Quality Assessment of In-the-Wild Videos

Dingquan Li
dingquanli@pku.edu.cn
NELVT, LMAM, School of Mathematical Sciences & BICMR, Peking University

Tingting Jiang
ttjiang@pku.edu.cn
NELVT, Department of Computer Science, Peking University

Ming Jiang
ming-jiang@pku.edu.cn
NELVT, LMAM, School of Mathematical Sciences & BICMR, Peking University

ABSTRACT

Quality assessment of in-the-wild videos is a challenging problem because of the absence of reference videos and shooting distortions. Knowledge of the human visual system can help establish methods for objective quality assessment of in-the-wild videos. In this work, we show two eminent effects of the human visual system, namely, content-dependency and temporal-memory effects, could be used for this purpose. We propose an objective no-reference video quality assessment method by integrating both effects into a deep neural network. For content-dependency, we extract features from a pre-trained image classification neural network for its inherent content-aware property. For temporal-memory effects, long-term dependencies, especially the temporal hysteresis, are integrated into the network with a gated recurrent unit and a subjectively-inspired temporal pooling layer. To validate the performance of our method, experiments are conducted on three publicly available in-the-wild video quality assessment databases: KoNViD-1k, CVD2014, and LIVE-Qualcomm, respectively. Experimental results demonstrate that our proposed method outperforms five state-of-the-art methods by a large margin, specifically, 12.39%, 15.71%, 15.45%, and 18.09% overall performance improvements over the second-best method VBLIINDS, in terms of SROCC, KROCC, PLCC and RMSE, respectively. Moreover, the ablation study verifies the crucial role of both the content-aware features and the modeling of temporal-memory effects. The PyTorch implementation of our method is released at https://github.com/lidq92/VSFA.

KEYWORDS

video quality assessment; human visual system; content dependency; temporal-memory effects; in-the-wild videos

ACM Reference Format:
Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality Assessment of In-the-Wild Videos. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351028

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351028

Figure 1: [Best viewed when zoomed in] Human judgments of visual quality are content-dependent. The first/second row shows a pair of in-focus/out-of-focus images. Every two images in a pair are taken in the same shooting condition, and they only differ in image content. However, user study shows that humans consistently prefer the left ones.

1 INTRODUCTION

Nowadays, most videos are captured in the wild by users with diverse portable mobile devices, which may contain annoying distortions due to out of focus, object motion, camera shake, or under/over exposure. Thus, it is highly desirable to automatically identify and cull low-quality videos, prevent their occurrence by quality monitoring processes during acquisition, or repair/enhance them with the quality-aware loss. To achieve this goal, quality assessment of in-the-wild videos is a precondition. However, this is a challenging problem due to the fact that the "perfect" source videos are not available and the shooting distortions are unknown. There is an essential difference between in-the-wild videos and synthetically-distorted videos, *i.e.*, the former contains a mass of content and may suffer from complex mixed real-world distortions that are temporally heterogeneous. On account of this, current state-of-the-art video quality assessment (VQA) methods (*e.g.*, VBLIINDS [35] and VIIDEO [28]) validated on traditional synthetic VQA databases [30, 38] fail in predicting the quality of in-the-wild videos [10, 23, 31, 42].

This work focuses on the problem "quality assessment of in-the-wild videos". Since humans are the end-users, we believe that knowledge of the human visual system (HVS) can help establish objective methods for our problem. Specifically, two eminent effects of HVS are incorporated into our method.

Human judgments of visual image/video quality depend on content, which is well known in many subjective experiments [1, 6, 26, 41, 43, 46, 53]. For images, Siahaan *et al.* show that scene and object categories influence human judgments of visual quality for JPEG compressed and blurred images [41]. Two compressed images with the same compression ratio may have different subjective quality if they contain different scenes [43], since the scene content can have different impact on the compression operations and the visibility of artifacts. For videos, similar content dependency can be found in compressed video quality assessment [26, 46] and quality-of-experience of streaming videos [1, 6]. Unlike quality assessment of synthetically-distorted images/videos, quality assessment of in-the-wild images/videos essentially requires to compare cross-content image/video pairs (*i.e.*, the pair from different reference images/videos) [25], which may be more strongly affected by content. To verify the correctness of this effect on our problem, we collect data and conduct a user study. We ask 10 human subjects to do the cross-content pairwise comparison for 201 image pairs. More than 7 of 10 subjects prefer one image to the other image in 82 image pairs. For illustration, two pairs of in-the-wild images are shown in Figure 1. Each image pair is taken in the same shooting conditions (*e.g.*, focus length, object distance). For the in-focus image pair in the first row, 9 of 10 subjects prefer the left one. For the out-of-focus image pair in the second row, 8 of 10 subjects prefer the left one to the right one. The only difference within a pair is the image content, so from our user study, we can infer that image content can affect human perception on quality assessment of in-the-wild images. We also conduct a user study for 43 video pairs, where every two videos in a pair are taken in similar settings. Similar results are found that video content could have impacts on judgments of visual quality for in-the-wild videos. In the supplemental material, we provide a video pair, for which all 10 subjects prefer the same video. Thus, we consider content-aware features in our problem to address the content dependency.

Human judgments of video quality are affected by their temporal memory. Temporal-memory effects indicate that human judgments of current frame rely on the current frame and information from previous frames. And this implies that long-term dependencies exist in the VQA problem. More specifically, humans remember poor quality frames in the past and lower the perceived quality scores for following frames, even when the frame quality has returned to acceptable levels [37]. This is called the temporal hysteresis effect. It indicates that the simple average pooling strategy overestimates the quality of videos with fluctuating frame-wise quality scores. Since the in-the-wild video contains more temporally-heterogeneous distortions than the synthetically-distorted video, human judgments of its visual quality reflect stronger hysteresis effects. Therefore, in our problem, modeling of temporal-memory effects should be taken into account.

In light of the two effects, we propose a simple yet effective no-reference (NR) VQA method with content-aware features and modeling of temporal-memory effects. To begin with, our method extracts content-aware features from deep convolutional neural networks (CNN) pre-trained on image classification tasks, for they are able to discriminate abundant content information. After that, it includes a gated recurrent unit (GRU) for modeling long-term dependencies and predicting frame quality. Finally, to take the temporal hysteresis effects into account, we introduce a differentiable subjectively-inspired temporal pooling model, and embed it as a layer into the network to output the overall video quality.

To demonstrate the performance of our method, we conduct experiments on three publicly available databases, *i.e.*, KoNViD-1k [12], LIVE-Qualcomm [10] and CVD2014 [31]. Our method is compared with five state-of-the-art methods, and its superior performance is proved by the experimental results. Moreover, the ablation study verifies the key role of each component in our method. This suggests that incorporating the knowledge of HVS could make objective methods more consistent with human perception.

