RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment

Pengfei Chen
cpf00790079@gmail.com
School of Artificial Intelligence,
Xidian University & School of
Information and Control Engineering,
China University of Mining and
Technology

Leida Li*
ldli@xidian.edu.cn
School of Artificial Intelligence,
Xidian University

Lei Ma
malei@mc-oe.com
Hangzhou Multi-Color
Optoelectronics Co., Ltd.

Jinjian Wu
Jinjian.wu@mail.xidian.edu.cn
School of Artificial Intelligence,
Xidian University

Guangming Shi
gmshi@xidian.edu.cn
School of Artificial Intelligence,
Xidian University

ABSTRACT

Video quality assessment (VQA), which is capable of automatically predicting the perceptual quality of source videos especially when reference information is not available, has become a major concern for video service providers due to the growing demand for video quality of experience (QoE) by end users. While significant advances have been achieved from the recent deep learning techniques, they often lead to misleading results in VQA tasks given their limitations on describing 3D spatio-temporal regularities using only fixed temporal frequency. Partially inspired by psychophysical and vision science studies revealing the speed tuning property of neurons in visual cortex when performing motion perception (*i.e.*, sensitive to different temporal frequencies), we propose a novel no-reference (NR) VQA framework named Recurrent-In-Recurrent Network (RIRNet) to incorporate this characteristic to prompt an accurate representation of motion perception in VQA task. By fusing motion information derived from different temporal frequencies in a more efficient way, the resulting temporal modeling scheme is formulated to quantify the temporal motion effect via a hierarchical distortion description. It is found that the proposed framework is in closer agreement with quality perception of the distorted videos since it integrates concepts from motion perception in human visual system (HVS), which is manifested in the designed network structure composed of low- and high- level processing. A holistic validation of our methods on four challenging video quality databases demonstrates the superior performances over the state-of-the-art methods.

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413717

CCS CONCEPTS

• **Computing methodologies → Image processing**; **Neural networks**.

KEYWORDS

video quality assessment; motion perception; speed tuning; temporal frequency

ACM Reference Format:
Pengfei Chen, Leida Li, Lei Ma, Jinjian Wu, and Guangming Shi. 2020. RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413717

1 INTRODUCTION

The spread of reliable and rapid internet connectivity have given rise to an ever-expanding global opportunity for new forms of consumer outreach [1]. The online sharing of user-generated content is now a day-to-day activity for many users around the world. As the main carrier of information transmission, video content is receiving widespread attention from the marketing industry as an upward trend in digital strategy, particularly among organizations that provide user-centric video services [9, 35]. Thus figuring out whether the videos after the production and distribution chains could fulfill the video receivers is of prime importance for the video providers. In order to yield estimates highly consistent with the human visual perception, the video quality assessment (VQA) measurements are urgently needed and have long been a bone of the contention. Among them, subjective metric relying on manual rating for distorted videos is the most reliable approach, where the average of the collecting opinions scores from testing subjects is known as the mean-opinion-score (MOS). However, its real-world applications are restricted for the sake of the time and labor it consumes. As an alternative, researchers pursue objective methods to automatically predict the visual quality of distorted videos.

Although great efforts have been devoted to VQA researches [3, 4, 36, 42, 46], most of them are narrow in scope and fall far short of expectations especially for those videos without references which

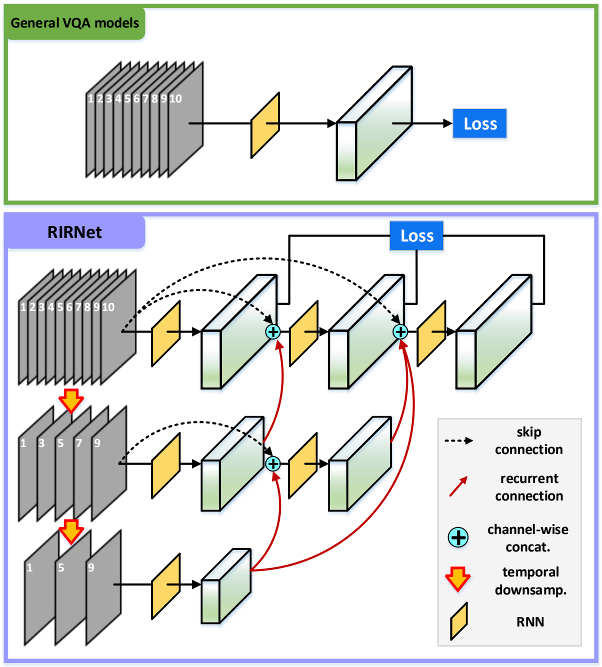

Figure 1: Upper: General deep VQA models which concentrate on fixed temporal frequency. Lower: The hierarchical distortion description is achieved by the proposed RIRNet. To integrate motion information with multiple temporal frequencies, the nested framework is composed of two kinds of connections, i) recurrent connections across low-to-high temporal frequencies, ii) skip connections across coarse-to-fine motion description. The total loss is accumulated by the intermediate products and final output through the deep supervision module.

are more valuable in practical applications. As opposed to the image communities [5, 13, 19], the fact that in video tasks 2D static images are expanded in temporal dimension to have motion information should account for this. Presentations of video sequences to human subjects induce visual experiences of motion and the perceived distortion in video sequences is a combination of both spatial and motion artifacts. Hence we argue that one possible solution to advance video quality prediction resides in an accurate representation of motion perception in video sequences.

Motion perception is a complicated procedure involving processing from low-level to high-level. Specifically, it begins in the striate cortex (Area V1), while the neurons in Area MT, which is driven by an extensive projection of V1 responses, is implicated in integrating local motion information computed by V1 neurons into an overall percept of motion [25]. Area MT is believed to play a role in guidance of some eye movements, segmentation and structure computation in 3D space, models of processing in Area MT is hence essential in VQA due to the critical role of these functions in the perception of videos by human observers. The response properties of neurons in Area MT are well studied in primates. A subset of neurons in Area MT have been explored to be speed-tuned, and their selectivity toward motion direction and speed inhibits the spatio-temporal frequency separability [28, 29]. To the best of our knowledge, few existing VQA models have put forth to take these models into account for motion perception. Hence in our design, it is desired to model the properties of neurons in Area MT through the deep-learning manner.

