No-Reference Video Quality Assessment Using Voxel-Wise fMRI Models of the Visual Cortex

Naga Sailaja Mahankali, *Senior Member, IEEE*, Mohan Raghavan, and Sumohana S. Channappayya, *Senior Member, IEEE*

***Abstract*—The performance of the human visual system is very efficient in many visual tasks such as identifying visual scenes, anticipating future actions based on the past observations, assessing the quality of visual stimuli, etc. A significant amount of effort has been directed towards finding quality aware representations of natural videos to solve the quality prediction task. In this work we present a novel no reference video quality assessment (NR-VQA) algorithm based on the functional Magnetic Resonance Imaging (fMRI) Blood Oxygen Level Dependent (BOLD) signal prediction with voxel-wise encoding models of the human brain. The voxel encoding models are learnt using deep features extracted from the AlexNet model to predict the fMRI response to natural video stimuli. We show that the curvature in the predicted voxel response time series provides good quality discriminability, and forms an important feature for quality prediction. Further, we show that the proposed curvature features in combination with the spatial index, temporal index and NIQE features deliver acceptable performance on the Video Quality Assessment (VQA) task on both synthetic and authentic distortion data-sets.**

***Index Terms*—Human visual system (HVS), functional magnetic resonance imaging (fMRI), blood oxygen level-dependent (BOLD), haemodynamic response function (HRF), video quality assessment (VQA).**

I. Introduction

THE evolution of cognitive research has uncovered major aspects of the human brain including its functionalities and architecture. Visual information processing is done in the visual cortex situated in the occipital lobe of the human brain. The human visual cortex is a hierarchically organized structure with several areas connected with both forward and backward connections. The information processing in the visual cortex is implemented in two paths [1]. One is the 'ventral stream' where object representation and form recognition tasks are carried out. The 'dorsal stream' pathway is said to perform motion and representation of object locations tasks. In this work we leverage the advances made in understanding the cortical responses using fMRI signals by applying it successfully to the perceptual quality assessment task.

Typically, perceptual quality assessment is performed using psycho-visual experiments conducted on human subjects. While these experiments provide the most accurate quality predictions, they are expensive and time-consuming. Objective quality assessment (QA) algorithms have been proposed to overcome these drawbacks while offering competitive performance. Objective QA algorithms come in a variety of flavours ranging from those based on models of the visual system [2] to those based on deep learning methods [3]. In this work, we propose an NR-VQA algorithm based on a model that predicts the fMRI responses of cortical regions to visual stimulus. Recently, methods have been proposed to assess perceptual image/video quality using electroencephalography (EEG) [4]–[6] signals. Modeling the voxels in the human visual cortex using fMRI BOLD signals is being studied extensively for the reconstruction of the perceived content from the brain activity [7]–[13]. We make the following contributions in this work: demonstrate the sensitivity of fMRI signals (predicted by a model) to perceptual quality, and their application to an NR-VQA algorithm that is shown to be competitive over synthetic and authentic distortions.

II. Related Work

We briefly review relevant literature of voxel models on the visual cortex, and NR-VQA algorithms in the following.

A. Voxel Models of Visual Cortex

Visual information perceived by the human eye passes through the lateral geniculate nucleus (LGN) in the thalamus and then reaches the primary visual cortex (V1) which is the first layer of the HVS. From the V1 region, the optical information is forwarded to the other major areas such as V2, V3, V4, Inferior Temporal region (IT), V5 or Middle Temporal region (MT), etc. of the visual cortex. Object detection occurs in the IT [14] and motion detection in the MT [15] areas of the human visual cortex. The pioneering work of Hubel and Wiesel [16] described the primary visual cortex as a hierarchical model composed of 'simple' and 'complex' cells. Rust *et al.* [17] and Simoncelli *et al.* [18] proposed linear-nonlinear computational models to mimic the neural responses in V1 and MT regions. fMRI has been extensively used to understand the functioning of the cortical regions in the HVS [8]. There are a variety of

Manuscript received October 15, 2021; revised November 30, 2021; accepted December 8, 2021. Date of publication December 17, 2021; date of current version January 28, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Victor Sanchez. *(Corresponding author: Naga Sailaja Mahankali.)*

Naga Sailaja Mahankali and Sumohana S. Channappayya are with the Department of Electrical Engineering, Indian Institute of Technology Hyderabad, Hyderabad, Telangana 500020, India (e-mail: ee16resch11004@iith.ac.in; sumohana@iith.ac.in).