The main contributions of this work are as follows:

- An objective NR-VQA method and the first deep learning-based model is proposed for in-the-wild videos.
- To our best knowledge, it is the first time that a GRU network is applied to model the long-term dependencies for quality assessment of in-the-wild videos and a differentiable temporal pooling model is put forward to account for the hysteresis effect.
- The proposed method outperforms the state-of-the-art methods by large margins, which is demonstrated by experiments on three large-scale in-the-wild VQA databases.

2 RELATED WORK

2.1 Video Quality Assessment

Traditional VQA methods consider structures [47, 48], gradients [21], motion [22, 36], energy [18], saliency [52, 54], or natural video statistics [9, 28, 35, 57]. Besides, quality assessment can be achieved by fusion of primary features [8, 19]. Recently, four deep learning-based VQA methods are proposed [15, 20, 55, 56]. Kim *et al.* [15] utilize CNN models to learn the spatio-temporal sensitivity maps. Liu *et al.* [20] exploit the 3D-CNN model for codec classification and quality assessment of compressed videos. Zhang *et al.* [55, 56] apply the transfer learning technique with CNN for video quality assessment. However, all these methods are trained, validated, and tested on synthetically distorted videos. Streaming video quality-of-experience is relevant to video quality but beyond the scope of this paper, and an interested reader can refer to the good surveys [14, 39].

Quality assessment of in-the-wild videos is a quite new topic in recent years [10, 12, 31, 42]. Four relevant databases have been constructed and corresponding subjective studies have been conducted. Overall, CVD2014 [31], KoNViD-1k [12], and LIVE-Qualcomm [10] are publicly available, while LIVE-VQC [42] will be available soon. Due to the fact that we cannot access the pristine reference videos in this situation, only NR-VQA methods are applicable. Unfortunately, the evaluation of current state-of-the-art NR-VQA methods [28, 35] on these video databases shows a poor performance [10, 23, 31, 42]. Existing deep learning-based VQA models are unfeasible in our problem since they either need the reference information [15, 55, 56] or only suit for compression artifacts [20]. Thus, this motivates us to propose the first deep learning-based model that is capable of predicting the quality of in-the-wild videos.

2.2 Content-Aware Features

Content-aware features can help addressing content-dependency on the predicted image/video quality, so as to improve the performance of objective models [13, 17, 41, 49]. Jaramillo *et al.* [13]

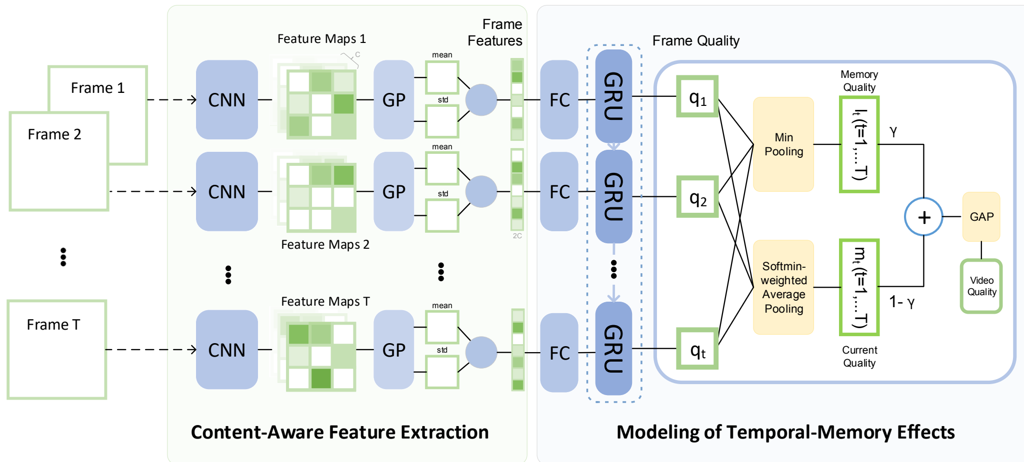


Figure 2: The overall framework of the proposed method. It mainly consists of two modules. The first module "content-aware feature extraction" is a pre-trained CNN with effective global pooling (GP) serving as a feature extractor. The second module "modeling of temporal-memory effects" includes two sub-modules: one is a GRU network for modeling long-term dependencies; the other is a subjectively-inspired temporal pooling layer accounting for the temporal hysteresis effects. Note that the GRU network is the unrolled version of one GRU and the parallel CNNs/FCs share weights.

extract handcrafted content-relevant features to tune existing quality measures. Siahaan *et al.* [41] and Wu *et al.* [49] utilize semantic information from the top layer of pre-trained image classification networks to incorporate with traditional quality features. Li *et al.* [17] exploit the deep semantic feature aggregation of multiple patches for image quality assessment. It is shown that these deep semantic features alleviate the impact of content on the quality assessment task. Inspired by their work, we consider using pre-trained image classification networks for content-aware feature extraction as well. Unlike the work in [17], to get the features, we directly feed the whole frame into the network and apply not only global average pooling but also global standard deviation pooling to the output semantic feature maps. Since our work aims at the VQA task, we further put forward a new module for modeling temporal characteristics of human behaviors when rating video quality.

2.3 Temporal Modeling

The temporal modeling in the VQA field can be viewed in two aspects, *i.e.*, feature aggregation and quality pooling.

In the feature aggregation aspect, most methods aggregate frame-level features to video-level features by averaging them over the temporal axis [8, 18, 22–24, 35]. Li *et al.* [19] adopt a 1D convolutional neural network to aggregate the primary features for a time interval. Unlike the previous methods, we consider using GRU network to model the long-term dependencies for feature integration.

In the quality pooling aspect, the simple average pooling strategy is adopted by many methods [20, 28, 36, 45, 57]. Several pooling strategies considering the recency effect or the worst quality section influence are discussed in [34, 40]. Kim *et al.* [15] adopt a convolutional neural aggregation network (CNAN) for learning frame weights, then the overall video quality is calculated by the weighted average of frame quality scores. Seshadrinathan and Bovik [37] notice the temporal hysteresis effect in the subjective experiments, and propose a temporal hysteresis pooling strategy for quality assessment. The effectiveness of this strategy has been verified in [3, 37, 50]. We also take account of the temporal hysteresis effects. However, the temporal pooling model in [37] is not differentiable. So we introduce a new one with subjectively-inspired weights which can be embedded into the neural network and be trained with back propagation as well. In the experimental part, we will show that this new temporal pooling model with subjectively-inspired weights is better than the CNAN temporal pooling [15] with learned weights.

3 THE PROPOSED METHOD

In this section, we introduce a novel NR-VQA method by integrating knowledge of the human visual system into a deep neural network. The framework of the proposed method is shown in Figure 2. It extracts content-aware features from a modified pre-trained CNN with global pooling (GP) for each video frame. Then the extracted frame-level features are sent to a fully-connected (FC) layer for dimensional reduction followed by a GRU network for long-term dependencies modeling. In the meantime, the GRU outputs the frame-wise quality scores. Lastly, to account for the temporal hysteresis effect, the overall video quality is pooled from these frame quality scores by a subjectively-inspired temporal pooling layer. We will detail each part in the following.

