Video Quality Assessment with Serial Dependence Modeling

Yongxu Liu, Jinjian Wu, Aobo Li, Leida Li, Weisheng Dong, Guangming Shi, *Fellow, IEEE*, Weisi Lin, *Fellow, IEEE*

***Abstract*—Video quality assessment (VQA) is much more challenging than image quality assessment, due to the difficulty of modeling temporal influence among frames. Most of the existing VQA methods usually isolate each moment within the video (i.e., it neglects the sequential nature), leading to a large gap from the subjective perception. Recent research on neuroscience suggests a serially dependent perception (SDP) mechanism in the human visual system (HVS). Namely, the HVS tends to incorporate the recent past visual experience to predict the present perception. Inspired by the SDP, we suggest that the HVS prefers stable and continuous degradations in videos due to their predictability, and exhibits less tolerance to interrupted and unpredictable disturbances. Thus, we introduce a novel serial dependence modeling (SDM) framework for full-reference VQA in this paper. Firstly, the instantaneous degradation is measured on both the static appearance and motion information for each glimpse of scenes. Since motion plays an important role in videos, two types of structures are extracted for motion representation, namely, an explicit content-based 3D structure and an implicit feature-based 2D structure. Next, an assessment-directed long-short term memory (A-LSTM) is proposed to capture the serial dependence among instantaneous degradations. With the consideration of the perceptual effect from the previous moment on the current one, especially the effect from the perceptually worst moment, the serially dependent degradation is characterized. Finally, by mimicking the subjective rating for video-viewing, an attention-based quality decision procedure is presented to acquire the final video quality. Experimental results on publicly available VQA databases demonstrate that the proposed method maintains good consistency with the subjective perception.**

***Index Terms*—Video Quality Assessment, Serial Dependence, Long-Short Term Memory, Attention**

I. Introduction

Digital videos have been incrementally active, and video-based applications have enriched our lives. Videos with high quality provide a stunning visual experience for human. However, under a limited condition, capturing, compression, delivery, and post-processing on videos cause degradations inevitably. Hence, reliable video quality assessment (VQA) is required for quality control to bring about a better subjective visual experience [1].

Yongxu Liu, Jinjian Wu, Aobo Li, Leida Li, Weisheng Dong, and Guangming Shi are with School of Artificial Intelligence, Xidian University, Xi'an, China. E-mail: yongxu.liu@stu.xidian.edu.cn; jinjian.wu@mail.xidian.edu.cn. (*Corresponding author: Jinjian Wu.*)

Weisi Lin is with the School of Computer Engineering, Nanyang Technological University, Singapore.

This work was partially supported by the National Key R&D Program of China (2018AAA0101400) and the National Natural Science Foundation of China (No. 62022063, No. 61772388 and No. 61632019).

When the human is the ultimate recipient, subjective ratings are the most preferred evaluation. However, due to the high cost and low efficiency, subjective VQA is impractical in real-world in-service applications, and this encourages researchers to seek suitable objective computational models that estimate the quality automatically and perform consistently in line with the human perception. Considering the full reference information for quality estimation, full-reference (FR) VQA [2, 3] can lead to more reliable prediction compared with reduced-reference (RR) VQA [4, 5] and non-reference (NR) VQA [6, 7], thus having attracted growing attention for research.

Due to the importance of motion in videos, during the last decade, much effort has been made to characterize the motion information and its degradation on VQA. In [8], a block-based motion estimation was incorporated into the structural similarity index measurement (SSIM), and the motion was directly employed as a weighting factor to tune the similarity score in motion regions. In [4, 9], the adjacent frame difference was treated as temporal information, and based on the hypothesis of Gaussian scale mixture distribution, the entropic differencing of frame difference was computed to measure the motion degradation. Further, with the natural 3D structure of video frames, 3D-filters [10, 11], 3D-gradient [12], and 3D-DCT [13] were used to represent the motion through a space-time decomposition or transformation, and the coefficients were used for degradation analysis. And heuristically, the video was measured as spatiotemporal slices in [14], and the 3D structure is viewed as multiple transformed 2D planes. In [15], through exploring the relation between optical flow and distortion contamination, the motion degradation in videos was measured by the variation of optical flow statistics. And recently, based on the observation that the human visual system (HVS) presents a continuous pursuit on moving objects, both motion velocity and motion content along motion trajectory were used for the dynamic degradation representation [16]. Although great improvements have been made, there's still a large gap between objective FR-VQAs and the human perception.

As convolutional neural networks (CNNs) have greatly improved many image/video-based tasks [17], some recent work tries to tackle the representation problem with a deep learning-based method. A feature-based transfer learning framework was employed on the basis that the features of images and videos shared a related latent subspace [18], and the learned knowledge from the task of image quality assessment (IQA) was transfered to VQA. In [19], a spatiotemporal sensitivity map was learned to modulate the error map, thus predicting

the frame-by-frame quality, and a similar sensitivity-learning step was employed along the temporal axis to get the video quality. And recently, a combination of 2D and 3D CNN was proposed to learn both the spatial and spatiotemporal features, and the residual frames were masked for further quality estimation [20]. However, these models lack of evident psychophysical basis, and try to directly fit the perceptual target with a straightforward data-driven method, which greatly weakens the particularity of subjective perception and limits the performance.

As the great challenge remains for current computational models on VQA, we trace back the problem to its related phenomenon in human visual perception. When a video displays, the HVS captures stimuli continuously and reacts constantly, creating a stable and continuous visual experience over time. Recent research in neuroscience reveals that this visual stability and contiguity are gained through a mechanism of serially dependent perception (SDP) [21]. It is found that there exists a continuity field in both the low-level and high-level of the HVS [22]. With such a continuity field, visual perception is serially dependent. The HVS uses both the recent and present inputs to inform the perception at the present moment, and this can bridge the temporal disruptions and facilitate the visual stability [23, 24].

The mechanism of SDP suggests two aspects in subjective perception: the sequential characteristic and the dependency relationship [23]. However, most of the existing VQA methods usually ignore the sequential nature and isolate each moment within the video. Some work calculates the frame-by-frame quality and set the average or Minkowski sum as the overall quality [25, 26], which totally discards the sequential characteristic. Since frames with lower quality have a noteworthy influence on perception, the work in [27] adopts a percentile average pooling to only concentrate some small portion of the whole frame-wise quality. The work in [28] classifies the frame-wise quality into two categories(i.e., higher quality and lower quality) with k-means clustering, and obtain the final video quality with adaptively weighting. Nevertheless, both of the methods still lack of sequential characterization and dependency modeling. The work in [29] observed a hysteresis effect in VQA through a subjective experiment. However, it lacks a solid theory to interpret this observation, and the built model with a discrete score-sorting fails to explore the internal sequential relation within the video. Some recent work directly employs the recurrent neural network in VQA [30, 31]. Although the sequential characteristic is captured roughly, the dependency relationship is not thoroughly explored. An overview of these temporal pooling methods can be found in [32].