Mohan Raghavan is with the Department of Biomedical Engineering, Indian Institute of Technology Hyderabad, Hyderabad, Telangana 500020, India (e-mail: mohanr@bme.iith.ac.in).

Digital Object Identifier 10.1109/LSP.2021.3136487

approaches in which voxel-wise models are built using fMRI signals. In [7], [8] Gabor filters are used for encoding the voxel responses. The authors of [9], [19] have shown an analogy between the layers of convolutional neural networks (CNNs) to the layers of the human visual cortex. Many advanced architectures such as auto-encoders [10] and generative adversarial networks (GANs) [11]–[13] are also used for encoding the voxel models from fMRI signals. Identification of the visual stimulus and its reconstruction from the cortical responses are two widely studied functions that are performed using these voxel models. In this work, we employ these voxel encoding models for video quality prediction.

B. NR-VQA Metrics

The human visual system (HVS) has the amazing capability of assessing the perceptual quality of a given video even in the absence of any reference. The challenge for NR-VQA algorithms is to be able to mimic this capability of the HVS. Many NR-VQA metrics have been proposed in the literature [2], [3], [20]–[25]. A simple method to address the VQA problem is by applying image quality assessment (IQA) metrics to each frame of the video and then pool the frame level quality scores to arrive at the video level quality score. NIQE [20], BRISQUE [21], FRIQUEE [22], are some of the no reference image quality assessment (NR-IQA) metrics which are used for VQA. These techniques are purely dependent on the spatial information. Since video is a spatio-temporal signal, the distortions involve both spatial and temporal artifacts. However, relying on spatial quality alone for NR-VQA has been shown to deliver sub-par performance [23]. The algorithms in [3], [23]–[28] are some of the NR-VQA metrics which implement both the spatial and temporal features for predicting video quality. V-BLIINDS [23], VIIDEO [24] and RAPIQUE [3] use various combinations of natural scene statistics (NSS)-based features to predict the quality score. In TLVQM [25] a large number of distortion-specific features are designed and combined to achieve a quality prediction score. In [2], 3-D MSCN features in combination with spatio-temporal Gabor features are used to formulate a quality metric. The proposed algorithm differs fundamentally from most NR-VQA algorithms in the literature in that we employ key features derived from a model that predicts physiological activity (i.e., blood oxygen level) in the HVS.

III. Proposed Approach

Our work draws inspiration from the perceptual straightening hypothesis proposed by Hénaff *et al.* [29], which states that the human visual system transforms complex visual inputs into perceptual representations which have straighter temporal trajectories. Further, motivated by the EEG based quality assessment [4]–[6], HVS inspired VQA algorithms and the fact that the voxel responses are slowly varying signals, we hypothesize that: (i) voxel responses in the visual cortex will vary with distortions in a video, (ii) the distortions in a video that affect its naturalness will in turn affect the smoothness of the voxel responses. Our proposed NR-VQA algorithm is built on these hypotheses.

A. Voxel Encoding Models

We employ CNN-based deep learning voxel-wise encoding models to predict the BOLD signal response to natural video stimulus. For encoding the voxel models we use the Vim-2 fMRI dataset developed by Nishimoto *et al.* [8] for natural video stimuli (of resolution $128 \times 128 \times 3$). For building the voxel-wise encoding models we follow the procedure proposed in [9]. The video frames of the training stimulus of Vim-2 dataset [8] are presented as input to an off the shelf AlexNet [30] model pre-trained on the imagenet data-set, which accepts an input of size 227×227. To handle the mismatch in input resolution, the frames are cropped to a resolution of 119×119 and the stride is changed from 4 to 2 for the kernels of the first layer of the AlexNet. Feature maps are extracted from all the convolutional and fully connected layers of the CNN. Principal component analysis (PCA) is applied to reduce the feature dimension such that 99% of the variance of feature space is preserved. The feature time series is convolved with a canonical Haemodynamic Response Function (HRF) to project the fast visual stimuli to slow changing BOLD signal space. Down-sampling at the video frame rate is applied to the feature time series to match the time resolution of fMRI. A linear regression model with L2-regularization is fitted using this feature time series, for each voxel at every layer of the CNN to predict the fMRI signal. Layer index and the regularization parameter are optimized for each voxel with a nine-fold cross-validation technique. The Vim-2 dataset contains nearly 10,000 voxel responses from the visual cortex for each of the three subjects. To reduce the computational complexity, a subset of 1500 voxels are selected in each subject. For this purpose, we consider the voxels with high correlation values for a sample population of natural video stimuli from the validation set of Vim-2. This is performed based on our hypothesis that a subset of voxels in the visual cortex that exhibit good correlation with the ground truth over a wide variety of video inputs are a good representative subset of the visual cortex. The process is explained in detail in our previous work [31].