3.1 Content-Aware Feature Extraction

For in-the-wild videos, the perceived video quality strongly depends on the video content as described in Section 1. This can be attributed to the fact that, the complexity of distortions, the human tolerance thresholds for distortions, and the human preferences could vary for different video content/scenes.

To evaluate the perceived quality of in-the-wild videos, the above observation motivates us to extract features that are not only perceptual (distortion-sensitive) but also content-aware. The image classification models pre-trained on ImageNet [4] using CNN possess the discriminatory power of different content information. Thus, the deep features extracted from these models (*e.g.* ResNet [11]) are expected to be content-aware. Meanwhile, the deep features are distortion-sensitive [5]. So it is reasonable to extract content-aware perceptual features from pre-trained image classification models.

Firstly, assuming the video has T frames, we feed the video frame $\mathbf{I}_t (t = 1, 2, \ldots, T)$ into a pre-trained CNN model and output the deep semantic feature maps $\mathbf{M}_t$ from its top convolutional layer:

$$\mathbf{M}_t = \mathrm{CNN}(\mathbf{I}_t). \tag{1}$$

$\mathbf{M}_t$ contains a total of C feature maps. Then, we apply spatial GP for each feature map of $\mathbf{M}_t$. Applying the spatial global average pooling operation ($\mathrm{GP}_{\mathrm{mean}}$) to $\mathbf{M}_t$ discards much information of $\mathbf{M}_t$. We further consider the spatial global standard deviation pooling operation ($\mathrm{GP}_{\mathrm{std}}$) to preserve the variation information in $\mathbf{M}_t$. The output feature vectors of $\mathrm{GP}_{\mathrm{mean}}$, $\mathrm{GP}_{\mathrm{std}}$ are $\mathbf{f}_t^{\mathrm{mean}}$, $\mathbf{f}_t^{\mathrm{std}}$ respectively.

$$\begin{aligned} \mathbf{f}_t^{\mathrm{mean}} &= \mathrm{GP}_{\mathrm{mean}}(\mathbf{M}_t), \\ \mathbf{f}_t^{\mathrm{std}} &= \mathrm{GP}_{\mathrm{std}}(\mathbf{M}_t). \end{aligned} \tag{2}$$

After that, $\mathbf{f}_t^{\mathrm{mean}}$ and $\mathbf{f}_t^{\mathrm{std}}$ are concatenated to serve as the content-aware perceptual features $\mathbf{f}_t$:

$$\mathbf{f}_t = \mathbf{f}_t^{\mathrm{mean}} \oplus \mathbf{f}_t^{\mathrm{std}}, \tag{3}$$

where $\oplus$ is the concatenation operator and the length of $\mathbf{f}_t$ is $2C$.

3.2 Modeling of Temporal-Memory Effects

Temporal modeling is another important clue for designing objective VQA models. We model the temporal-memory effects in two aspects. In the feature integration aspect, we adopt a GRU network for modeling the long-term dependencies in our method. In the quality pooling aspect, we propose a subjectively-inspired temporal pooling model and embed it into the network.

Long-term dependencies modeling. Existing NR-VQA methods cannot well model the long-term dependencies in the VQA task. To handle this issue, we resort to GRU [2]. It is a recurrent neural network model with gates control which is capable of both integrating features and learning long-term dependencies. Specifically, in this paper, we consider using GRU to integrate the content-aware perceptual features and predict the frame-wise quality scores.

The extracted content-aware features are of high dimension, which is not easy for training GRU. Therefore, it is better to perform dimension reduction before feeding them into GRU. It could be beneficial by performing dimension reduction with other steps in the optimization process jointly. In this regard, we perform dimension reduction using a single FC layer, that is:

$$\mathbf{x}_t = \mathbf{W}_{fx}\mathbf{f}_t + \mathbf{b}_{fx}, \tag{4}$$

where $\mathbf{W}_{fx}$ and $\mathbf{b}_{fx}$ are the parameters in the single FC layer. Without the bias term, it acts as a linear dimension reduction model.

After dimension reduction, the reduced features $\mathbf{x}_t (t = 1, \cdots, T)$ are sent to GRU. We consider the hidden states of GRU as the integrated features, whose initial values are $\mathbf{h}_0$. The current hidden state $\mathbf{h}_t$ is calculated from the current input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$, that is:

$$\mathbf{h}_t = \mathrm{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}). \tag{5}$$

With the integrated features $\mathbf{h}_t$, we can predict the frame quality score q_t by adding a single FC layer:

$$q_t = \mathbf{W}_{hq}\mathbf{h}_t + \mathbf{b}_{hq}, \tag{6}$$

where $\mathbf{W}_{hq}$ and $\mathbf{b}_{hq}$ are the weight and bias parameters.

Subjectively-inspired temporal pooling. In subjective experiments, subjects are intolerant of poor quality video events [32]. More specifically, temporal hysteresis effect is found in the subjective experiments, *i.e.*, subjects react sharply to drops in video quality and provide poor quality for such time interval, but react dully to improvements in video quality thereon [37].

A temporal pooling model is adopted in [37] to account for the hysteresis effect. Specifically, a memory quality element is defined as the minimum of the quality scores over the previous frames; a current quality element is defined as a sort-order-based weighted average of the quality scores over the next frames; the approximate score is calculated as the weighted average of the memory and current elements; the video quality is computed as the temporal average pooling of the approximate scores. However, there are some limitations on directly applying this model to the NR quality assessment of in-the-wild videos. First, this model requires the reliable frame quality scores as input, which cannot be provided in our task. Second, the model in [37] is not differentiable due to the sort-order-based weights in the definition of the current quality element. Thus it cannot be embedded into the neural network. In our problem, since we only have access to the overall subjective video quality, we need to learn the neural network without frame-level supervision. Thus, to connect the predicted frame quality score q_t to the video quality Q, we put forward a new differentiable temporal pooling model by replacing the sort-order-based weight function in [37] with a differentiable weight function, and embed it into the network. Details are as follow.

To mimic the human's intolerance to poor quality events, we define a memory quality element l_t at the t-th frame as the minimum of quality scores over the previous several frames:

$$\begin{aligned} l_t &= q_t, & &\text{for } t = 1, \\ l_t &= \min_{k \in V_{prev}} q_k, & &\text{for } t > 1, \end{aligned} \tag{7}$$

where $V_{prev} = \{\max(1, t - \tau), \cdots, t - 2, t - 1\}$ is the index set of the considered frames, and τ is a hyper-parameter relating to the temporal duration.