Motivated by the SDP, in this work, we suggest modeling the problem of VQA in a serially dependent way. Intuitively, we suggest that under the same distortion level, the HVS prefers the degradation with more contiguity and stability due to its serial dependence and predictability (e.g., Motion JPEG compression distortion), but resists the unpredictable and interrupted one (e.g., packet loss). Hence, we propose a novel serial dependence modeling (SDM)-based framework for FR-VQA, which contains three stages: snippet-based instantaneous degradation estimation, SDM-based serially dependent degradation characterization, and attention-based quality decision. Firstly, considering the importance of motion in videos, two types of structures (i.e., an explicit content-based 3D structure and an implicit feature-based 2D structure) are extracted to represent motion information. The 3D structure is obtained by stacking video frames, and the 2D structure is formed from features among frames, thus capturing motion information in both content-based and feature-based domains. By taking both the static appearance and motion information into account, the instantaneous degradation is estimated for each video snippet, which reflects the pristine feeling at every moment of video. Next, towards the specific VQA problem, a recall procedure is introduced into the long-short term memory (LSTM), and an assessment-directed LSTM (A-LSTM) is proposed to capture the strong serial dependence among instantaneous degradations. With a consideration of the perceptual effect from the previous moment on the current one, especially the effect from the perceptually worst moment, the serially dependent degradation is characterized, and the VQA problem is modeled as a serially dependent process. Finally, by mimicking the behavior of subjective rating for video-viewing, an attention-based quality decision procedure is introduced to highlight the perceptually worst moment and integrate the time-varying perceptual degradations into the final quality score.

The main contributions of this work are as follows.

1) Driven by SDP, we find that the sequential influence is important for VQA, and the HVS prefers a stable time-varying degradation due to its predictability, but resists the interrupted and unpredictable disturbances.
2) A novel framework focusing on SDM is therefore proposed, which explores the sequential influence of time-varying degradation, and VQA is regarded as a serially dependent problem.
3) We also propose a task-specific A-LSTM cell to capture the dual dependency from both the last moment and the perceptually worst moment in the previous, as well as an attention-based quality decision procedure to adaptively modulate the video quality.

The rest of this paper is organized as follows. In Section II, a further explanation of SDP is given. The detailed implementation of the proposed SDM is introduced in Section III. In Section IV, the effectiveness of the proposed method is demonstrated. And we draw the conclusion of this paper in Section V.

II. Serial dependence perception

In this section, a brief introduction about the serial dependence in subjective perception is first given. Then according to the SDP, we discuss its relation to the quality assessment task, thus guiding the VQA modeling.

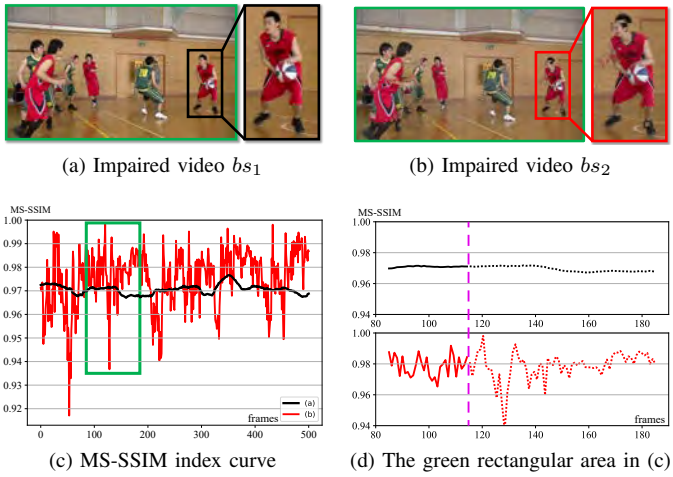
The well-perceived scenes in video by the HVS are always stable and continuous, and the explanation about this stability is SDP. Research shows that the brain builds a continuity field, with which the human perception is serially dependent and promotes the visual stability over time [21, 23]. Serial

dependence during human perception indicates that, previously relevant information can linger and impact what we currently perceive. Since the recent visual past is typically a good indicator of the future [24], the brain always attempts to use this previous information to interpret the current visual input in different perceptual levels [21, 22, 33], thus creating a stable and continuous perception over time.

The serial dependence suggests two aspects in subjective perception: the sequential characteristic and the dependency relationship [23]. The sequential characteristic indicates the importance of the order each moment presents along the time axis, and the dependence contributes to the predictability. Based on the sequential characteristic and dependency relationship when perceiving dynamic stimuli, with both the recent and present visual inputs to inform the current perception, the HVS facilitates a continuous and stable visual experience. With respect to VQA, distortions on degraded videos will corrupt such predictability, which disturbs the serial dependence during visual perception. If we set the distortion level with some criteria, such as peak signal-to-noise ratio (PSNR) for noise, or the multi-scale structural similarity index measurement (MS-SSIM) for structure [2], we find that at the same level of distortion degradation, a stable and predictable time-varying degradation is much preferable to the HVS, and tends to be with a higher perceptual quality, while a disrupted and unpredictable one shows worse. This is because the former shows good consistency with SDP in the HVS, and the SDP lowers the perceptual feeling of the latter.

An illustration is shown in Fig. 1, where (a) and (b) are two impaired videos named as bs_1 and bs_2 respectively (from CSIQ video quality database [14]). bs_1 is degraded with compression distortion, while bs_2 is contaminated with packet loss. Though they share the same level of distortion, their perceptual quality values are different. As shown in the figure, due to the contamination of compression artifact, there exists obvious blurring and blockiness around the basketball player enlarged in Fig. 1(a). Also, due to the impact of packet loss, apparent content loss and shifting are observed in Fig. 1(b). Fig. 1(c) gives the corresponding MS-SSIM scores for the two videos: the black line is for (a) bs_1, while the red is for (b) bs_2. To better understand how the SDP affects VQA task, the green rectangular area in (c) is enlarged and shown in (d) to give a more detailed analysis. We assume that the purple dash line in Fig. 1(d) indicates the current moment, then the solid parts (on the left) of the two curves (the black and the red) represent the past, while the dashed parts (on the right) represent the future.