B. Video Quality Prediction

Inspired by the perceptual straightness hypothesis in [29], we in turn hypothesize that the curvature of the voxel response time series is sensitive to the perceptual quality of the video stimulus. We empirically verified this hypothesis with our voxel encoding models. To do so, we constructed an artificial video sequence of 65 frames in which we considered two frames of different scenes and recursively interpolated the middle frames with pixel average of the two frames. Thus we created a video having zero curvature (as defined in (3)) in the pixel domain. The time-series responses of the voxels of a particular region of interest (RoI) such as V1 or MT are vectorized and used for computing the perceptual curvature of natural and artificial video sequences. This is illustrated in Fig. 1. We now describe our feature extraction procedure. Let $\mathbf{x}_t$ represent the signal vector at time instant t. In our work this could either be the pixel intensity vector in the pixel domain or the voxel response vector from a set of voxels in a RoI in the perceptual domain. The first order difference between two successive signal vectors $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$ is

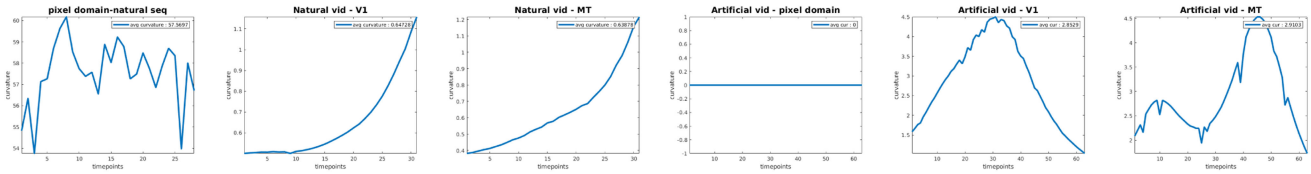

Fig. 1. Plots of the curvature of the frames of a natural video and an artificial video in pixel and perceptual (V1 and MT) domains. Three left images correspond to the natural video, and the remaining correspond to the artificial video. The smoothness of the perceptual domain curvature of a natural video and the relative loss of smoothness in the artificial video is clearly illustrated.

defined as the difference vector

$$\dot{\mathbf{x}}_t = \mathbf{x}_{t+1} - \mathbf{x}_t. \tag{1}$$

The normalized first order difference vector is given by

$$\hat{\dot{\mathbf{x}}}_t = \frac{\dot{\mathbf{x}}_t}{\|\dot{\mathbf{x}}_t\|}. \tag{2}$$

The second order difference is similarly defined as $\ddot{\mathbf{x}}_t = \dot{\mathbf{x}}_{t+1} - \dot{\mathbf{x}}_t$. The curvature is defined as the angle between the successive normalized difference vectors

$$c_t = \arccos\langle \hat{\dot{\mathbf{x}}}_t, \hat{\dot{\mathbf{x}}}_{t+1} \rangle. \tag{3}$$