Based on this observation, we present a novel framework to deal with the no-reference VQA task, termed Recurrent-In-Recurrent Network (RIRNet), intuitively enabling the network to model the motion perception of the neurons in Area MT. The resulting framework and its difference from the general deep VQA models are illustrated in Figure 1. The underlying hypothesis behind our model is that, resorting to the recurrent operations performed by recurrent neural networks (RNNs), motion information in higher temporal frequency could be gradually enriched by those details from lower ones, prior to fusion with the retrospective motion contents of the corresponding time scale. In this case, not only the motion information of previous frames can be captured by current frame, but also the motion information of itself in lower temporal frequencies. Moreover, the skip connections are responsible for recover the lost information in the downsampling operations, forming the building block to depict a coarse-to-fine motion description. The designed temporal modeling mechanism is considered to operate in a recurrent-in-recurrent way, where the output of the current frame is not only used to create prediction candidates for the next frame, but the current state would further propagate to the same frame with higher temporal frequencies.

The contributions of this work could be summarized with the following points:

- We propose a novel deep learning-based framework to solve the NR-VQA task with only input video frames. The framework can be further divided into two parts (*i.e.*, quality degradation learning and motion effect modeling), which can better match quality perception of the distorted videos in human visual system.
- We introduce a hierarchical temporal modeling scheme to model the speed-tuned property of visual neurons in Area MT, benefiting from motion information corresponding to multiple temporal frequencies.
- We demonstrate that the proposed metric outperforms the state-of-the-art for video distortions from both artificial and authentic, as supported by the quantitative and qualitative evaluation on four challenging video quality benchmark databases.

2 RELATED WORK

On the basis of the availability of reference information, objective VQA methods can be further classified as full-reference (FR), reduced-reference (RR), and no-reference (NR) VQA metrics. Entire or partial information of reference videos is attainable in FR/RR-VQA metrics, impelling an appreciable correlation between the predicted results of state-of-the-art FR/RR methods [2, 31, 40] and human visual perception. Contrarily, NR-VQA metrics exploit distortion-specific or natural video statistical models without any information from original videos, which is the major advantage in practical applications and also the primary concern in this work.

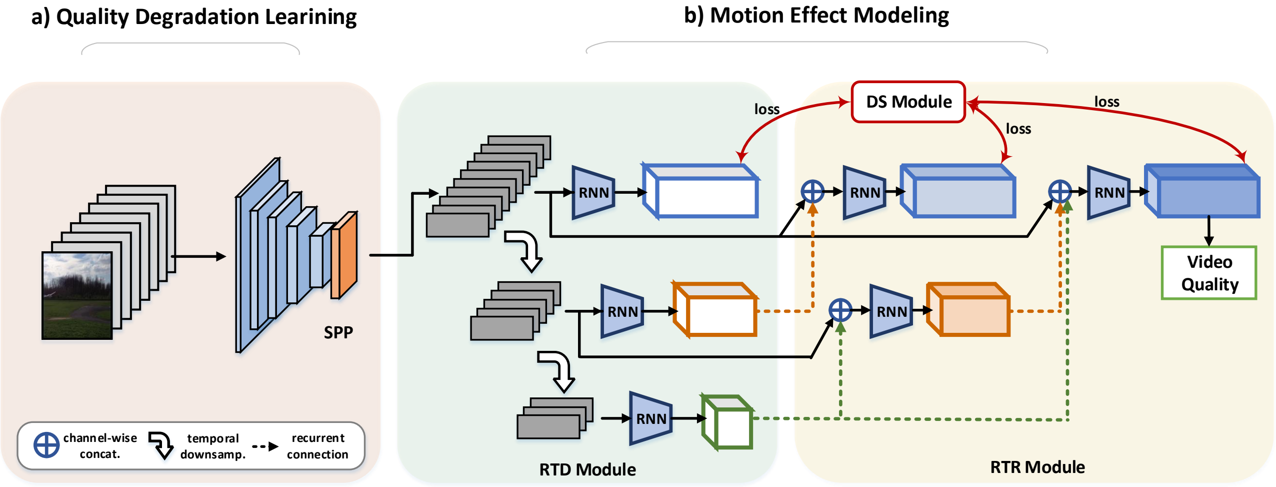

Figure 2: The overall structure of the RIRNet. The model consists of two parts: a) quality degradation learning sub-network for extracting distortion-aware features from the raw videos; b) motion effect modeling sub-network can further be divided into the RTD module for allocating the extracted feature vectors and performing temporal downsampling, the RTR module for fusing the multiple motion information with different temporal frequencies and DS module for tackling the problem of gradient fracture. Best viewed with color and zoom-in.

Existing NR-VQA metrics are mainly directed against the distortion-specific problems, such as rate adaptation and motion blur [3, 42]. These metrics demonstrate the advantages for the specific distortions, but not for other situations. The general-purpose method is another type of NR-VQA dealing with diversified distortions. Recently, benefiting from effective feature extraction algorithms, some successful general-purpose NR-VQA metrics have been proposed and shown promising performance. Saad *et al.* [30] proposed V-BLIINDS where a model in the discrete cosine transform (DCT) domain and a motion model that quantifies motion coherency were combined to predict video quality. Mittal *et al.* [23] proposed a metric called VIIDEO which models the intrinsic statistical regularities to quantify disturbances introduced by distortions. However, as the extension of the images in temporal dimension, videos are characterized exhaustively not only by spatial features but also by temporal ones, which leads to the failure when it comes to videos with more complicated spatio-temporal regularities for conventional general-purpose metrics.