As depicted in Fig. 1(c), since the compression artifact in this illustration always leads to a constant degradation, bs_1 (the black curve) presents a much stable quality curve, and shows a strong serial dependence during visual perception from the perspective of SDP. According to the mechanism of SDP, the HVS can inform the perception at the present moment with the previous perception. As shown in Fig. 1(d), given the current moment (the purple dash line) and the previous (the solid black curve), since the black curve is smooth and stable, the following perceptual degradation can be well predicted with the previous experience in an easy way. In this way, the



(a) Impaired video bs_1 (b) Impaired video bs_2

(c) MS-SSIM index curve (d) The green rectangular area in (c)

Fig. 1: An illustration of how VQA task is affected by SDP. (a) and (b) are two impaired videos with similar distortion level (measured by MS-SSIM here), and (c) shows the corresponding MS-SSIM curves. The black curve (bs_1) is highly serially dependent and predictable (DMOS—from 0 to 100, and the higher the worse perceived quality: 44.898), while the red (bs_2) is interrupted and more unpredictable. The unpredictability makes bs_2 a lower quality (DMOS: 68.497) . The capability of prediction for the two curves is further illustrated in (d), which is an enlarged version of the green rectangular area shown in (c).

perceptual feeling in the HVS can be greatly contiguous and predictable with little surprise, thus causing a higher perceptual quality (DMOS—Difference Mean Opinion Scores: 44.898, the lower the better).

On the contrary, as a result of packet loss, the degradation is always random and highly unpredictable. Just as given in Fig. 1(c), bs_2 (the red curve) shows a more unstable quality variation with obviously less serial dependence. The pop-out distortion always disturbs the prediction in the HVS since the previously perceived degradation information cannot be a good indicator for the current perception any more. As also shown in Fig. 1(d), even if given the past of bs_2 (the solid red curve), the subsequent curve (the red dashed) cannot be predictable with the previous experience. This unpredictability breaks the SDP in the HVS, and even though the two videos share a similar distortion level, the perceptual quality of the video (b) is much lower (DMOS: 68.497).

Therefore, according to the analysis above, we suggest that the HVS tends to prefer those video frames with strong serial dependence (thus easy to be predicted), and resist the disrupted and unpredictable ones. Considering this strong serial dependence during visual perception, we tackle the VQA task as a serially dependent problem in this work as to be explored in the next section.

III. Proposed Method

In this section, inspired by SDP in the HVS, a novel framework is proposed for the VQA task. In this work, the

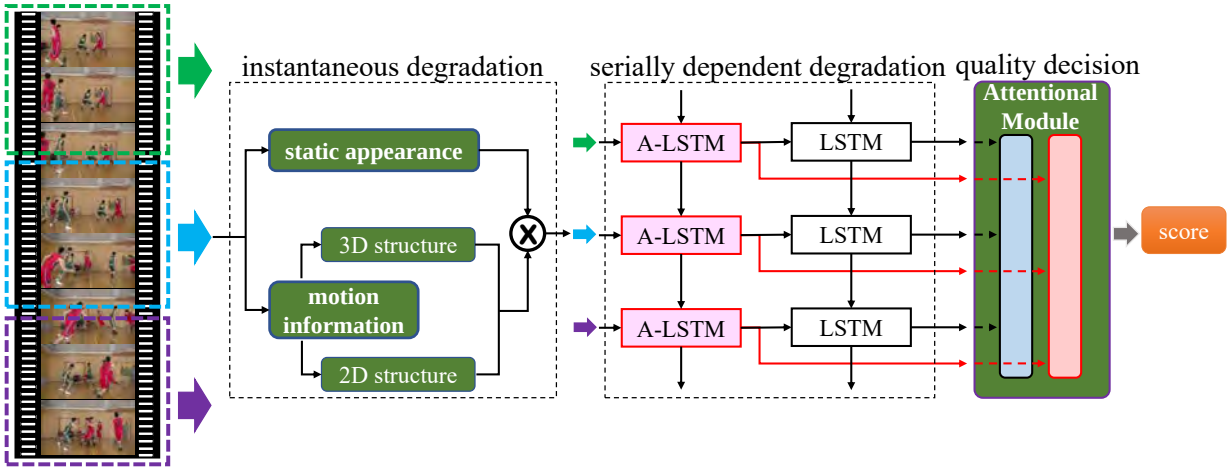


Fig. 2: Flow-chart of the proposed SDM FR-VQA. The video data are first segmented into several short snippets, base on which the instantaneous degradations are estimated. Then, the isolated instantaneous degradations among all snippets are fed into a two-layer network consisting of A-LSTM and LSTM to model the serially dependent degradations. An attention-based quality decision procedure is followed to make the final decision for the video quality.

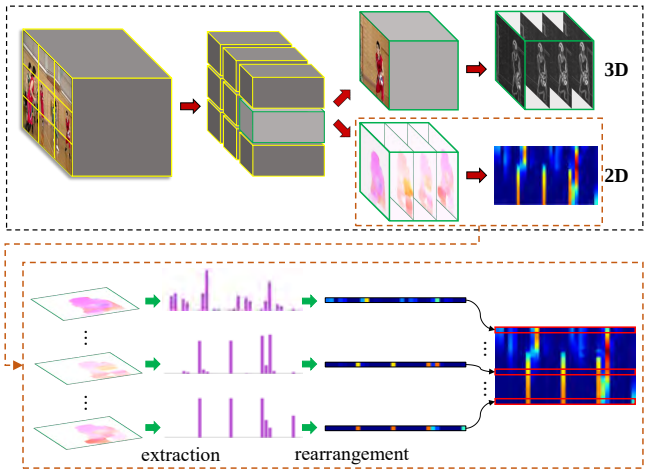


Fig. 3: Motion representation for instantaneous degradation. Video data is first partitioned into spatiotemporal tubes. The frames within each tube are stacked to construct a content-based 3D structure, and the features of corresponding optical flow maps in the tube shape a 2D structure.

proposed SDM highlights the strong dependence in serial perception, and tackles VQA as a serially dependent problem. The video quality is estimated with three stages. The instantaneous degradation is first estimated. Then, the serially dependent degradation is modeled. Finally, by mimicking the subjective rating after video-viewing, a quality decision procedure is introduced to obtain the overall video quality. The flowchart of the proposed model is shown in Fig. 2.

A. *Instantaneous degradation*

Instantaneous degradation is seen as the feeling of stimuli at first sight without any intervention of previous working memory. Based on the well-known two-pathway visual processing in the HVS, the instantaneous degradation is characterized with two aspects: the static appearance (corresponding to the ventral stream), and the motion information (corresponding to the dorsal stream). The former is always measured with a well-developed IQA method, but how to solve the latter is still an open challenge.