The curvature is calculated between all the adjacent pairs of normalized difference vectors of a video (i.e., over all the frames) and the average is considered as the curvature of the video. As described previously, the curvature plots of the natural and artificial videos in both the pixel and the perceptual domains are shown in Fig. 1. We have considered the V1 and MT regions for the perceptual domain representation since they are two important areas of the HVS representing the low and mid-level cortical responses. It is evident from the plots that the trajectory of a natural video sequence is more curved in the pixel domain whereas the curvature is linear in the perceptual domain. For the artificial video sequence whose curvature is zero in the pixel domain, the perceptual curvature is highly non-linear. For predicting the quality of a video, the voxels (from the selected subset of voxels) belonging to each RoI are regrouped as separate vectors. Curvature is calculated for these vectors in 16 RoIs: V1, V2, V3, V3a, V3b, MT and lateral occipital area (latocc) of the visual cortex in both the right and left hemispheres. We considered only these regions, as they were consistent over all the three subjects in the Vim-2 data-set. This will form the curvature feature vector $\mathbf{c}_t$ of length 16. The average of the first difference vector $\bar{\dot{\mathbf{x}}}_t$ and the second difference vector $\bar{\ddot{\mathbf{x}}}_t$ over all the voxels of each of the 16 RoI are taken as a measure of perceptual straightness in addition to the curvature features, which together constitute our fMRI model-based features. In addition to these 48 features, we also use the average NIQE score denoted as $\bar{Q}_s$, an indicator of spatial quality, spatial and temporal indices (denoted by SI and TI respectively) [32] which measure the average spatial and temporal activity of a video as quality discerning features. In this work, we experiment with three different combinations of these features and evaluate their efficacy for the NR-VQA task. The first feature vector $\mathbf{f_1} = [\bar{\mathbf{c}}, \bar{\dot{\mathbf{x}}}, \bar{\ddot{\mathbf{x}}}]^t$ is composed purely of fMRI model-based components. The length of this feature vector is $16 \times 3 = 48$ where the 16 RoI identified earlier are chosen. Note that $\bar{\mathbf{c}}, \bar{\dot{\mathbf{x}}}, \bar{\ddot{\mathbf{x}}}$ are the time averages of $\mathbf{c}_t, \bar{\dot{\mathbf{x}}}_t, \bar{\ddot{\mathbf{x}}}_t$ respectively over the frames of the video. The second feature vector $\mathbf{f_2} = [\bar{\mathbf{c}}, \bar{\dot{\mathbf{x}}}, \bar{\ddot{\mathbf{x}}}, \bar{Q}_s, \text{SI}, \text{TI}]^t$ includes the average NIQE score as well as the SI and TI values, with $\bar{\mathbf{c}}, \bar{\dot{\mathbf{x}}}, \bar{\ddot{\mathbf{x}}}$ as defined for $\mathbf{f_1}$. The length of $\mathbf{f_2}$ is 51. The third feature vector $\mathbf{f_3} = [\bar{c}, \bar{\dot{x}}, \bar{\ddot{x}}, \bar{Q}_s, \text{SI}, \text{TI}]^t$ is of length 6. Here all the voxels (i.e., no specific RoI) are chosen resulting in one dimensional $\mathbf{c}_t$, $\bar{c}$, $\bar{\dot{x}}$ and $\bar{\ddot{x}}$. For quality prediction, we apply Support Vector Regression (SVR) with a radial basis function (RBF) kernel to map these features to the mean opinion score (MOS) of videos.

IV. Results and Discussion

Our algorithm is evaluated on diverse data-sets composed of both traditional/synthetic (EPFL [33], LIVE-SD [34], LIVE-MOBILE [35]) and authentic/in-capture (KoNViD [36], CVD2014 [37], LIVE-VQC [38]) distortions. The spatial resolution of the videos in these data-sets is different from each other and also from the expected input resolution of the AlexNet model. The frames of the videos are resized to 227×227 to match the input resolution of the voxel encoding model (AlexNet). We use the Pearson Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC) between subjective scores and the predicted objective scores to quantify the performance of the proposed no reference VQA technique. For obtaining the quality score, we trained an SVR with $80:20$ train-test split with the video-level spatial and temporal features and MOS scores as labels. The correlation results in Table I are the median scores over 100 trials. We report results for the fMRI model trained for each of the three subjects in the Vim-2 data-set. The performance of the average NIQE scores as a standalone metric is presented in the first row of the table.