With the advent of deep learning, extracting discriminative and semantic features automatically has come to reality. However, few deep learning-based NR-VQA metrics have emerged mainly due to the fact that conventional 2D-CNNs are not capable of processing the raw videos with three-dimensional spatio-temporal regularities. Notably, Li *et al.* [20] extracted 3D shearlet transform features of distorted videos to analyze natural scene statistics, then the features evolved by CNNs make the discriminative parts of the primary features exaggerated. Zhang *et al.* [44] applied weakly supervised learning with CNN and resampling strategy for VQA. Commonly, RNNs and 3D-CNNs are two widely-used approaches for dealing with the spatio-temporal information. Liu *et al.* [21] exploited the 3D-CNN model for codec classification and quality assessment of compressed videos. In [18], a NR-VQA method for in-the-wild videos by incorporating content-dependency and temporal-memory effects was validated. However, the performance and application scope of these algorithms are constrained without effective extraction of motion information in video quality perception, given the fact that they are executed at only a fixed temporal frequency. Therefore, it is highly desired to develop a general-purpose NR metric which could take advantage of motion information with different temporal frequencies.

3 RECURRENT-IN-RECURRENT NETWORK

In this section, we introduce the framework of the proposed RIRNet. To better match the motion processing in primate cortex, it starts with the quality degradation learning sub-network, and followed by the motion effect modeling sub-network which could be further disassembled into three parts, namely Recurrent Temporal Dimension (RTD) module, Recurrent Temporal Resolution (RTR) module and Deep Supervision (DS) module. Figure 2 illustrates the concept of the proposed model. Given the feature vectors calculated from the quality degradation learning sub-network, the motion effect modeling sub-network specializes in integrating multi-scale motion information from different temporal frequencies. The extracted features are downsampled in temporal dimension in the RTD module. The RTR module, coupled with the DS module which furnishes latent supervision in side-outputs, allows a hierarchical spatio-temporal distortion description and provides a coarse-to-fine video quality prediction.

3.1 Quality Degradation Learning

The goal of the quality degradation learning sub-network is to perform low-level motion processing and derive local motion information. The extracted distortion-aware features from individual frames will further be fed to the downstream sub-network. Encouraged by the positive results that deep semantic features obtained from a large number of training data achieved in helping address the content-dependency issues on predicted image quality [33, 41], we consider the expansion in the video field where the network can not only obtain information in a single static image, but also capture the sophisticated evolution along the temporal dimension. There are multiple options for the precise details of the extended connectivity in temporal dimension. Among three broad connectivity pattern categories [15] (early fusion, late fusion and slow fusion) which are distinguished by when to fuse the extracted features between different frames. In this work, the late fusion pattern is chosen to fuse the extracted features which enables direct employ of the pre-trained model in the image field and perform information fusion in the late stage.

Furthermore, a sparse sampling strategy instead of dense frames is adopted in this work in terms of the observation that successive frames are highly redundant for video quality assessment [39]. In detail, 4 frames are selected at equal intervals in each video segment lasting one second as the input to our network.

Assuming each input consists of T selected frames, we feed the video frames $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_T\}$ into a pre-trained backbone model and output the deep semantic feature maps $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ from its top convolutional layer:

$$\mathbf{X} = \mathrm{CNN}(\mathbf{V}). \tag{1}$$

Then we apply spatial pyramid pooling (SPP) [10] which aims to discard redundant information and produce a fixed-length feature vector for each feature map.

After that, given that the extracted feature vectors are of high dimension and not feasible for training, they have been dimensionally reduced through the fully connected (FC) layer, before being fed into temporal modeling sub-network:

$$\mathbf{T} = \mathbf{W}_{XT} \cdot \mathbf{X} + \mathbf{b}_{XT}, \tag{2}$$

where $\mathbf{W}_{XT}$ and $\mathbf{b}_{XT}$ are the parameters in the single FC layer.

3.2 Motion Effect Modeling

Motion effect modeling is another crucial clue for designing objective VQA models, integrating the obtained local motion information into a global percept. Differing from the majority of existing works which focus on specific temporal frequency [18, 44], the proposed scheme is designed to yield a more comprehensive motion representation through information at different temporal frequencies. The temporal modeling sub-network can be separated in two aspects. In the feature integration aspect, feature vectors with different temporal frequencies are assigned and aggregated by RTD module. In the motion refine aspect, the RTR module consisting of nested skip and recurrent connections is built to capture the fine-grained motion perception. Furthermore, the introduction of deep supervision provides additional supervision to solve the gradient back propagation problem when training. Figure 3 shows the high-level overview of the temporal modeling sub-network.

Recurrent Temporal Dimension Module. The RTD module aims to design a downsampling strategy in temporal dimension and perform feature integration for each frame in the current state based on retrospective ones. First, in order to obtain multiple motion information with different temporal frequencies, an allocation strategy is introduced to obtain feature sequences with different temporal frequencies. Formally, we consider the calculated feature vectors $\mathbf{T}$ as the first element of the feature vector sequence and we rename it as $\mathbf{F}^0$ which contains T feature vectors $\{\mathbf{f}_1^0, \mathbf{f}_2^0, ..., \mathbf{f}_T^0\}$, the rest elements can be calculated as:

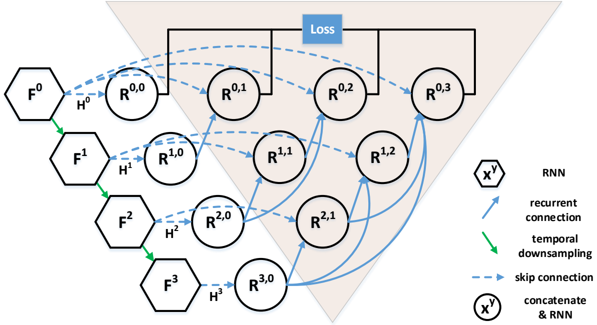

Figure 3: A high-level overview of the motion effect modeling sub-network. Since it is composed of three components, in order to distinguish them, the RTR module accompanied by the DS module is marked with a triangle shadow.

$$\mathbf{F}^{p+1} = \|\mathbf{F}^p\|, \tag{3}$$

where $\|\cdot\|$ denotes the uniformly downsampling operation. With repeating the downsampling operations N times, we can get the feature vector sequence $\mathbf{F} = \{\mathbf{F}^n, n = 0, 1, ..., N\}$.