In many video-related research areas, motion in videos is generally extracted and represented via two ways: one is a 3D structure learned from 3D CNNs [34], and the other is a 2D structure with motion estimation [35]. Following these studies, we propose two types of structures extracted for motion representation: an explicit content-based 3D structure and an implicit feature-based 2D structure. Via stacking video frames into a 3D structure, the explicit motion content is acquired and further analyzed by 3D convolution. For a more in-depth characterization of motion, we also extract features from optical flow maps. By stacking features among multiple frames, the motion information can be represented implicitly with a 2D structure. The 3D structure shows the motion content directly, while the 2D structure implies the motion in feature-based domain. And the work in [16] also verifies the complementation of the both.

Specifically, reference and distorted videos are first divided into short snippets, and each snippet contains $\mathcal{T}$ frames. This snippet-based method is under the hypothesis that, when a video displays, the perception of human is not based on a single frame, but a short-range of frames. Typically, the frame rate of common videos is about 30, and the brain cannot react to every single frame within $1/30$ second [36, 37]. Therefore, the perceptual responding is believed to be based on a short-range snippet, and this hypothesis is reasonable. As illustrated in Fig. 3, within each snippet, a window with the size of $w \times w$ is set to segment the snippet. Consequently, a snippet with $\mathcal{T}$ frames is divided into multiple $w \times w \times \mathcal{T}$ spatiotemporal tubes. Based on these spatiotemporal tubes, both an explicit content-based 3D structure and an implicit feature-based 2D structure are extracted, and the degradation on motion information is

TABLE I: A light-weight CNN architecture for feature-based motion degradation

stage	kernel size	channel	downsample	output ($C \times W \times H$)
raw input	-	1	-	$1 \times 32 \times 18$
conv2d	3	16	False	$16 \times 32 \times 16$
conv2d	3	16	True	$16 \times 16 \times 8$
conv2d	1	32	False	$32 \times 16 \times 8$
conv2d	3	32	True	$32 \times 8 \times 4$
conv2d	3	64	True	$64 \times 4 \times 2$
conv2d	1	16	False	$16 \times 4 \times 2$
mean	-	-	-	$16 \times 1 \times 1$
fc	-	1	-	1

characterized.

Towards the 3D structure, due to the limited data size in the VQA related databases, it is hard to deploy a 3D CNN to analyze the degradation directly. Since the HVS is highly sensitive to content change, the previous work in [16] employed the 3D Prewitt operators to convolve the stacked video frames and achieved promising performance. Thus, in this work, instead of training an overloaded 3D CNN model, the 3D Prewitt operators are directly employed to convolve the data. Hence, the resulting magnitude of convolution for each tube can be obtained as e. Then, the similarity of the magnitude is computed as

$$\mathcal{S} = \frac{2 \cdot e_r \cdot e_d}{e_r^2 + d_d^2}, \tag{1}$$

where e_r (e_d) denotes the magnitude of the reference (distorted) snippet after convolution. Following the suggestion in [38], the degradation on the content-based motion information is obtained with a standard deviation pooling strategy $\mathcal{D}_c = \mathrm{STD}(\mathcal{S})$.

Apart from the content-based 3D structure, motion is also represented by a feature-based 2D structure with motion estimation. Since the HVS presents strong selectivity for motion velocity (both speed and direction) [39], we present a feature-based motion representation to characterize the velocity with a 2D space-time view. Firstly, within each spatiotemporal tube, the motion displacement is obtained with a dense optical flow computation [40]. Then, the histogram of optical flow (HoF) for every two adjacent frames is calculated as motion features due to its good ability on the characterization of motion speed and direction. As the HoF is a natural 1D signal, the motion within a spatiotemporal tube can be shown in a 2D view by vertically stacking the histograms along the time axis. As is shown in Fig. 3, within the 2D space-time structure, each row corresponds to the HoF features of a single frame in the spatiotemporal tube, and each column corresponds to the same histogram pattern among multiple frames. We calculate the similarity of the extracted feature-based motion representation for both the reference and the distorted video with Eq. (1). Benefiting from the 2D space-time arrangement, the feature-based motion degradation can be further explored with a light-weight but effective 2D CNN. The CNN architecture is specified in Tab. I, and some necessary paddings are added to fully convolve the information. After six convolutional layers and a fully-connected layer, the degradation on the feature-based motion information is acquired as $\mathcal{D}_f$.

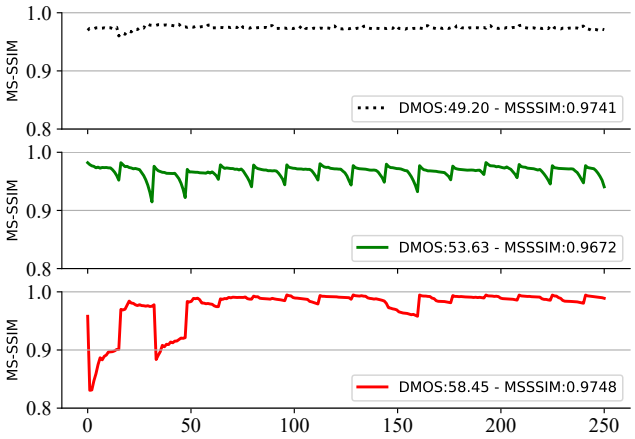


Fig. 4: An illustration for the effect of the perceptually worst moment on perceptual quality. The three curves presents MS-SSIM values for three degraded videos from LIVE database [41], and their corresponding DMOS and MS-SSIM values are given in the bottom right corner.

We aggregate both the content-based and feature-based motion degradations in each snippet as

$$\mathcal{D}_{motion} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{D}_{c,i} \cdot \mathcal{D}_{f,i}, \tag{2}$$

where M is the number of tubes within the snippet. Similar to the existing other work, the degradation on the static appearance is estimated with the mean value of a frame-by-frame IQA algorithm [38], denoted as $\mathcal{D}_{appearance}$. Then, the instantaneous degradation is given as a simple combination:

$$\mathcal{D}_{ins} = \mathcal{D}_{appearance} \cdot \mathcal{D}_{motion}. \tag{3}$$

All the related parameter settings follow from the work in [16]: $\mathcal{T}$ is set to 18 (corresponding to about 0.6 second for a 30 fps video) and w is 48. The number of bins in HoF is set as 8, and that of cells is 4. Hence, the space-time feature-based motion representation is with a size of 18×32, as the input size given in the Tab. I. And the detailed discussion of these parameters are given in the previous work.

B. Serially dependent degradation

As introduced in Sec. II, the HVS presents a strong serial dependence during visual perception, using the previous information to guide the prediction at the present. And disturbance of the serial dependence can cause a negative effect on the perceptual quality. As there's a natural dependency in video sequences, we suggest tackling the VQA problem with a serially dependent method.