From Table I we see that $\mathbf{f_1}$ shows competent performance over the synthetic distortion data-sets while it under-performs over authentic distortion data-sets. $\mathbf{f_2}$ performs consistently over the entire range of the data-sets. The quality prediction using $\mathbf{f_3}$ is better than the other two over all the authentic distortion datasets. Further, it performs well over the EPFL [33] data-set while slightly under-performing as compared to the other two combinations on the LIVE-MOBILE [35] data-set. In the case of LIVE-SD [34] this combination does not work well. To explain this behavior, the t-SNE visualizations of the $\mathbf{f_3}$ features for the distorted videos of the six data-sets with different levels of distortion are shown in Fig. 2. Except for LIVE-SD [34], the grouping of the videos with same range of quality scores is clearly illustrated over EPFL [33], LIVE-Mobile [35] and CVD2014 [37] data-sets. In the case of KoNViD-1K [36] and LIVE-VQC [38], only some ranges of quality scores are grouped together. The quality discriminability provided by $\mathbf{f_3}$ reflects in its performance over the considered data-sets. Further, these plots show that the six simple components of $\mathbf{f_3}$ provide an

TABLE I
PERFORMANCE EVALUATION OF PROPOSED APPROACH ON DATA-SETS WITH SYNTHETIC AND AUTHENTIC DISTORTIONS. THE ITALICISED SCORES ARE TAKEN FROM THE LITERATURE AND THE BOLD ONES ARE THE BEST SCORES

	EPFL [33]		LIVE-SD [34]		LIVE-MOBILE [35]		KoNViD [36]		CVD2014 [37]		LIVE-VQC [38]	
	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC	LCC	SROCC
NIQE [20]	0.516	0.499	0.370	0.378	0.364	0.34107	0.532	0.523	0.564	0.386	0.425	0.415
V-BLIINDS [23]	*0.752*	*0.807*	***0.881***	***0.759***	*0.437*	*0.439*	*0.565*	*0.572*	*0.71*	*0.70*	*0.694*	*0.718*
VIIDEO [24]	*0.184*	*0.205*	*0.651*	*0.624*	*0.245*	*0.216*	*0.280*	*0.275*	*0.341*	*0.239*	*0.274*	*0.173*
TLVQM [25]	*0.896*	***0.897***	*0.685*	*0.504*	*0.898*	*0.868*	*0.780*	*0.780*	*0.727*	*0.694*	***0.799***	***0.803***
RAPIQUE [3]	0.908	0.886	0.801	0.746	**0.951**	**0.937**	0.816	0.807	**0.776**	**0.745**	0.782	0.764
3D-MSCN + ST-Gabor [2]	***0.928***	*0.883*	*0.598*	*0.588*	*0.841*	*0.807*	*0.653*	*0.642*	*0.653*	*0.615*	-	-
SIONR [27]	-	-	-	-	-	-	***0.818***	***0.811***	-	-	*0.782*	*0.736*
$\mathbf{f_1}$–Subject-1	0.900	0.850	0.578	0.507	0.843	0.796	0.297	0.263	0.502	0.456	0.302	0.281
$\mathbf{f_1}$–Subject-2	0.881	0.826	0.616	0.584	0.855	0.798	0.201	0.175	0.434	0.389	0.342	0.305
$\mathbf{f_1}$–Subject-3	0.899	0.844	0.517	0.498	0.851	0.811	0.305	0.278	0.462	0.431	0.271	0.244
$\mathbf{f_2}$–Subject-1	0.924	0.869	0.604	0.568	0.866	0.830	0.568	0.552	0.625	0.598	0.538	0.520
$\mathbf{f_2}$–Subject-2	0.897	0.84	0.619	0.593	0.850	0.817	0.553	0.542	0.595	0.539	0.547	0.520
$\mathbf{f_2}$–Subject-3	0.913	0.865	0.531	0.502	0.851	0.826	0.572	0.556	0.583	0.537	0.505	0.500
$\mathbf{f_3}$–Subject-1	0.906	0.843	0.470	0.445	0.801	0.773	0.592	0.583	0.672	0.628	0.613	0.574
$\mathbf{f_3}$–Subject-2	0.905	0.827	0.469	0.446	0.850	0.817	0.582	0.572	0.667	0.615	0.612	0.573
$\mathbf{f_3}$–Subject-3	0.910	0.838	0.471	0.438	0.796	0.766	0.595	0.589	0.683	0.631	0.607	0.574