In this paper, we consider applying RNN to integrate the extracted features. The current hidden state $\mathbf{h}_t^n$ whose initial values are $\mathbf{h}_0^n$ is calculated from the current input $\mathbf{f}_t^n$ and the previous hidden state $\mathbf{h}_{t-1}^n$:

$$\mathbf{h}_t^n = \mathrm{RNN}(\mathbf{f}_t^n, \mathbf{h}_{t-1}^n). \tag{4}$$

We then feed the assemble of the current hidden state $\mathbf{h}_t^n$ in all time steps, $\mathbf{H}^n$, to the RTR module, as the input of node $\mathbf{R}^{0,0}$.

Recurrent Temporal Resolution Module. Unlike previous works which focus temporal modeling on fixed temporal frequency, the RTR module is formulated to take full advantage of motion information with different temporal frequencies to get fine-grained motion perception related to the video quality. The realization of this function mainly depends on two kinds of connection: 1) Inspired by the anatomical evidences which have shown that recurrent synapses typically outnumber the feed-forward and feedback synapses in the neocortex [7], we establish recurrent connections between diverse temporal frequencies in a nested way, building a gradually deeper hierarchical structure; 2) moreover, to overcome the inherent restrictions during downsampling operations, skip connections are filled to recover the lost information, establishing a foundation for the coarse-to-fine prediction. We claim that with this framework shallower and deeper layers are combined to learn richer combinations that span more of the feature hierarchy.

Formally, we formulate the framework as follows. Let $\mathcal{R}^{i,j}$ denote the output of node $\mathbf{R}^{i,j}$ where i indexes the downsampling iterations along the temporal dimension and j indexes the number of the multi-scale motion information fused besides itself. The stack of feature vectors represented by $\mathcal{R}^{i,j}$ $(j > 1)$ is computed as:

$$\mathcal{R}^{i,j} = \mathcal{N}\left[\mathbf{H}^{i}, \mathcal{P}^{i,j}\right], \tag{5}$$

where function $\mathcal{N}(\cdot)$ is implemented by a RNN block, and $[\cdot]$ denotes the concatenation layer. The node $\mathbf{R}^{i,j}$ receive the information $\mathbf{H}^{i}$ from the same temporal frequency through the skip connections, and the information from lower frequencies $\mathcal{P}^{i,j}$ through the recurrent connections, where the $\mathcal{P}^{i,j}$ could be generated as followed:

$$\mathcal{P}^{i,j} = \left[C(\mathcal{R}^{i+k,j-k})\right]_{k=1}^{j}, \tag{6}$$

where $C(\cdot)$ denotes the recurrent connection composed of a frame-wise concatenation followed by a 1×1 convolutional layer to ensure dimensional invariance. The forming framework enables the network to more effectively capture fine-grained motion information in video sequences, where motion contents in higher temporal frequencies are gradually enriched by the corresponding motion contents whose temporal frequencies are relatively lower.

After the final aggregated feature vectors $\mathcal{R}^{0,N}$ are available, with the last element of the sequence, $\mathbf{h}_T^N$, we end up with a single FC layer for calculating the overall video quality score Q:

$$Q = \mathbf{W}_{hQ} \cdot \mathbf{h}_T^N + \mathbf{b}_{hQ}, \tag{7}$$

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

Deep Supervision Module. Considering the inevitable gradient transfer fracture in Figure 3, the introduction of deep supervision [17] enables to operate integrated direct supervision to each side-output, rather than the standard approach of providing supervision only at final output. The hypothesis is that with the help of this structure, the network makes the back propagation of gradient possible, and the optimizer would face an easier optimization problem when the additional supervised layers weakens the gradient vanishing problem by preserving gradients from early stage.

Owing to the designed structure, the designed structure is capable of fusing motion information at multiple temporal frequencies as $\mathcal{R}^{0,k}$ $(k \in \{0, 1, ..., N\})$, which are amenable to deep supervision. Given the input distorted video Y, we describe the loss function of the deep supervision as a weighted summation of the losses of several side-outputs, and the overall loss for prediction $\widehat{Y}$ can be calculated as:

$$\mathcal{L}(Y, \widehat{Y}) = \sum_{m=1}^{N+1} \alpha_m \mathcal{L}_m + \beta \mathcal{L}_{out}, \tag{8}$$

where $\mathcal{L}_m$ corresponds to the loss between the m_{th} side-output and the label, and α_m assigns each component its own weight coefficient. $\mathcal{L}_{out}$ corresponds to the loss associated with the final output layer. β is a hype-parameter to balance two losses and is tuned on a held-out validation set.

4 EXPERIMENTAL RESULTS

4.1 Experimental Protocols

Database. In the experiments, we relied on 4 subjective video databases which contain a large variety of distortion types. They can be further divided into two categories, LIVE VQA and CSIQ VQA are composed of videos with artificial distortion, while the contents in KoNViD-1k database and CVD2004 database suffer from natural distortion.

LIVE Video Quality database [32]. The database contains 160 videos divided into 10 groups with a resolution of 768×432. Each group contains one reference video and its corresponding 15 distorted videos whose length are 10s. 15 distorted videos in each group are generated from four different distortions: wireless distortions, IP distortions, H.264 compression, and MPEG-2 compression.

CSIQ Video Quality database [38]. This database contains 12 reference videos and 216 distorted videos generated from 6 distortion types: H.264/AVC compression, H.264 video with packet loss rate, MJPEG compression, Wavelet compression, White noise and HEVC compression.

KoNViD-1k database [12]. This database aims at natural distortions. It comprises a total of 1,200 videos of resolution 960×540 that are fairly filtered multiple steps from a large public video dataset, YFCC100M. The videos are 8s long with 24/25/30fps. The MOS ranges from 1.22 to 4.64.

CVD2014 video database [26]. This database aims at complex distortions introduced during video acquisition. It contains 234 videos of resolution 640×480 or 1280×720. 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.