As SDP shows that previous perception impacts the current perception, at each moment, we propose to take both the perception at the last moment in the previous and the current input into account. Besides, we empirically demonstrate that the perceptually worst moment in the previous experience greatly affects the perceptual quality. As is shown in Fig. 4, the three

curves correspond to the frame-by-frame MS-SSIM three degraded videos from LIVE video quality dat As mentioned in Sec. II, the one shown on th black dash line) represents a video with stable d and thus obtains a better perceptual quality (DM due to the strong serial dependence in serial degrac middle one (the green line, DMOS: 53.63) repr video with disrupted degradation, is perceptually the top because of the resistance of SDP. The thir red line in the bottom, DMOS: 58.45) represents which the perceptual quality is first annoying but in the subsequent frames. Although the average value of the third video (MS-SSIM: 0.9748) is h that of the second video (0.9672), the perceptual the third one is much worse. The larger DMOS of the third video indicates that the perceptually worst moment in the previous can be recalled in the subsequent and greatly affect the video quality. Therefore, in our practical modeling, we take the perceptual effect from the previous degradations into account explicitly (both the last moment and the perceptually worst moment in the previous), and propose an assessment-directed LSTM (A-LSTM) to cater to the VQA problem.

LSTM [36] is extensively used in many tasks to capture the sequential characteristics. The classical computation of a single LSTM cell is given as

$$\begin{aligned} i_t &= \sigma(W_{ii}x_t + W_{hi}h_{t-1}) \\ f_t &= \sigma(W_{if}x_t + W_{hf}h_{t-1}) \\ g_t &= \tanh(W_{ig}x_t + W_{hg}h_{t-1}) \\ o_t &= \sigma(W_{io}x_t + W_{ho}h_{t-1}) \\ c_t &= (f_t \cdot c_{t-1} + i_t \cdot g_t) \\ h_t &= o_t \cdot \tanh(c_t) \end{aligned} \tag{4}$$

where x_t is the input at current moment, h_{t-1} is the last output (also known as the hidden representation) of the cell, and c_{t-1} is the last memory state in the cell that updates itself iteratively [36]. Given the current input x_t and the last output h_{t-1}, i_t, f_t and o_t are classical gates to control the input, memory and output separately. g_t shows the current state, and c_t updates its state with the current state g_t and the previous state c_{t-1}. And finally the cell outputs the current representation h_t. All the W with subscripts in the equation are weights to be decided. LSTM receives the output from the previous moment and updates the cell state itself at every moment, thus forming the characteristic of serial dependence naturally.

Towards the VQA task, as illustrated in Fig. 4, the perceptually worst moment in the previous affects the perceptual quality greatly. Therefore, beyond the classical computation in the LSTM cell, we propose an A-LSTM cell where a recall procedure is injected into the cell to explicitly highlight the importance of the perceptually worst moment. At each moment, the hidden representation of the perceptually worst moment in the previous τ duration will be recalled, and affect the current output. Hence, regarding to Eq. (4), we introduce a new variable r_t in the cell, which is computed as

$$r_t = \text{RECALL}(h_{t-\tau}, h_{t-\tau+1}, ..., h_{t-1}). \tag{5}$$

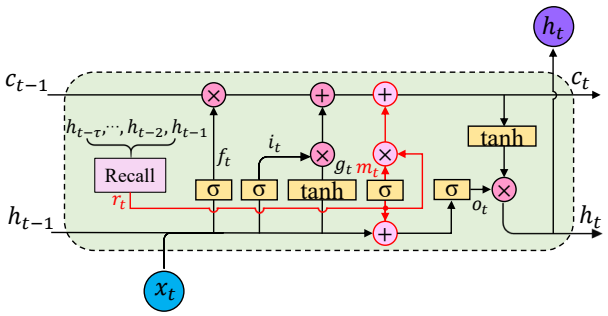


Fig. 5: The architecture of A-LSTM cell. Different from the classical LSTM cell, given the last output h_{t-1}, the last state c_{t-1}, and the current input x_t, the cell recalls the hidden representation of the perceptually worst moment in the previous τ duration. With a degradation gate m_t to control the importance of this recall, r_t would affect the current state c_t and output h_t. The procedure is highlighted in red.

Here, r_t indicates the hidden representation corresponding to the perceptually worst moment in the previous duration. This recall is a natural perception processing in the assessment task for videos. During the assessment task, the perceptual feeling is related not only to the last moment, but also the perceptually worst moment in the previous duration. Hence we suggest that this worst perception would generally affect the current cell state as well as the output at present. As a result, the output gate o_t in Eq. (4) is proposed to be computed additionally with the involvement of r_t. Thus, the output gate is reconsidered as

$$\hat{o}_t = \sigma(W_{io}x_t + W_{ho}h_{t-1} + W_{ro}r_t). \tag{6}$$

Here, $\hat{o}_t$ represents the modified output gate to distinguish that in Eq. (4). By introducing a degradation gate m_t to control the importance of this recall adaptively, the cell state $\hat{c}_t$ and the output $\hat{h}_t$ are updated as

$$\begin{aligned} m_t &= \sigma(W_{rm}r_t) \\ \hat{c}_t &= (f_t \cdot \hat{c}_{t-1} + i_t \cdot g_t + r_t \cdot m_t) \\ \hat{h}_t &= \hat{o}_t \cdot \tanh(\hat{c}_t). \end{aligned} \tag{7}$$

The overall architecture of the A-LSTM cell is shown in Fig. 5, where the introduced recall procedure is highlighted in red. For a general purpose, we directly use the symbols o_t, c_t, and h_t in the figure, rather than those in Eq. (6) and Eq. (7).

As shown in Fig. 2, the instantaneous degradation $\mathcal{D}_{ins}$ in each video snippet is characterized sequentially, and fed into such an A-LSTM layer serially. The hidden size of A-LSTM is set to 4. Hence, after the A-LSTM layer, a sequentially dependent hidden representation $\mathcal{H} = \{h_1, h_2, ..., h_t\}$ is generated, each with a dimension of 4. Then, an additional standard LSTM layer is stacked to reduce the dimension (from 4 to 1). Through exploring the serial dependence during visual perception thoroughly, the serially dependent degradations (i.e., the continuous perceptual quality values) are acquired as $\mathcal{Q} = \{q_1, q_2, ..., q_t\}$ over the video.

