. RAPIQUE: Rapid and Accurate Video Quality Prediction of User Generated Content. *IEEE Open J. Signal Process.* 2 (2021), 425–440.
- [46] VQEG. 2000. *Final report from the video quality experts group on the validation of objective models of video quality assessment*. Technical Report.
- [47] Yilin Wang, Sasi Inguva, and Balu Adsumilli. 2019. YouTube UGC Dataset for Video Compression Research. In *MMSP*. 1–5.
- [48] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. 2021. Rich Features for Perceptual Quality Assessment of UGC Videos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 13435–13444.
- [49] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. FAST-VQA: Efficient End-to-end Video Quality Assessment with Fragment Sampling. *arXiv preprint arXiv: 2207.02595* (2022).
- [50] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. DisCoVQA: Temporal Distortion-Content Transformers for Video Quality Assessment. *arXiv preprint arXiv: 2206.09853* (2022).
- [51] Jinjian Wu, Yongxu Liu, Weisheng Dong, Guangming Shi, and Weisi Lin. 2019. Quality Assessment for Video With Degradation Along Salient Trajectories. *IEEE Trans. Multimed.* 21, 11 (2019), 2738–2749.
- [52] Jing Xiao, Ruimin Hu, Liang Liao, Yu Chen, Zhongyuan Wang, and Zixiang Xiong. 2016. Knowledge-Based Coding of Objects for Multisource Surveillance Video Data. *IEEE Trans. Multimed.* 18, 9 (2016), 1691–1706.
- [53] Jing Xiao, Zhongyuan Wang, Yu Chen, Liang Liao, Jun Xiao, Gen Zhan, and Ruimin Hu. 2017. A sensitive object-oriented approach to big surveillance data compression for social security applications in smart cities. *Softw. Pract. Exp.* 47, 8 (2017), 1061–1080.

- [54] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C. Bovik, and Xiangchu Feng. 2014. Blind Image Quality Assessment Using Joint Statistics of Gradient Magnitude and Laplacian Features. *IEEE Trans. Image Process.* 23, 11 (2014), 4850–4862.
- [55] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* 111, 23 (2014), 8619–8624.
- [56] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. 2021. Patch-VQ: 'Patching Up' the Video Quality Problem. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 14019–14029.
- [57] Junyong You. 2021. Long Short-term Convolutional Transformer for No-Reference Video Quality Assessment. In *Proc. ACM Int. Conf. Multimedia.* 2112–2120.
- [58] Junyong You and Jari Korhonen. 2019. Deep Neural Networks for No-Reference Video Quality Assessment. In *Proc. IEEE Conf. Image Process.* 2349–2353.
- [59] Lin Zhang, Lei Zhang, and Alan C. Bovik. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Trans. Image Process.* 24, 8 (2015), 2579–2591.
- [60] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. 2021. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Trans. Image Process.* 30 (Mar. 2021), 3474–3486.
- [61] Yi Zhang and Damon M. Chandler. 2013. No-reference image quality assessment based on log-derivative statistics of natural scenes. *J. Electronic Imaging* 22, 4 (2013), 043025.
- [62] Xian Zhong, Wenxin Huang, Ruiqi Luo, and Can Wang. 2020. Video Human Behavior Recognition Based on ISA Deep Network Model. *Int. J. Pattern Recognit. Artif. Intell.* 34, 14 (2020), 2056012:1–2056012:17.
- [63] Xian Zhong, Zhuo Zhou, Wenxuan Liu, Kui Jiang, Xuemei Jia, Wenxin Huang, and Zheng Wang. 2022. VCD: View-Constraint Disentanglement for Action Recognition. In *Proc. IEEE Conf. Acoust. Speech Signal Process.* 2170–2174.