Evaluation criteria. Two common performance criteria to evaluate the performance of the proposed metric are employed in our experiments: the Pearson linear correlation coefficient (PLCC) to evaluate the accuracy of the prediction results, and Spearman rank-order correlation coefficient (SRCC) for measuring the monotonicity of the results. It is noted that higher index means better prediction effect, and a well-performing quality assessment method is expected to deliver PLCC, SRCC values close to 1. Considering the inconsistency of the scale between objective predictions and the subjective scores, we refer to the suggestion of Video Quality Experts Group (VQEG) [37] before calculating PLCC values, and adopt a four-parameter logistic function for mapping the objective score to the subjective score as outlined in [14].

4.2 Implementation Details

Full-resolution input. Since perceptual quality is sensitive to the variation of input scale, our proposed approach supports by design arbitrary input size with the introduction of the SPP [10] layer which enables the network to get rid of the fixed-size input constraint and output the fixed-length feature vector for varying input sizes (21 in this work).

Training. We choose ResNet-50 [11] pre-trained on ImageNet [8] as the backbone network. For recurrent layer, we resort to GRU [6] which is a recurrent neural network model with gates control. In order to speed up the convergence of the optimization scheme, several RNNs with different sizes (corresponding to the sequence length) are individually trained and saved before being fine-tuned together. The proposed model is implemented with PyTorch [27]. The Adam [16] optimizer with an initial learning rate $1e-4$ is deployed for minimizing the ℓ_2 loss, and every 40 epochs the learning rate could be decayed by a factor of 0.2. We set the training batch size as 16 for training our model. To provide a fair comparison for all comparison methods, all tests are carried out on a computer with a E5-2630 CPU, 12G NVIDIA TITAN Xp GPU and 64 GB RAM.

Table 1: Quantitative results of different methods on four publicly available databases. Larger PLCC and SRCC indicate better performance and the best results are marked in boldface.

Database	Criterion	BRISQUE [22]	NIQE [24]	V-BLIINDS [30]	V-CORNIA [43]	VIIDEO [23]	VSFA [18]	RIRNet(Ours)
LIVE VQA [32]	PLCC	0.4170	0.3981	0.7482	0.7440	0.6518	0.7278	**0.8091**
	SRCC	0.4008	0.3476	0.7244	0.7324	0.6240	0.7001	**0.7828**
CSIQ VQA [38]	PLCC	0.5338	0.3696	0.7710	0.7533	0.5447	0.7816	**0.8426**
	SRCC	0.5245	0.3599	0.7843	0.7721	0.4906	0.7980	**0.8574**
KoNViD-1k [12]	PLCC	0.5896	0.4105	0.6273	0.6518	0.3058	0.7391	**0.7812**
	SRCC	0.6254	0.3782	0.6158	0.6795	0.3412	0.7452	**0.7755**
CVD2014 [26]	PLCC	0.6157	0.3981	0.7222	0.6716	0.2083	0.8277	**0.8780**
	SRCC	0.6298	0.4995	0.7068	0.6583	0.1544	0.8431	**0.8891**

4.3 Performance Measure and Comparison

In this paper, seven popular objective NR-I/VQA metrics (BRISQUE [22], NIQE [24], V-BLLINDS [30], V-CORNIA [43], VIIDEO [23] and VSFA [18]) are chosen for comparison. Experiments on each database are processed by k-fold ($k = 5$) cross-validation, ensuring the training sets and testing sets are not overlapped in content. This procedure is repeated 5 times and the average values of PLCC and SRCC results across all repetitions for the mentioned competitors and the proposed algorithms are given in Table 1. The best results on each database are marked in boldface. To be noted, IQA metrics (BRISQUE, NIQE) are performed frame-by-frame of the video and the overall index is computed as the average of the frame level quality scores. Among all the metrics, NIQE and VIIDEO are training-free models while others require training. Support vector regressor (SVR) [34] is employed to learn the mapping from their feature spaces to the ground truth for conventional metrics. All experiments are conducted under the same conditions. Note that VSFA is the only open-source NR deep learning-based metric available here, other metrics which need partial information from the references [44] or have not been released [20, 21] are not feasible for comparison.

It can be observed from Table 1 that the proposed metric achieves the best performance both in predicting accuracy (PLCC) and monotonicity (SRCC) on all databases, which confirms our model to be generalize to diverse impairments. It is not surprisingly that IQA metrics (BRISQUE and NIQE) inevitably perform poorly since they do not consider the temporal relation between frames and treat them independently. Among four benchmarks, KoNViD-1k [12] is the most challenging one due to the diversified environments. However, our model yields substantially better quantitative results (5.70% in terms of PLCC) than the second-best method (VSFA) on it, which is attributed to the way we designed the temporal modeling scheme, *i.e.*, fusion of motion information with different temporal frequencies to achieve a hierarchical motion perception. Besides, considering the case without VSFA and comparing our metric with the rest conventional methods, it is worth noting that compared with the artificial distortions, the proposed method possesses a greater advantage than the conventional methods in the authentic distortion databases. Specifically, contrast to the best performing conventional method, 19.85% improvements on KoNViD-1k (*vs.* V-CORNIA) and 21.57% improvements on CVD2014 (*vs.* V-BLIINDS) are achieved, compared with 8.14% increase on LIVE VQA (*vs.* V-BLIINDS) and 9.29% increase on CSIQ VQA (*vs.* V-BLIINDS). This confirms that the data-driven feature representation is equally effective for VQA tasks, especially for the more complicated situation of natural distortion where conventional methods cannot perform effective feature extraction.

4.4 Cross-database Testing

The generalization ability measuring the performances when coming across the unprecedented videos is of vital importance for a well-performing VQA model. In this section, we implement a different experiment scheme to demonstrate the robustness of the proposed RIRNet. Specifically, among four participating databases, the quality prediction model is trained on one of them and tested on the other three for each test. To further make a comprehensive comparison, other four competing learning-based models are also included. The validation PLCC results are listed in Table 2.

Compared with the conventional metrics, the deep models achieve more robust performances, ascribing to the eminent feature learning ability of the DNNs. Without the training samples from the target databases, the proposed method could still be in the leading position among the comparison methods (8 times the best performance in all 12 tests, far ahead of the 3 times achieved by the second place model, VSFA), confirming the well adaptability of the proposed RIRNet to unknown samples due to the integrity description of spatiotemporal distortion. We can thus conclude that the proposed model do not depend on the databases.