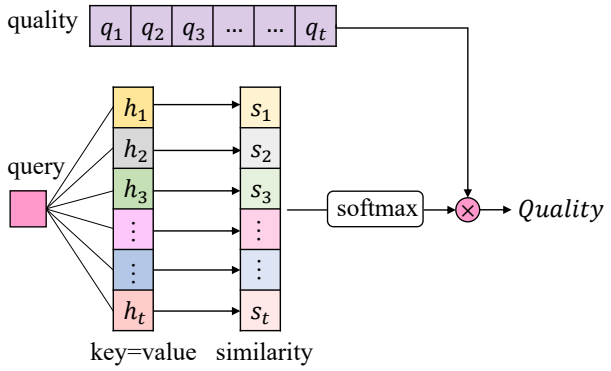


Fig. 6: The architecture of attention-based quality decision

C. Quality decision

In a typical subjective VQA process, viewers are asked to give their opinions for the test video, which calls for a quality decision procedure. And during the quality decision, viewers recall the sequential degradations and tend to give more attention to the perceptually worst moment. This phenomenon influences many pooling methods, such as the percentile pooling and classified-based weighting. The percentile pooling sorts the quality values, picks up the mean value of the worst part and discards the other quality values. And classified-based weighting is to classify the quality values and assigns different weighting factors.

Instead of discarding part of the quality values or assigning hard weighting factors, in this paper, we propose an attention-based quality decision which takes all the values into consideration and adaptively tunes the importance of quality values. Since video viewers tend to focus more on the perceptually worst moment when they consider and decide the video quality, the proposed quality decision procedure traverses the recollection, highlights the worst perceptual quality, and gives the quality score of the video. A common idea is shared between this decision procedure and the aforementioned A-LSTM that the perceptually worst moment plays an important role in VQA, but occurs in different phases. A-LSTM is for the modeling of SDP when a video displays, while this attention-based procedure is for the overall quality decision after the video-viewing.

Fig. 6 gives the general architecture of the procedure. As the worst moment captures more attention, in this work, the hidden representation of the worst moment in A-LSTM layer is set as a query h_Q. Then, the query goes through the keys consisting of the whole representations $\mathcal{H} = \{h_1, h_2, ..., h_t\}$ over video. The similarity between the keys and the query is calculated via an useful scaled dot-product method [42] as

$$S = \frac{h_Q \mathcal{H}^T}{\sqrt{d_h}}, \tag{8}$$

where d_h is the dimension of hidden representation (as mentioned in III-B, $d_h = 4$ in this paper). Eq. (8) actually computes the similarity of hidden representations as their inner product but with a scale factor. This computation enables a adaptive and soft weighting for the serially dependent quality, which is different from the hard weighting assignment and tends to be more flexible. Afterwards, a softmax processing is followed to normalize the similarity and the final video quality is obtained as

$$Quality = \mathcal{Q}^T S = \sum_i s_i q_i, \tag{9}$$

where $\mathcal{Q}$ is the estimated serially dependent quality obtained in Sec. III-B, and q_i is the quality in the i-th video snippet.

IV. Experimental Result Analysis

In this section, we first give a brief description of databases and protocols for performance comparison. Then, a comprehensive comparison between the proposed SDM and the existing FR-VQA methods is conducted. A further ablation experiment is followed to show the significance of each component of the SDM. Afterwards, a cross dataset test is followed to present the capability of generalization. Finally, the SDM is transplanted to other FR-VQA/IQA methods and the experimental result shows its effectiveness.

A. Databases and protocols

To conduct a comprehensive performance comparison and demonstration, four publicly available VQA databases are used in this paper, i.e., the LIVE video database [41], the CSIQ video database [14], the IVPL video database [43], and the IVC-IC database [44]. A detailed description about the databases can be found in [16].

Three widely-used metrics are adopted for performance comparison, which are Spearman rank order correlation coefficient (SROCC), Kendall's rank order correlation coefficient (KROCC), and Pearson linear correlation coefficient (PLCC). SROCC and KROCC evaluate the monotonicity, while PLCC indicates the accuracy of prediction. All these three metrics range from 0 to 1, and the higher value means the better performance. Considering the limited space in the table, the root mean square error (RMSE) is not included here.

Five-fold cross validation experiment is applied in this paper. For each database, the video data is split into five parts considering the reference videos. Four in five of the videos are used to train the model, and the rest data validate the performance. Due to the limited quantity of the reference videos in LIVE, CSIQ and IVPL, in these three databases, we traverse every possible training-testing combination, and exhaustively repeat the procedure to get more robust performance. That is, as the number of reference videos in LIVE and IVPL is 10, we repeat the training-testing procedure for 45 rounds considering every possible combination. And for CSIQ, 66 rounds are processed since the number of reference videos is 12. As for IVC-IC, the number of possible combinations is large enough, so we take 20 rounds randomly instead. The median value among all the rounds is adopted as the final performance.

The proposed SDM is implemented with PyTorch 1.1.0 in an environment of Ubuntu 2016 and Python 3.6.5. The learning rate is set to 0.0008, and Adam optimizer is employed. L_1 loss is adopted for a supervised learning, and we also fix the random seed of SDM to keep a stable reproducibility.