Accounting for the fact that subjects react sharply to the drops in quality but react dully to the improvements in quality, we construct a current quality element m_t at the t-th frame, using the weighted quality scores over the next several frames, where larger weights are assigned for worse quality frames. Specifically, we define the weights w_t^k by a differentiable softmin function (a composition of the negative linear function and the softmax function).

$$\begin{aligned} m_t &= \sum_{k \in V_{next}} q_k w_t^k, \\ w_t^k &= \frac{e^{-q_k}}{\sum_{j \in V_{next}} e^{-q_j}}, k \in V_{next}, \end{aligned} \tag{8}$$

where $V_{next} = \{t, t+1, \cdots, \min(t+\tau, T)\}$ is the index set of the related frames.

In the end, we approximate the subjective frame quality scores by linearly combining the memory quality and current quality elements. The overall video quality Q is then calculated by temporal global average pooling (GAP) of the approximate scores:

$$q'_t = \gamma l_t + (1-\gamma) m_t, \tag{9}$$

$$Q = \frac{1}{T}\sum_{t=1}^{T} q'_t, \tag{10}$$

where γ is a hyper-parameter to balance the contributions of memory and current elements to the approximate score.

Note that we model the temporal-memory effects with both a global module (*i.e.*, GRU) and a local module (*i.e.*, subjectively-inspired temporal pooling with a window size of $2\tau + 1$). The long-term dependency is always considered by GRU, no matter which value of τ in the temporal pooling is chosen.

3.3 Implementation Details

We choose ResNet-50 [11] pre-trained on ImageNet [4] for the content-aware feature extraction, and the feature maps are extracted from its 'res5c' layer. In this instance, the dimension of $\mathbf{f}_t$ is 4096. The long-term dependencies part is a single FC layer that reduces the feature dimension from 4096 to 128, followed by a single-layer GRU network whose hidden size is set as 32. The subjectively-inspired temporal pooling layer contains two hyper-parameters, τ and γ, which are set as 12 and 0.5, respectively. We fix the parameters in the pre-trained ResNet-50 to ensure that the content-aware property is not altered, and we train the whole network in an end-to-end manner. The proposed model is implemented with PyTorch [33]. The L_1 loss and Adam [16] optimizer with an initial learning rate 0.00001 and training batch size 16 are used for training our model.

4 EXPERIMENTS

We first describe the experimental settings, including the databases, compared methods and basic evaluation criteria. Next, we carry out the performance comparison and result analysis of our method with five state-of-the-art methods. After that, an ablation study is conducted. Then, we show results of different choices of feature extractor and temporal pooling strategy. Finally, the adding value of motion information and computational efficiency are discussed.

4.1 Experimental Settings

Databases. There are four databases constructed for our problem: LIVE Video Quality Challenge Database (LIVE-VQC) [42], Konstanz Natural Video Database (KoNViD-1k) [12], LIVE-Qualcomm Mobile In-Capture Video Quality Database (LIVE-Qualcomm) [10], and Camera Video Database (CVD2014) [31]. The latter three are now publicly available, while the first one is not accessible now. So we conduct experiments on KoNViD-1k, LIVE-Qualcomm and CVD2014. Subjective quality scores are provided in the form of mean opinion score (MOS).

KoNViD-1k [12] aims at natural distortions. To guarantee the video content diversity, it comprises a total of 1,200 videos of resolution 960×540 that are fairly sampled from a large public video dataset, YFCC100M. The videos are 8s with 24/25/30fps. The MOS ranges from 1.22 to 4.64.

LIVE-Qualcomm [10] aims at in-capture video distortions during video acquisition. It includes 208 videos of resolution 1920×1080 captured by 8 different smart-phones and models 6 in-capture distortions (artifacts, color, exposure, focus, sharpness and stabilization). The videos are 15s with 30fps. The realignment MOS ranges from 16.5621 to 73.6428.

CVD2014 [31] also aims at complex distortions introduced during video acquisition. It contains 234 videos of resolution 640×480 or 1280×720 recorded by 78 different cameras. The videos are 10-25s with 11-31fps, which are a wide range of time span and fps. The realignment MOS ranges from -6.50 to 93.38.

Compared methods. Because only NR methods are applicable for quality assessment of in-the-wild videos, we choose five state-of-the-art NR methods (whose original codes are released by the authors) for comparison: VBLIINDS [35], VIIDEO [28], BRISQUE [27][1], NIQE [29], and CORNIA [51]. Note that we cannot compare with the three recent deep learning-based general VQA methods, since [55] needs scores of full-reference methods and [15, 56] are full-reference methods, which are unfeasible for our problem.

Basic evaluation criteria. Spearman's rank-order correlation coefficient (SROCC), Kendall's rank-order correlation coefficient (KROCC), Pearson's linear correlation coefficient (PLCC) and root mean square error (RMSE) are the four performance criteria of VQA methods. SROCC and KROCC indicate the prediction monotonicity, while PLCC and RMSE measure the prediction accuracy. Better VQA methods should have larger SROCC/KROCC/PLCC and smaller RMSE. When the objective scores (*i.e.*, the quality scores predicted by a VQA method) are not the same scale as the subjective scores, we refer to the suggestion of Video Quality Experts Group (VQEG) [44] before calculating PLCC and RMSE values, and adopt a four-parameter logistic function for mapping the objective score o to the subjective score s:

$$f(o) = \frac{\tau_1 - \tau_2}{1 + e^{-\frac{o-\tau_3}{\tau_4}}} + \tau_2, \tag{11}$$

where τ_1 to τ_4 are fitting parameters initialized with $\tau_1 = \max(s)$, $\tau_2 = \min(s)$, $\tau_3 = \text{mean}(o)$, $\tau_4 = \text{std}(o)/4$.

4.2 Performance Comparison

For each database, 60%, 20%, and 20% data are used for training, validation, and testing, respectively. There is no overlap among these three parts. This procedure is repeated 10 times and the mean and standard deviation of performance values are reported in Table 1. For VBLIINDS, BRISQUE and our method, we choose the models with the highest SROCC values on the validation set during the training phase. NIQE, CORNIA, and VIIDEO are tested on the same 20% testing data after the parameters in Eqn. (11) are optimized with the training and validation data.

Table 1 summarizes the performance values on the three databases, and the overall performance values (indicated by the weighted performance values) as well. Our method achieves the best overall performance in terms of both the prediction monotonicity (SROCC, KROCC) and the prediction accuracy (PLCC, RMSE), and have a

[1]Video-level features of BRISQUE are the average pooling of its frame-level features.

Table 1: Performance comparison on the three VQA databases. Mean and standard deviation (std) of the performance values in 10 runs are reported, *i.e.*, mean (± std). 'Overall Performance' shows the weighted-average performance values over all three databases, where weights are proportional to database-sizes. In each column, the best and second-best values are marked in boldface and underlined, respectively.