4.5 Ablation Study

To evaluate the contribution of each component in the proposed model, we conduct ablation experiments which begins with Quality Degradation Learning (QDL) sub-network, followed by the Motion Effect Modeling (MEM) sub-network. Table 3 presents the experimental results.

QDL sub-network. We train our model from scratch when we remove the features extracted from the pre-trained model. As is evident from the results, the removal of the pre-trained model is accompanied by significant performance drop in all four databases (0.1998, 0.2202, 0.1471 and 0.1878 respectively in PLCC), which verifies that pre-trained model is a force to be reckoned with in the successful extraction of quality-related features. Note that the

Table 2: Cross-database validation of the proposed metric with the competing learning-based models. The best performances are marked in boldface.

Training database	Test database	BRISQUE [22]	V-BLIINDS [30]	V-CORNIA [43]	VSFA [18]	RIRNet(Ours)
LIVE VQA [32]	CSIQ VQA [38]	0.2973	0.4960	0.4736	0.5987	**0.6230**
	KoNViD-1k [12]	0.1985	0.2738	0.2950	**0.4022**	0.3953
	CVD2014 [26]	0.2551	0.2812	0.3389	**0.4903**	0.4788
CSIQ VQA [38]	LIVE VQA [32]	0.3243	0.5590	**0.5846**	0.5822	0.5771
	KoNViD-1k [12]	0.1804	0.2976	0.3144	0.3690	**0.3721**
	CVD2014 [26]	0.1623	0.2714	0.2875	0.4395	**0.4508**
KoNViD-1k [12]	LIVE VQA [32]	0.3980	0.4913	0.4682	0.6380	**0.6621**
	CSIQ VQA [38]	0.3761	0.6318	0.6220	0.7170	**0.7354**
	CVD2014 [26]	0.4360	0.5325	0.5980	0.6958	**0.7712**
CVD2014 [26]	LIVE VQA [32]	0.1742	0.4356	0.3882	0.5760	**0.6033**
	CSIQ VQA [38]	0.4283	0.5912	0.6174	**0.7208**	0.7126
	KoNViD-1k [12]	0.2954	0.5381	0.5244	0.5920	**0.6454**

Table 3: Alation study of different sub-networks. Performance measured by PLCC of our metric equipped with different components on all four databases for ablation study.

Sub-net		Database			
QDL	MEM	LIVE	CSIQ	KoNViD	CVD2014
×	✓	0.6093	0.6224	0.6341	0.6902
✓	×	0.7160	0.6932	0.6674	0.7135
✓	✓	**0.8091**	**0.8426**	**0.7812**	**0.8780**

Table 4: Alation study of different components of the QDL sub-network. Performance measured by PLCC of our metric equipped with different components on all four databases for ablation study.

Components of QDL			Database			
RTD	RTR	DS	LIVE	CSIQ	KoNViD	CVD2014
×	×	×	0.6860	0.7332	0.6576	0.7335
✓	×	×	0.7121	0.7695	0.6901	0.7633
✓	✓	×	0.8075	0.8307	0.7754	0.8716
✓	✓	✓	**0.8091**	**0.8426**	**0.7812**	**0.8780**

other pre-trained model that is available and potentially helpful can be readily deployed as the backbone network but it is beyond the scope of this work.

MEM sub-network. Then the effectiveness of the temporal modeling sub-network is measured by simply decoding the extracted features with RNNs (equivalent to the output of $\mathcal{R}^{0,0}$). As can be witnessed that the TM sub-network boosts the performance of video quality assessment by 0.1231, 0.1094, 0.1236 and 0.1445 in terms of PLCC on the four databases, respectively. They are indeed substantial improvements considering the progresses reported in recent years by state-of-the-art methods on VQA tasks (refer to Table 1).

Moreover, the MEM sub-network is further composed of three parts, *i.e.*, the RTD module, the RTR module and the DS module.

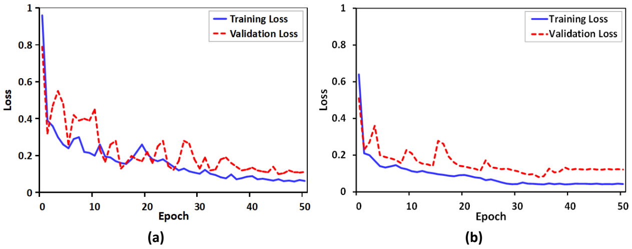

Figure 4: The loss curves in LIVE VQA database [32] (a) without DS module and (b) with DS module.

To further investigate the contribution from each of them to the whole system, we then incrementally augment the system with each individual module. The results presented in Table 4 show that the RTR module brings about the major performance boost (78.80%, 66.82%, 73.71% and 79.38%) while the RTD module yields another minor gain on four databases, which confirms the validity of the proposed scheme for making full use of motion information with different temporal frequencies. Although the effect of the DS module on the overall performance is not obvious, it has played a positive role in the convergence of the network as can be observed in Figure 4.

4.6 Qualitative Evaluation of RIRNet

To gain more insight into how the structure of RIRNet gradually improves the prediction, we also present examples to qualitatively evaluate the proposed approach, and include the single temporal frequency baseline (equivalent to the output of $\mathcal{R}^{0,0}$) for comparison. The visual results combined with their class activation maps (CAMs) [45], which vividly exhibit the intensity distribution of gradient changes, are exhibited in Figure 5.