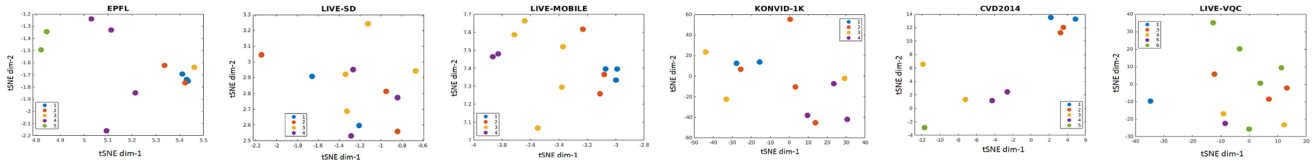

Fig. 2. t-SNE visualizations of $\mathbf{f_3}$ over the videos at varying qualities from the VQA data-sets considered in this work.

TABLE II
STATISTICAL HYPOTHESIS TESTING OF MODEL PERFORMANCE WITH RESPECT TO LCC / SROCC ON THE EPFL, LIVE-SD, LIVE-MOBILE, KONVID-1 K, CVD AND LIVE-VQC DATA-SETS (IN THAT ORDER FROM LEFT TO RIGHT)

	NIQE	TLVQM	RAPIQUE	$\mathbf{f_3}$
NIQE	—	000000 / 000000	000000 / 000000	000000 / 000000
TLVQM	111111 / 111111	—	000001 / 100001	010111 / 110111
RAPIQUE	111111 / 111111	111110 / 011110	—	011111 / 111111
$\mathbf{f_3}$	111111 / 111111	101000 / 001000	100000 / 000000	—

acceptable level of performance over synthetic and authentic distortions.We also observe from Table I that the scores are fairly consistent over all the three subjects. This shows that our subset of voxels is effective on the vastly different VQA data-sets chosen for performance evaluation. This observation illustrates the generalization capability of the voxel encoding model as well. We do note that the performance of the proposed NR-VQA algorithm while promising, ranks below the current state-of-the-art models like RAPIQUE [3] and TLVQM [25] as shown in Table II. Nevertheless, through our work we would like to illustrate the potential that models for brain activity (specifically like the fMRI model for BOLD signal prediction) hold for solving computer vision tasks. Further, the simplicity and explainability offered by such models is a useful by-product. We compare the computational complexity of different NR-VQA metrics using their Matlab implementations that are publicly available. All the algorithms are run on the same computer for the time complexity comparison. We used CIF, resolution videos from EPFL PoliMi [33] to check the running time. The average running time of 10 runs is reported on a per-frame basis in Table III. The hardware and software specifications of the computer are Intel(R) Core(TM) i7-4720HQ CPU @ 2.60 GHz, GM107 M [GeForce GTX 960M], 16 GB RAM and MATLAB R2018a. The proposed algorithm takes roughly 0.45 seconds to process one frame. The most time is spent on dimensionality reduction in the fMRI model.

TABLE III
PER-FRAME TIME COMPLEXITY AT CIF RESOLUTION

Algorithm	Time in sec
NIQE [20]	0.0673
VIIDEO [24]	0.1479
V-BLIINDS [23]	0.4324
TLVQM [25]	0.0809
RAPIQUE [3]	0.0329
3D-MSCN + ST-Gabor [2]	4.8634
$\mathbf{f_1}$	0.3607
$\mathbf{f_2}$	0.4581
$\mathbf{f_3}$	0.4477

V. CONCLUSIONS AND FUTURE WORK

We proposed a novel no-reference VQA metric based on the curvature smoothness of the fMRI response of visual cortex voxels to natural videos. The proposed algorithm delivers acceptable levels of performance on synthetic and authentic distortions. To the best of our knowledge, applying fMRI voxel models for the NR-VQA task has not been attempted previously. This novel metric is shown to be consistent across subjects on all the considered synthetic and authentic VQA data-sets. We believe that future work on fMRI VQA subjective studies will greatly help in building better and more efficient VQA metrics.

REFERENCES

[1] L. G. Ungerleider and M. Mishkin, "Two cortical visual systems," in *Analysis of Visual Behavior*, I. DJ, G. MA, and M. RJ, Eds. Cambridge, MA, USA: MIT Press, 1982, pp. 549–558.