TABLE II: Performance comparison with FR-VQA methods on databases

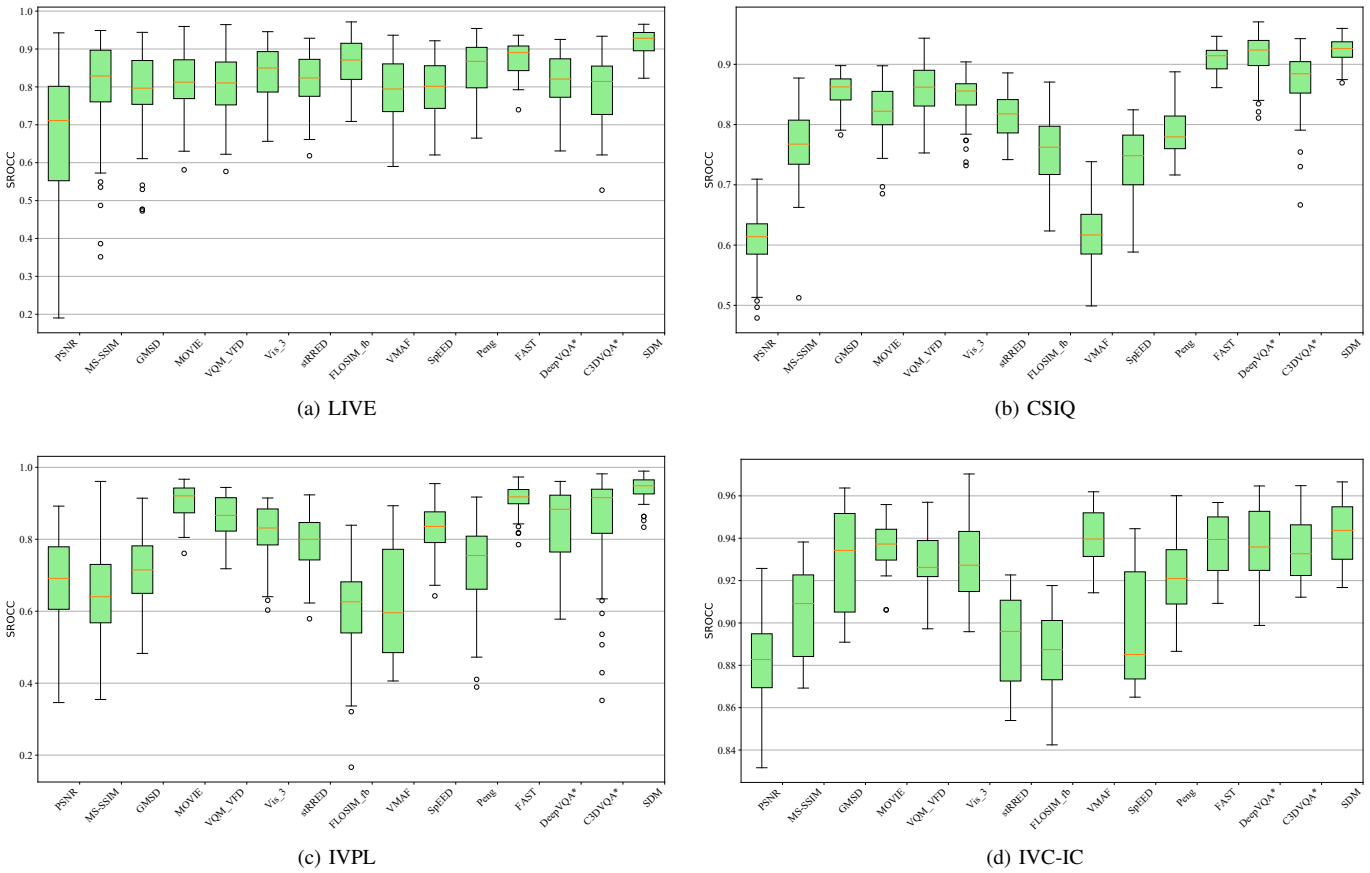
	LIVE (150)			CSIQ (216)			IVPL (128)			IVC-IC (240)			Weighted Average		
	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC
PSNR	0.7112	0.5402	0.7490	0.6139	0.4317	0.6235	0.6913	0.5362	0.7099	0.8827	0.7023	0.8920	0.7352	0.5606	0.7520
MS-SSIM	0.8287	0.6552	0.8608	0.7676	0.5762	0.7771	0.6404	0.5026	0.6584	0.9091	0.7414	0.9200	0.8042	0.6335	0.8202
GMSD	0.7967	0.6184	0.8305	0.8625	0.6762	0.8468	0.7148	0.5435	0.7333	0.9342	0.7908	0.9487	0.8467	0.6787	0.8570
MOVIE	0.8125	0.6368	0.8423	0.8221	0.6365	0.8117	0.9206	0.7681	0.9370	0.9373	0.7916	0.9502	0.8750	0.7102	0.8851
VQM_VFD	0.8105	0.6460	0.8369	0.8620	0.6683	0.8650	0.8664	0.6931	0.9007	0.9262	0.7700	0.9439	0.8732	0.7013	0.8913
Vis3	0.8499	0.6828	0.8701	0.8557	0.6698	0.8498	0.8314	0.6508	0.8751	0.9273	0.7709	0.9410	0.8737	0.7022	0.8882
stRRED	0.8237	0.6460	0.8418	0.8178	0.6254	0.8091	0.8000	0.6190	0.8091	0.8960	0.7219	0.8909	0.8415	0.6600	0.8425
FLOSIM_fb	0.8710	0.6904	0.8905	0.7624	0.5619	0.7737	0.6262	0.4762	0.6660	0.8874	0.7059	0.9005	0.8017	0.6203	0.8202
VMAF	0.7944	0.6092	0.8221	0.6169	0.4402	0.6323	0.5957	0.4783	0.5913	0.9396	**0.8068**	**0.9538**	0.7550	0.6013	0.7691
SpEED	0.8016	0.6368	0.8174	0.7484	0.5476	0.7507	0.8357	0.6561	0.8465	0.8851	0.7112	0.8408	0.8192	0.6382	0.8105
Peng	0.8674	0.7011	0.8773	0.7795	0.5841	0.7781	0.7548	0.5870	0.7767	0.9210	0.7660	0.9373	0.8394	0.6680	0.8502
FAST	0.8904	0.7273	0.9033	0.9143	0.7476	0.8930	0.9179	0.7672	0.9240	0.9394	0.8021	0.9441	0.9183	0.7647	0.9172
DeepVQA	0.9152	-	0.8952	0.9123	-	0.9135	-	-	-	-	-	-	-	-	-
C3DVQA	0.9261	-	0.9122	0.9152	-	0.9043	-	-	-	-	-	-	-	-	-
DeepVQA*	0.8207	0.6368	0.8355	0.9237	0.7603	**0.9238**	0.8835	0.7143	0.8980	0.9359	0.7881	0.9437	0.8996	0.7361	0.9078
C3DVQA*	0.8143	0.6230	0.8224	0.8843	0.7000	0.8751	0.9157	0.7460	0.9136	0.9327	0.7899	0.9404	0.8913	0.7217	0.8924
SDM	**0.9284**	**0.7779**	**0.9184**	**0.9260**	**0.7683**	0.9002	**0.9487**	**0.8188**	**0.9458**	**0.9437**	0.8048	0.9472	**0.9362**	**0.7910**	**0.9272**

B. Performance comparison

With the above-mentioned VQA databases, the performance of the proposed SDM is compared with 12 classical models (i.e., PSNR, MS-SSIM [2], GMSD [38], MOVIE [10], VQM_VFD [45], Vis3 [14], stRRED [4], FLOSIM_fb [15], VMAF [46], SpEED [9], Peng [11] and FAST [16]), as well as two deep learning-based FR-VQA methods (i.e., DeepVQA [19] and C3DVQA [20]). The performance of classical methods is also obtained via the same testing rounds expounded above for a fair comparison. As for the two deep learning-based methods (i.e., DeepVQA, and C3DVQA), the results are first adopted directly from the corresponding literatures. On the other hand, since the source code of these two methods can be totally accessible, the methods are also verified in the same training-testing environment with the default settings, which is denoted with * (i.e., DeepVQA*, and C3DVQA*). It is worth noting that, due to the huge consumption of computation and time in the training phase of the two methods, especially in our setting that exhausts all the training-testing possibilities, the hyperparameters are directly used as it is in the code and not tuned further.