Compared to the single temporal frequency baseline (indicated by the blue box), the proposed RIRNet (indicated by the red box) is observed to output more precise quality predictions, benefiting from the designed recurrent-in-recurrent temporal modeling mechanism. Take Figure 5 (a) for an example, the single temporal frequency

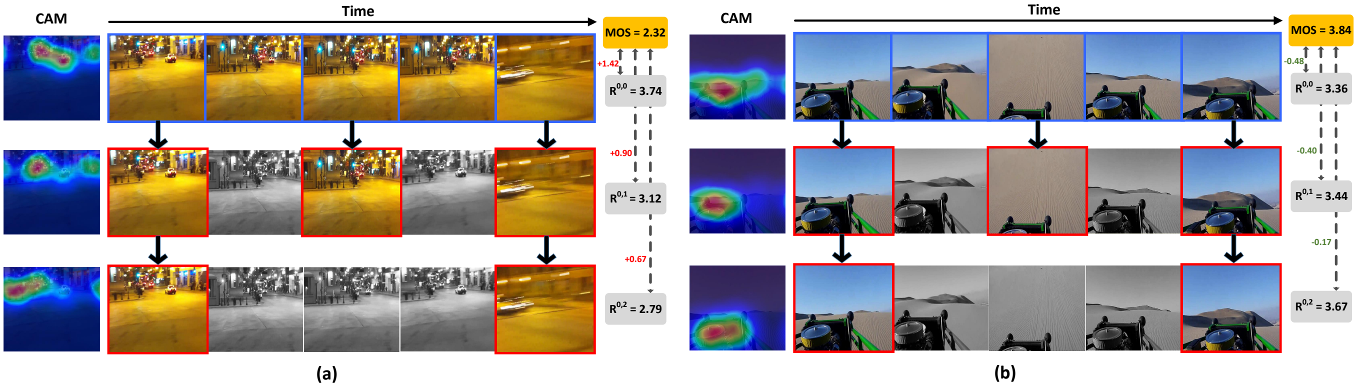

Figure 5: Qualitative comparison between the basis RNN model and our RIRNet by examples from KoNViD-1k [12] database. For each example, the sampled frames are processed by the proposed RIRNet, and output of $\mathcal{R}^{0,0}$ is equivalent to the the basis RNN model (indicated by the blue box). We also give the class activation map (CAM) [45] of the first frame in each video sequence to better illustrate the effectiveness of our method.

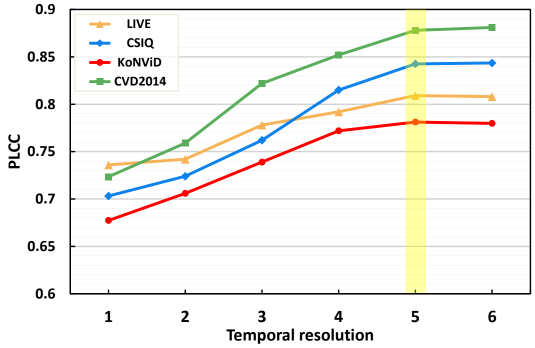

Figure 6: Impact of temporal resolution on quality prediction. The downsampling iterations employed in our implementation is five whose performances are marked in yellow.

baseline provides an unacceptable prediction for the video (3.74, compared with the MOS 2.32). On the contrary, as multiple motion information with different temporal frequencies is taken into account by the RIRNet, more accurate quality prediction is achieved by degrees, as can be witnessed in the figure (from 3.74 to 3.12 to 2.79). Simultaneously, from the given CAMs, we could observe that the focus of the network gradient change gradually shifts to the position with stronger distortion in the proposed model.

4.7 Impact of Temporal Resolution

To illustrate the effectiveness of motion representation based on multi temporal frequencies and verify whether the accumulation of more temporal resolutions is beneficial for video quality prediction, the models are trained with various numbers of temporal resolutions (ranging from 1 to 6) in this section. Figure 6 exhibits the experimental results on all the four databases.

There is an obvious performance gap between the models as can be witnessed. This indicates that the designed temporal modeling scheme for which the network is originally trained for does affect the performance. Therefore, the fact that motion representation with multi temporal frequencies is highly needed for effective evaluation of video quality has been confirmed. Moreover, all curves turn out incrementally trend up to a certain number of iterations (5 in this work), but tend to be saturated or even slightly declining after that (except for CVD2014 due to the large video length distribution). In other words, more accumulated temporal resolutions do not necessarily lead to consistently better results. Therefore in our implementation, 5 different scales are employed to build the comprehensive motion representation considering both the accuracy and computational consumption.

5 CONCLUSION

In this paper, we propose the RIRNet for NR-VQA task building upon insights of the motion perception in primate cortex. By jointly exploiting multiple motion information with different temporal frequencies when performing high-level motion processing, the proposed framework encodes the extracted local motion information into powerful motion representation that is closely related to the sophisticated spatio-temporal distortions in source videos. It turned out the proposed model enables a coarse-to-fine quality prediction from a human perception point of view. We show the superior performance of the RIRNet both quantitatively and qualitatively on four publicly available video quality databases. It is expected that the proposed framework also has a potential for other tasks such as video segmentation which require effective and hierarchical motion perception.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 61771473, 61991451 and 61379143, the Key Project of Shanxi Provincial Department of Education (Collaborative Innovation Center) under Grant 20JY024, the Science and Technology Plan of Xi'an under Grant 20191122015 KYPT011JC013, the Natural Science Foundation of Jiangsu Province under Grant BK20181354, and the Six Talent Peaks High-level Talents in Jiangsu Province under Grant XYDXX-063.

REFERENCES

[1] S. Bae and T. Lee. 2011. Product type and consumers' perception of online consumer reviews. *Electronic Markets* 21, 4 (2011), 255–266.

[2] C. G. Bampis, Z. Li, and A. C. Bovik. 2018. Spatiotemporal feature integration and model fusion for full reference video quality assessment. *IEEE Trans. Circuits Syst. Video Technol.* 29, 8 (2018), 2256–2270.

[3] T. Brandão and M. P. Queluz. 2010. No-reference quality assessment of H. 264/AVC encoded video. *IEEE Trans. Circuits Syst. Video Technol.* 20, 11 (2010), 1437–1447.

[4] P. Chen, L. Li, Y. Huang, F. Tan, and W. Chen. 2019. QoE Evaluation for Live Broadcasting Video. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 454–458.

[5] P. Chen, L. Li, X. Zhang, S. Wang, and A. Tan. 2019. Blind quality index for tone-mapped images based on luminance partition. *Pattern Recognit.* 89 (2019), 108–118.