Method	Overall Performance				LIVE-Qualcomm [10]				
	SROCC↑	KROCC↑	PLCC↑	RMSE↓	SROCC↑	p-value (<0.05)	KROCC↑	PLCC↑	RMSE↓
BRISQUE [27]	0.643 (± 0.059)	0.465 (± 0.047)	0.625 (± 0.053)	3.895 (± 0.380)	0.504 (± 0.147)	1.21E-04	0.365 (± 0.111)	0.516 (± 0.127)	10.731 (± 1.335)
NIQE [29]	0.526 (± 0.055)	0.369 (± 0.041)	0.542 (± 0.054)	4.214 (± 0.323)	0.463 (± 0.105)	5.28E-07	0.328 (± 0.088)	0.464 (± 0.136)	10.858 (± 1.013)
CORNIA [51]	0.591 (± 0.052)	0.423 (± 0.043)	0.595 (± 0.051)	4.139 (± 0.300)	0.460 (± 0.130)	4.98E-06	0.324 (± 0.104)	0.494 (± 0.133)	10.759 (± 0.939)
VIIDEO [28]	0.237 (± 0.073)	0.164 (± 0.050)	0.218 (± 0.070)	5.115 (± 0.285)	0.127 (± 0.137)	9.77E-11	0.082 (± 0.099)	-0.001 (± 0.106)	12.308 (± 0.881)
VBLIINDS [35]	0.686 (± 0.035)	0.503 (± 0.032)	0.660 (± 0.037)	3.753 (± 0.365)	0.566 (± 0.078)	1.02E-05	0.405 (± 0.074)	0.568 (± 0.089)	10.760 (± 1.231)
Ours	**0.771** (± 0.028)	**0.582** (± 0.029)	**0.762** (± 0.031)	**3.074** (± 0.448)	**0.737** (± 0.045)	-	**0.552** (± 0.047)	**0.732** (± 0.0360)	**8.863** (± 1.042)

Method	KoNViD-1k [12]					CVD2014 [31]				
	SROCC↑	p-value	KROCC↑	PLCC↑	RMSE↓	SROCC↑	p-value	KROCC↑	PLCC↑	RMSE↓
BRISQUE [27]	0.654 (± 0.042)	6.00E-06	0.473 (± 0.034)	0.626 (± 0.041)	0.507 (± 0.031)	0.709 (± 0.067)	7.03E-07	0.518 (± 0.060)	0.715 (± 0.048)	15.197 (± 1.325)
NIQE [29]	0.544 (± 0.040)	7.31E-11	0.379 (± 0.029)	0.546 (± 0.038)	0.536 (± 0.010)	0.489 (± 0.091)	1.73E-10	0.358 (± 0.064)	0.593 (± 0.065)	17.168 (± 1.318)
CORNIA [51]	0.610 (± 0.034)	6.77E-09	0.436 (± 0.029)	0.608 (± 0.032)	0.509 (± 0.014)	0.614 (± 0.075)	5.69E-09	0.441 (± 0.058)	0.618 (± 0.079)	16.871 (± 1.200)
VIIDEO [28]	0.298 (± 0.052)	4.22E-15	0.207 (± 0.035)	0.303 (± 0.049)	0.610 (± 0.012)	0.023 (± 0.122)	3.02E-14	0.021 (± 0.081)	-0.025 (± 0.144)	21.822 (± 1.152)
VBLIINDS [35]	0.695 (± 0.024)	6.75E-05	0.509 (± 0.020)	0.658 (± 0.025)	0.483 (± 0.011)	0.746 (± 0.056)	2.94E-06	0.562 (± 0.0570)	0.753 (± 0.053)	14.292 (± 1.413)
Ours	**0.755** (± 0.025)	-	**0.562** (± 0.022)	**0.744** (± 0.029)	**0.469** (± 0.054)	**0.880** (± 0.030)	-	**0.705** (± 0.044)	**0.885** (± 0.031)	**11.287** (± 1.943)

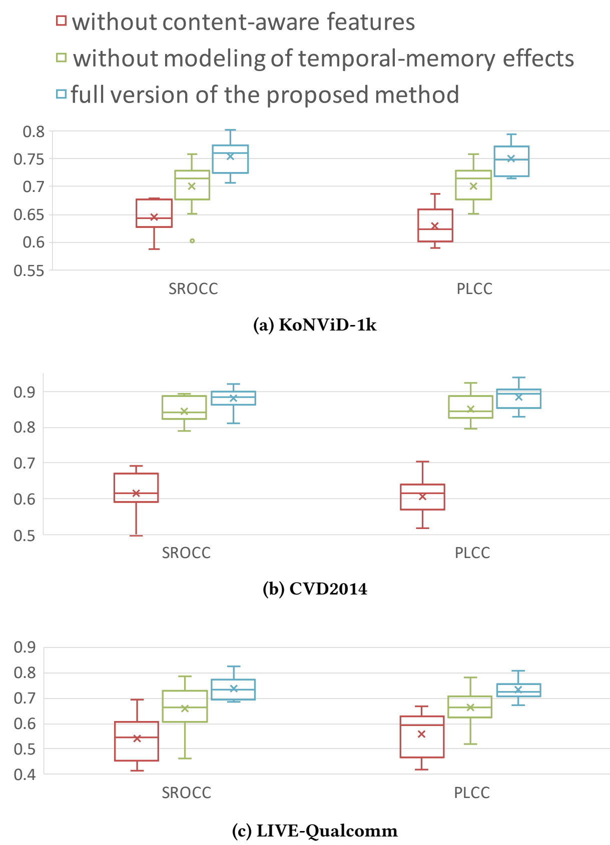
large gain over the second-best method VBLIINDS. VIIDEO fails because it is based only on temporal scene statistics and cannot model the complex distortions. For all individual databases, our method outperforms the other compared methods by a large margin. For example, compared to the second-best method VBLIINDS, in terms of SROCC, our method achieves 30.21% improvements on LIVE-Qualcomm, 8.63% improvements on KoNViD-1k and 17.96% improvements on CVD2014. Among the three databases, LIVE-Qualcomm is the most challenging one for the compared methods and our method—not only mean performance values are small but also standard deviation values for all methods are large. This verifies the statement in [10] that videos in LIVE-Qualcomm challenge both human viewers and objective VQA models.

Statistical significance. We further carry out the statistical significance test to see whether the results shown in Table 1 are statistical significant or not. On each database, the paired t-test is conducted at 5% significance level using the SROCC values (in 10 runs) of our method and of the compared one. The p-values are shown in Table 1. All are smaller than 0.05 and prove our method is significantly better than all the other five state-of-the-art methods.

4.3 Ablation Study

To demonstrate the importance of each module in our framework, we conduct an ablation study. The overall 10-run-results are shown in the form of box plots in Figure 3.

Content-aware features. We first show the performance drop due to the removal of the content-aware features. When we remove the content-aware features extracted from CNN, we use BRISQUE [27] features instead (red). The removal of the content-aware features causes significant performance drop in all three databases. p-values are 1.10E-05, 1.76E-08, 2.47E-06, and 14.57%, 30.00%, 26.87% decrease in terms of SROCC are found on KoNViD-1k, CVD2014 and LIVE-Qualcomm respectively. Content-aware perceptual features contribute most to our method, which verifies that content-aware perceptual features are crucial for assessing the perceived quality of in-the-wild videos.