[2] 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.

[3] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," *IEEE Open J. Signal Process.*, vol. 2, pp. 425–440, 2021.

[4] S. Scholler *et al.*, "Toward a direct measure of video quality perception using EEG," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2619–2629, May 2012.

[5] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "EEG correlates during video quality perception," in *Proc. 22nd Eur. Signal Process. Conf.*, 2014, pp. 2135–2139.

[6] S. Bosse *et al.*, "Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1694–1706, Aug. 2018.

[7] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, pp. 352–355, 2008.

[8] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Curr. Biol.*, vol. 21, no. 19, pp. 1641–1646, 2011.

[9] H. Wen, J. Shi, Y. Zhang, K. Lu, J. Cao, and Z. Liu, "Neural encoding and decoding with deep learning for dynamic natural vision," *Cereb Cortex*, vol. 20, pp. 1–25, 2017.

[10] R. Beliy, G. Gaziv, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 6517–6527.

[11] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. van Gerven, "Generative adversarial networks for reconstructing natural images from brain activity," *NeuroImage*, vol. 181, pp. 775–785, 2018.

[12] G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani, "End-to-end deep image reconstruction from human brain activity," *Front. Comput. Neurosci.*, vol. 13, p. 21, 2019.

[13] G. St-Yves and T. Naselaris, "Generative adversarial networks conditioned on brain activity reconstruct seen images," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2018, pp. 1054–1061.

[14] E. J. Holmes and C. G. Gross, "Effects of inferior temporal lesions on discrimination of stimuli differing in orientation," *J. Neurosci.*, vol. 4, pp. 3063–3068, 1984.

[15] J. Maunsell and V. D. Essen, "Functional properties of neurons in middle temporal visual area of the macaque monkey. I. selectivity for stimulus direction, speed, and orientation," *J. Neuriophysiol.*, vol. 49, no. 5, pp. 1127–1147, 1983.

[16] H. DH and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, pp. 106–154, 1962.

[17] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon, "How MT cells analyze the motion of visual patterns," *Nature Neurosci.*, vol. 9, no. 11, pp. 1421–1431, 2006.

[18] D. J. Heeger, E. P. Simoncelli, and J. A. Movshon, "Computational models of cortical visual processing," *Proc. Nat. Acad. Sci.*, vol. 93, no. 2, pp. 623–627, 1996.

[19] M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion, "Seeing it all: Convolutional network layers map the function of the human visual system," *NeuroImage*, vol. 152, pp. 184–194, 2017.

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

[21] 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.

[22] D. Ghadiyaram and A. C. Bovik, "Feature maps driven no-reference image quality prediction of authentically distorted images," in *Proc. SPIE*, 2015, vol. 9394, pp. 158–171.

[23] 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.

[24] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Jan. 2016.

[25] 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.

[26] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2021.3088505.

[27] W. Wu, Q. Li, Z. Chen, and S. Liu, "Semantic information oriented no-reference video quality assessment," *IEEE Signal Process. Lett.*, vol. 28, pp. 204–208, 2021.

[28] Y. Zhang, Z. Liu, Z. Chen, X. Xu, and S. Liu, "No-reference quality assessment of panoramic video based on spherical-domain features," in *Proc. Picture Coding Symp.*, 2021, pp. 1–5.

[29] O. J. Hénaff, R. L. Goris, and E. P. Simoncelli, "Perceptual straightening of natural videos," *Nature Neurosci.*, vol. 22, pp. 984–991, 2019.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1097–1105.

[31] N. S. Mahankali and S. S. Channappayya, "Video quality prediction using voxel-wise fMRI models of the visual cortex," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 2125–2129.

[32] P. ITU, "910. subjective video quality assessment methods for multimedia applications," Int. Telecommun. Union Telecommun. Sector, 1999.

[33] F. D. Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," in *Proc. Int. Workshop Qual. Multimedia Experience*, 2009, pp. 204–209.

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

[35] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[36] V. Hosu *et al.*, "The Konstanz natural video database (KoNViD-1 k)," in *Proc. 9th Int. Conf. Qual. Multimedia Experience*, 2017, pp. 1–6.

[37] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014-A database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073–3086, Jul. 2016.

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