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[7] P. Dayan, L. F. Abbott, and L. Abbott. 2001. Theoretical neuroscience: computational and mathematical modeling of neural systems. (2001).

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, 248–255.

[9] S. Gehlen. [n.d.]. Experts agree: live streaming video for brands is one of 2016's biggest marketing trends. [Online]. Available: www.yourbrandlive.com/blog/live-streaming-video-marketingtrends-2016. Accessed: Nov.5, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 9 (2015), 1904–1916.

[11] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, 770–778.

[12] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szir¢nyi, S. Li, and D. Saupe. 2017. The Konstanz natural video database (KoNViD-1k). In *Proc. Int. Conf. Quality Multimedia Exper. (QoMEx)*. IEEE, 1–6.

[13] B. Hu, L. Li, and J. Qian. 2018. Internal Generative Mechanism Driven Blind Quality Index for Deblocked Images. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 2476–2480.

[14] ITU. 2011. Recommendation ITU-R BT.500.13: Methodology for the subjective assessment of the quality of television pictures. *International Telecommunications Union* (2011).

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. IEEE, 1725–1732.

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

[17] C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. 2015. Deeply-supervised nets. In *Artificial intelligence and statistics*. 562–570.

[18] D. Li, T. Jiang, and M. Jiang. 2019. Quality Assessment of In-the-Wild Videos. In *Proc. ACM Int. Conf. Multimedia (ACM MM)*. ACM, 2351–2359.

[19] L. Li, Y. Zhou, K. Gu, W. Lin, and S. Wang. 2017. Quality assessment of DBIR-synthesized images by measuring local geometric distortions and global sharpness. *IEEE Trans. Multimedia* 20, 4 (2017), 914–926.

[20] Y. Li, L. Po, C. Cheung, X. Xu, L. Feng, F. Yuan, and K. Cheung. 2015. No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.* 26, 6 (2015), 1044–1057.

[21] W. Liu, Z. Duanmu, and Z. Wang. 2018. End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks. In *Proc. ACM Int. Conf. Multimedia (ACM MM)*. ACM, 546–554.

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 21, 12 (2012), 4695–4708.

[23] A. Mittal, M. A. Saad, and A. C. Bovik. 2015. A completely blind video integrity oracle. *IEEE Trans. Image Process.* 25, 1 (2015), 289–300.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik. 2012. Making a ¡°Completely Blind¡± Image Quality Analyzer. *IEEE Signal Process. Lett.* 20, 3 (2012), 209–212.

[25] J. A. Movshon and W. T. Newsome. 1996. Visual response properties of striate cortical neurons projecting to area MT in Macaque monkeys. *J. Neurosci.* 16, 23 (1996), 7733–7741.

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

[27] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. 2017. Automatic differentiation in pytorch. (2017).

[28] J. A. Perrone and A. Thiele. 2001. Speed skills: Measuring the visual speed analyzing properties of primate MT neurons. *Nature Neurosci.* 4, 5 (2001), 526–532.

[29] N. J. Priebe, S. G. Lisberger, and J. A. Movshon. 2006. Tuning for spatiotemporal frequency and speed in directionally selective neurons of macaque striate cortex. *J. Neurosci.* 26, 11 (2006), 2941–2950.

[30] M. A. Saad, A. C. Bovik, and C. Charrier. 2014. Blind prediction of natural video quality. *IEEE Trans. Image Process.* 23, 3 (2014), 1352–1365.

[31] K. Seshadrinathan and A. C. Bovik. 2009. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. Image Process.* 19, 2 (2009), 335–350.

[32] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. 2010. Study of subjective and objective quality assessment of video. *IEEE Trans. Image Process.* 19, 6 (2010), 1427–1441.

[33] E. Siahaan, A. Hanjalic, and J. A. Redi. 2018. Semantic-aware blind image quality assessment. *Sig. Process.: Image Commun.* 60 (2018), 237–252.

[34] A. J. Smola and B. Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.

[35] Facebook Live Statistics. [n.d.]. Facebook Video Statistics. [Online]. Available: http://mediakix.com/2016/08/facebook-video-statistics-everyone-needs-know. Accessed: Nov.5, 2019.

[36] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro. 2011. No-reference pixel video quality monitoring of channel-induced distortion. *IEEE Trans. Circuits Syst. Video Technol.* 22, 4 (2011), 605–618.

[37] VQEG. 2000. FINAL REPORT FROM THE VIDEO QUALITY EXPERTS GROUP ON THE VALIDATION OF OBJECTIVE MODELS OF VIDEO QUALITY ASSESSMENT March 2000. (2000).

[38] Phong V Vu and Damon M Chandler. 2014. ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging* 23, 1 (2014), 013016.

[39] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 20–36.

[40] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin. 2019. Quality Assessment for Video with Degradation Along Salient Trajectories. *IEEE Trans. Multimedia* 21, 11 (2019), 2738–2749.

[41] J. Wu, J. Zeng, W. Dong, G. Shi, and W. Lin. 2019. Blind image quality assessment with hierarchy: Degradation from local structure to deep semantics. *J. Vis. Commun. Image Represent.* 58 (2019), 353–362.

[42] Q. Wu, H. Li, F. Meng, and K. N. Ngan. 2018. Toward a blind quality metric for temporally distorted streaming video. *IEEE Trans. Broadcast.* 64, 2 (2018), 367–378.

[43] J. Xu, P. Ye, Y. Liu, and D. Doermann. 2014. No-reference video quality assessment via feature learning. In *Proc. IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 491–495.

[44] Y. Zhang, X. Gao, L. He, W. Lu, and R. He. 2018. Blind video quality assessment with weakly supervised learning and resampling strategy. *IEEE Trans. Circuits Syst. Video Technol.* 29, 8 (2018), 2224–22255.

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. Learning deep features for discriminative localization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, 2921–2929.

[46] Y. Zhou, L. Li, S. Wang, J. Wu, and Y. Zhang. 2018. No-reference quality assessment of DIBR-synthesized videos by measuring temporal flickering. *J. Vis. Commun. Image Represent.* 55 (2018), 30–39.