(a) KoNViD-1k

(b) CVD2014

(c) LIVE-Qualcomm

Figure 3: Box plots of the ablation study.

Modeling of temporal-memory effects. To verify the effectiveness of modeling of temporal-memory effects, we compare the full version of our proposed method (blue) with the whole temporal modeling module removed (green). Temporal modeling provides 7.70%, 4.14%, 12.01% SROCC gains on KoNViD-1k, CVD2014 and LIVE-Qualcomm respectively, where the p-values are 4.00E-04, 1.11E-04, and 8.49E-03. In view of PLCC, it leads to 5.98%, 4.00%, 10.41% performance improvements on KoNViD-1k, CVD2014 and LIVE-Qualcomm respectively. We further do the ablation study on KoNViD-1k for the two individual temporal sub-modules separately. Removal of long-term dependencies modeling leads to 2.12% decrease in terms of SROCC, while removal of subjectively-inspired temporal pooling leads to 2.68% decrease in terms of SROCC. This indicates the two temporal sub-modules (one is global and the other is local) are complementary.

4.4 Choice of Feature Extractor

There are many choices for content-aware feature extraction. In the following, we mainly consider the pre-trained image classification models and the global standard deviation (std) pooling.

Pre-trained image classification models. In our implementation, we choose ResNet-50 as the content-aware feature extractor. It is interesting to explore other pre-trained image classification models for feature extraction. The results in Table 2 show that VGG16 have similar performance with ResNet-50 (p-values of paired t-test using SROCC values are greater than 0.05, actually 0.1011). However, ResNet-50 has less parameters than AlexNet and VGG16.

Table 2: Performance of different pre-trained image classification models on KoNViD-1k.

Pre-trained model	SROCC↑	KROCC↑	PLCC↑
ResNet-50	0.755 (±0.025)	0.562 (±0.022)	0.744 (±0.029)
AlexNet	0.732 (±0.040)	0.540 (±0.036)	0.731 (±0.035)
VGG16	0.745 (±0.024)	0.554 (±0.023)	0.747 (±0.022)

Global std pooling. When the global std pooling is removed, the performance on KoNViD-1k drops as shown in Figure 4. mean SROCC drops from 0.755 to 0.701, while mean PLCC drops significantly from 0.744 to 0.672. This verifies that global std pooling preserves more information and thus results in good performance.

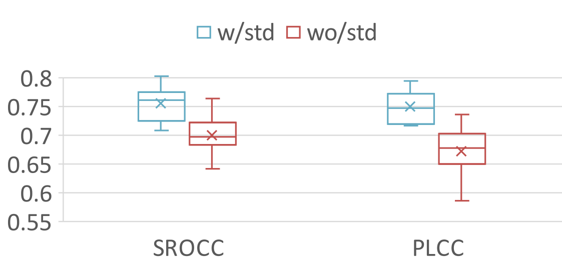


Figure 4: Effectiveness of global std pooling on KoNViD-1k.

4.5 Choices of Temporal Pooling Strategy

Here, we explore different choices of temporal pooling strategy.

Hyper-parameters in subjectively-inspired temporal pooling. The subjectively-inspired temporal pooling contains two hyper-parameters, τ and γ. Figure 5 shows results of different choices of the two parameters. In the left figure, τ is fixed to 12, and γ varies from 0.1 to 0.9 with a step size 0.1. SROCC fluctuates up and down around 0.75, and achieves the best with $\gamma = 0.5$. This is because smaller γ overlooks the memory quality while larger γ overlooks the current quality. In the right figure, γ is fixed to 0.5, and τ varies from 6 to 30 with a step size 6. The highest SROCC value is obtained with $\tau = 12$, which suggests temporal hysteresis effect may lasts about one second for videos with a frame rate of 25fps.

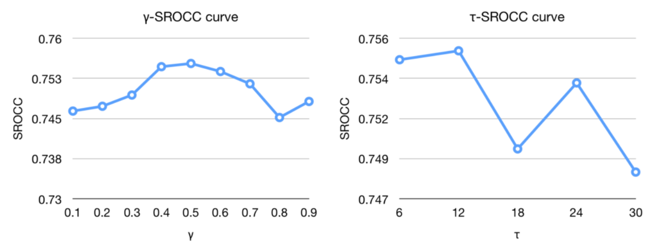


Figure 5: Performance on KoNViD-1k of different hyper-parameters in subjectively-inspired temporal pooling

Pooling in subjective-inspired temporal pooling. To verify the effectiveness of min pooling, we compare it with average pooling. The results on KoNViD-1k are shown in Table 3. And we can see that average pooling is statistically worse than min pooling (p-value is 3.04E-04). This makes sense since min pooling accounts for "humans are quick to criticize and slow to forgive".

Table 3: Effectiveness of min pooling in subjective-inspired temporal pooling on KoNViD-1k.

pooling	SROCC↑	p-value	KROCC↑	PLCC↑
min	**0.755** (±0.025)	-	**0.562** (±0.022)	**0.744** (±0.029)
average	0.736 (±0.031)	3.04E-4	0.543 (±0.027)	0.740 (±0.027)

Handcrafted weights vs. learned weights. Our subjectively-inspired temporal pooling can be regarded as a weighted average pooling strategy, where the weights are designed by hand (see Eqn. (7), (8) and (9)) to mimic the temporal-memory effects. One interesting question is whether the performance can be further improved by making the weights learnable. One possible way is using a temporal CNN (TCNN) to learn the approximate scores $\mathbf{q}'$ from the frame quality scores $\mathbf{q}$, *i.e.*,

$$\mathbf{q}' = \mathrm{TCNN}(\mathbf{q}, \mathbf{kernel\_size} = 2\tau + 1) = \mathbf{w} \otimes \mathbf{q},$$

where $\otimes$ means the convolutional operator, and $\mathbf{w}$ is the learnable weights of TCNN with length $2\tau + 1$ (the same size as ours).

Another way is by the convolutional neural aggregation network (CNAN) introduced in [15]. It is formulated as follow:

$$\boldsymbol{\omega} = \mathrm{softmax}(\mathbf{w}_m \otimes \mathbf{q}),\ Q = \boldsymbol{\omega}^T \mathbf{q},$$

where $\mathbf{w}_m$ is a memory kernel, $\boldsymbol{\omega}$ is the learned frame weights normalized by a softmax function and Q is the overall video quality.

In Figure 6, we report the mean and standard deviation of SROCC values among these three temporal pooling models (including ours) on the three databases. It can be seen that the two models with the learned weights (TCNN and CNAN) underperform the model with handcrafted weights (Ours). This may be explained by the fact that the handcrafted weights are manually designed to mimic the temporal hysteresis effects, while the learned weights do not capture the patterns well.

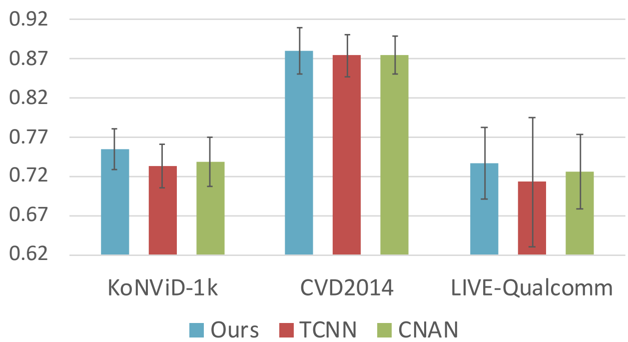


Figure 6: SROCC comparison between temporal pooling models with learned weights or handcrafted weights.

4.6 Motion information

Motion information is important for video processing. In this subsection, we would like to see whether the performance can be further improved with the motion information added. We extract the optical flow using the initialized TVNet [7] without finetuning, and calculate the optical flow statistics as described in [22], then concatenate the statistics to the content-aware features. The performance comparison of our model with/without motion information on KoNViD-1k is shown in Figure 7. Motion information can further improve the performance a little. However, we should note that optical flow computation is very expensive, which makes the small improvements seem unnecessary. It is desired to explore effective and efficient motion-aware features in the VQA task.

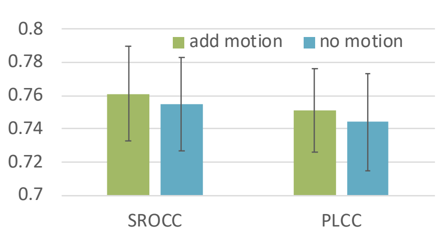


Figure 7: The performance comparison of our model with/without motion information on KoNViD-1k.

4.7 Computational efficiency

Besides the performance, computational efficiency is also crucial for NR-VQA methods. To provide a fair comparison for the computational efficiency of different methods, all tests are carried out on a desktop computer with Intel Core i7-6700K CPU@4.00 GHz, 12G NVIDIA TITAN Xp GPU and 64 GB RAM. The operating system is Ubuntu 14.04. The compared methods are implemented with MATLAB R2016b while our method is implemented with Python 3.6. The default settings of the original codes are used without any modification. From the three databases, we select four videos with different lengths and different resolutions for test. We repeat the tests ten times and the average computation time (seconds) for each method is shown in Table 4. Our method is faster than VBLIINDS—the method with the second-best performance. It is worth mentioning that our method can be accelerated to **30x faster or more** (The larger resolution is, the faster acceleration is.) by simply switching the CPU mode to the GPU mode.

Table 4: The average computation time (seconds) for four videos selected from the original databases. {xxx}frs@{yyy}p indicates the video frame length and the resolution.

Method	240frs@540p	364frs@480p	467frs@720p	450frs@1080p
BRISQUE [27]	12.6931	12.3405	41.2220	79.8119
NIQE [29]	45.6477	41.9705	155.9052	351.8327
CORNIA [51]	225.2185	325.5718	494.2449	616.4856
VIIDEO [28]	137.0538	128.0868	465.2284	1024.5400
VBLIINDS [35]	382.0657	361.3868	1390.9999	3037.2960
Ours	269.8371	249.2085	936.8452	2081.8400

5 CONCLUSION AND FUTURE WORK

In this work, we propose a novel NR-VQA method for in-the-wild videos by incorporating two eminent effects of HVS, *i.e.*, content-dependency and temporal-memory effects. Our proposed method is compared with five state-of-the-art methods on three publicly available in-the-wild VQA databases (KoNViD-1k, CVD2014, and LIVE-Qualcomm), and achieves 30.21%, 8.63%, and 17.96% SROCC improvements on LIVE-Qualcomm, KoNViD-1k, and CVD2014, respectively. Experiments also show that content-aware perceptual features and modeling of temporal-memory effects are of importance for in-the-wild video quality assessment. However, the correlation values of the best method are still less than 0.76 on KoNViD-1k and LIVE-Qualcomm. This indicates that there is ample room for developing an objective model which correlates well with human perception. In the further study, we will consider embedding the spatio-temporal attention models into our framework since they could provide information about when and where the video is important for the VQA problem.

ACKNOWLEDGMENTS

This work was partially supported by the National Basic Research Program of China (973 Program) under contract 2015CB351803, the Natural Science Foundation of China under contracts 61572042, 61520106004, and 61527804. We acknowledge the High-Performance Computing Platform of Peking University for providing computational resources.

REFERENCES

[1] Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. 2017. Study of temporal effects on subjective video quality of experience. *TIP* 26, 11 (Nov. 2017), 5217–5231.

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[3] Lark Kwon Choi and Alan Conrad Bovik. 2018. Video quality assessment accounting for temporal visual masking of local flicker. *SPIC* 67 (2018), 182–198.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. IEEE, 248–255.

[5] Samuel Dodge and Lina Karam. 2016. Understanding how image quality affects deep neural networks. In *QoMEX*. IEEE.

[6] Zhengfang Duanmu, Kede Ma, and Zhou Wang. 2017. Quality-of-experience of adaptive video streaming: Exploring the space of adaptations. In *ACM MM*. ACM, 1752–1760.

[7] Lijie Fan, Wenbing Huang, Stefano Ermon Chuang Gan, Boqing Gong, and Junzhou Huang. 2018. End-to-End Learning of Motion Representation for Video Understanding. In *CVPR*. 6016–6025.

[8] Pedro Garcia Freitas, Welington YL Akamine, and Mylène CQ Farias. 2018. Using multiple spatio-temporal features to estimate video quality. *SPIC* 64 (2018), 1–10.

[9] Deepti Ghadiyaram, Chao Chen, Sasi Inguva, and Anil Kokaram. 2017. A no-reference video quality predictor for compression and scaling artifacts. In *ICIP*. IEEE, 3445–3449.

[10] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush K Moorthy, Prasanjit Panda, and Kai-Chieh Yang. 2018. In-Capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE TCSVT* 28, 9 (Sept. 2018), 2061–2077.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[12] 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*. IEEE.

[13] Benhur Ortiz Jaramillo, Jorge Oswaldo Niño-Castañeda, Ljiljana Platiša, and Wilfried Philips. 2016. Content-aware objective video quality assessment. *JEI* 25, 1 (2016), 013011.

[14] Parikshit Juluri, Venkatesh Tamarapalli, and Deep Medhi. 2015. Measurement of quality of experience of video-on-demand services: A survey. *IEEE Commun. Surv. Tutor.* 18, 1 (2015), 401–418.

[15] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. 2018. Deep Video Quality Assessor: From Spatio-temporal Visual Sensitivity to A Convolutional Neural Aggregation Network. In *ECCV*. 219–234.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. 2019. Which Has Better Visual Quality: The Clear Blue Sky or a Blurry Animal? *IEEE TMM* 21, 5 (May 2019), 1221–1234.

[18] Xuelong Li, Qun Guo, and Xiaoqiang Lu. 2016. Spatiotemporal statistics for video quality assessment. *IEEE TIP* 25, 7 (July 2016), 3329–3342.

[19] Yuming Li, Lai-Man Po, Chun-Ho Cheung, Xuyuan Xu, Litong Feng, Fang Yuan, and Kwok-Wai Cheung. 2016. No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE TCSVT* 26, 6 (June 2016), 1044–1057.

[20] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. 2018. End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks. In *ACM MM*. ACM, 546–554.

[21] Wen Lu, Ran He, Jiachen Yang, Changcheng Jia, and Xinbo Gao. 2019. A spatiotemporal model of video quality assessment via 3D gradient differencing. *Information Sciences* 478 (2019), 141–151.

[22] K Manasa and Sumohana S Channappayya. 2016. An optical flow-based no-reference video quality assessment algorithm. In *ICIP*. IEEE, 2400–2404.

[23] Hui Men, Hanhe Lin, and Dietmar Saupe. 2017. Empirical evaluation of no-reference VQA methods on a natural video quality database. In *QoMEX*. IEEE.

[24] Hui Men, Hanhe Lin, and Dietmar Saupe. 2018. Spatiotemporal Feature Combination Model for No-Reference Video Quality Assessment. In *QoMEX*. IEEE.

[25] Aliaksei Mikhailiuk, María Pérez-Ortiz, and Rafal Mantiuk. 2018. Psychometric scaling of TID2013 dataset. In *QoMEX*.

[26] Milan Mirkovic, Petar Vrgovic, Dubravko Culibrk, Darko Stefanovic, and Andras Anderla. 2014. Evaluating the role of content in subjective video quality assessment. *The Scientific World Journal* 2014 (2014).

[27] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE TIP* 21, 12 (Dec. 2012), 4695–4708.

[28] Anish Mittal, Michele A Saad, and Alan C Bovik. 2016. A completely blind video integrity oracle. *IEEE TIP* 25, 1 (Jan. 2016), 289–300.

[29] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2013. Making a "completely blind" image quality analyzer. *IEEE SPL* 20, 3 (Mar. 2013), 209–212.

[30] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana. 2012. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE JSTSP* 6, 6 (Oct. 2012), 652–671.

[31] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. 2016. CVD2014—A database for evaluating no-reference video quality assessment algorithms. *IEEE TIP* 25, 7 (July 2016), 3073–3086.

[32] Jincheol Park, Kalpana Seshadrinathan, Sanghoon Lee, and Alan Conrad Bovik. 2013. Video quality pooling adaptive to perceptual distortion severity. *IEEE TIP* 22, 2 (Feb. 2013), 610–620.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

[34] Snjezana Rimac-Drlje, Mario Vranjes, and Drago Zagar. 2009. Influence of temporal pooling method on the objective video quality evaluation. In *BMSB*. IEEE, 1–5.

[35] Michele A Saad, Alan C Bovik, and Christophe Charrier. 2014. Blind prediction of natural video quality. *IEEE TIP* 23, 3 (Mar. 2014), 1352–1365.

[36] Kalpana Seshadrinathan and Alan Conrad Bovik. 2010. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE TIP* 19, 2 (Feb. 2010), 335–350.

[37] Kalpana Seshadrinathan and Alan C Bovik. 2011. Temporal hysteresis model of time varying subjective video quality. In *ICASSP*. IEEE, 1153–1156.

[38] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. 2010. Study of subjective and objective quality assessment of video. *IEEE TIP* 19, 6 (June 2010), 1427–1441.

[39] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2014. A survey on quality of experience of HTTP adaptive streaming. *IEEE Commun. Surv. Tutor.* 17, 1 (2014), 469–492.

[40] Michael Seufert, Martin Slanina, Sebastian Egger, and Meik Kottkamp. 2013. "To pool or not to pool": A comparison of temporal pooling methods for HTTP adaptive video streaming. In *QoMEX*. IEEE, 52–57.

[41] Ernestasia Siahaan, Alan Hanjalic, and Judith A Redi. 2018. Semantic-aware blind image quality assessment. *SPIC* 60 (2018), 237–252.

[42] Zeina Sinno and Alan C Bovik. 2019. Large scale study of perceptual video quality. *IEEE TIP* 28, 2 (Feb. 2019), 612–627.

[43] Sophie Triantaphillidou, Elizabeth Allen, and R Jacobson. 2007. Image quality comparison between JPEG and JPEG2000. II. Scene dependency, scene analysis, and classification. *JIST* 51, 3 (2007), 259–270.

[44] VQEG. 2000. Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment.

[45] Phong V Vu and Damon M Chandler. 2014. ViS_3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *JEI* 23, 1 (2014), 013016.

[46] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, Yun Zhang, Jiwu Huang, Sam Kwong, and C.-C. Jay Kuo. 2017. VideoSet: A large-scale compressed video quality dataset based on JND measurement. *JVCIR* 46 (2017), 292–302.

[47] Yue Wang, Tingting Jiang, Siwei Ma, and Wen Gao. 2012. Novel spatio-temporal structural information based video quality metric. *IEEE TCSVT* 22, 7 (July 2012), 989–998.

[48] Zhou Wang, Ligang Lu, and Alan C Bovik. 2004. Video quality assessment based on structural distortion measurement. *SPIC* 19, 2 (2004), 121–132.

[49] Jinjian Wu, Jichen Zeng, Weisheng Dong, Guangming Shi, and Weisi Lin. 2019. Blind image quality assessment with hierarchy: Degradation from local structure to deep semantics. *JVCIR* 58 (2019), 353–362.

[50] Jingtao Xu, Peng Ye, Yong Liu, and David Doermann. 2014. No-reference video quality assessment via feature learning. In *ICIP*. IEEE, 491–495.

[51] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *CVPR*. IEEE, 1098–1105.

[52] Junyong You, Touradj Ebrahimi, and Andrew Perkis. 2014. Attention driven foveated video quality assessment. *IEEE TIP* 23, 1 (2014), 200–213.

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*. 586–595.

[54] Wei Zhang and Hantao Liu. 2017. Study of saliency in objective video quality assessment. *IEEE TIP* 26, 3 (Mar. 2017), 1275–1288.

[55] Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, and Ran He. 2018. Blind Video Quality Assessment with Weakly Supervised Learning and Resampling Strategy. *TCSVT* (2018).

[56] Yu Zhang, Xinbo Gao, Lihuo He, Wen Lu, and Ran He. 2019. Objective Video Quality Assessment Combining Transfer Learning With CNN. *TNNLS* (2019).

[57] Yun Zhu, Yongfang Wang, and Yuan Shuai. 2017. Blind video quality assessment based on spatio-temporal internal generative mechanism. In *ICIP*. IEEE, 305–309.