The performance comparison among the four databases is shown in Tab. II. The best performance is highlighted in bold and the second best is with underline. By taking SDP during visual processing procedure into consideration, the proposed SDM can achieve higher performance over existing methods, and make a further improvement on the consistency of quality prediction. Concerning the number of parameters, SDM (about 32K) is 4x lighter than DeepVQA (about 141K), and about 7x lighter than C3DVQA (about 227K). Even with much less parameters, SDM still outperforms the two newly proposed deep learning-based methods on LIVE and CSIQ databases, and shows obvious superiority to the classical methods on the first three video quality databases. As for IVC-IC, SDM is closer to VMAF, since the performance is slightly higher than that of VMAF on SROCC, but faintly lower on KROCC and PLCC. The possible reasons are two-fold: 1) IVC-IC only contains compression artifacts, and VMAF is also trained on large-scale compressed videos, which makes VMAF more suitable than other VQA methods; 2) As we suggest that the main contribution of our work is the consideration of the temporal influence in quality evaluation, the compression artifacts generally result in stable degradation along time axis, thus limiting the performance of SDM (we will examine this later). Even though VMAF shows good performance on IVC-IC, the performance on LIVE, CSIQ and IVPL elucidates its shortage. However, the proposed SDM can always keep good consistency with human perception (over 0.92 on SROCC) throughout the databases, and this shows a better stability and stronger generalization capability. Note that the performances of DeepVQA* and C3DVQA* on LIVE are not as impressive as they present in the literatures, and the possible reasons are four folds. 1) First, both of the two methods are experimented with limited training-testing rounds randomly in the literatures, and the randomness would cause the fluctuation of performance. 2) Second, our setting is to exhaust all possible training-testing combinations (except for IVC-IC) for a stable comparison, thus the different test conditions would cause the variation of performance. 3) Third, both of the two methods have not fixed the random seed in their codes, thus the performance would not be fixed since the weights are initialized randomly every time, which contributes to the fluctuation. 4) Finally, the hyperparameters might be further tuned among different databases. However, this is not practical in our actual experiments. The training phase of both DeepVQA and C3DVQA is very time-consuming. Specifically, training C3DVQA on CSIQ would cost more than 11 hours each training-testing round with 2 GPUs, and training DeepVQA on CSIQ would cost about 12 hours each training-testing round. In our setting, the training-testing procedure would iterates for all the combinations (except for IVC-IC) and all the databases. In this condition, we can hardly tune the hyperparameters of the two methods.

We further report a more comprehensive and holistic comparison with the average performance, which is weighted by the number of samples in each database. Through the weighted average performance in the right of Tab. II, some



(a) LIVE

(b) CSIQ

(c) IVPL

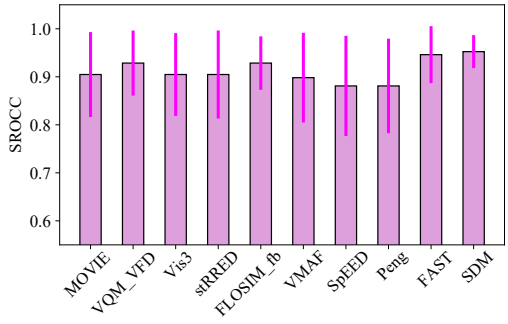
(d) IVC-IC

Fig. 7: Box plot for performance (SROCC) comparison on databases

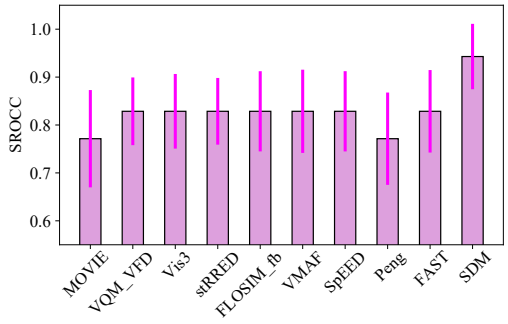
traditional VQA methods (i.e., MOVIE, VQM_VFD, Vis3, and FAST) show obviously stable performance among databases. And due to the imbalanced behaviors among databases, other algorithms (such as VMAF and SpEED) present weaker performance. The proposed SDM notably outperforms existing methods (including the classical and deep learning-based ones) and shows reliable consistency with the human perception, which verifies the robustness and stability.

As mentioned in Sec. IV-A, for a fair comparison, the validation set is traversed throughout the database and algorithms are tested for all the rounds. In order to get an intensive comparison, Fig. 7 gives the box plot of performance (SROCC) among all the rounds on the four video databases. From Tab. II, MS-SSIM presents a surprising performance on LIVE (SROCC is 0.8287), which is even better than MOVIE and VQM_VFD (the classical algorithms specified to the VQA problem). However, as illustrated in Fig. 7(a), although MS-SSIM presents a higher median SROCC, the performance among the 45 rounds are decentralized and show more outliers. The worst SROCC of MS-SSIM is 0.3515, while that of MOVIE is 0.5812. This suggests that the median performance cannot be the exclusive criterion for comparison. From this point of view, it is found that among the traditional VQA algorithms, FAST and FLOSIM present better performance on LIVE video database (with the higher values of both the median performance and the worst performance). As shown in Fig. 7(b), comparing to our proposed method, although DeepVQA* achieves a competitive performance on CSIQ, there are more outliers in DeepVQA*, which indicates that the proposed method is more robust and stable than DeepVQA*. The similar phenomena can be also found in the other databases. Among the four databases, the proposed SDM generally shows a better median performance and a higher worst performance. The advantageous performance and the compact distribution among the databases demonstrate the effectiveness and robust of the proposed method.

As the aforementioned in Sec. II, artifacts with predictable degradations (some compression artifacts) are more preferable to the HVS, while ones with disrupted and disordered disturbances (usually packet loss distortions) show more reluctance. To show a deeper understanding of how serial dependence affects the assessment of video quality, a further performance (SROCC) comparison is shown in Fig. 8. In the comparison, the performance of two distortions in LIVE database is listed, which are MPEG compression (more predictable) and IP-based packet loss (more unpredictable). As the degradation resulting from MPEG compression distortion is more stable and predictable (see Fig. 1 for illustration), SDM shows a limited advantage over the other VQA methods. As illustrated in Fig. 8(a), SDM obtains a narrower variation, but a very close median SROCC (0.9524) to those of other VQA methods (0.9461 is the best). However, when it comes

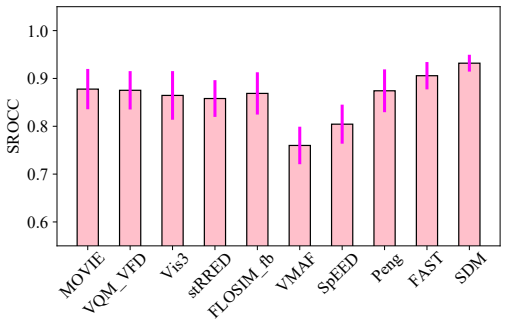


(a) MPEG compression

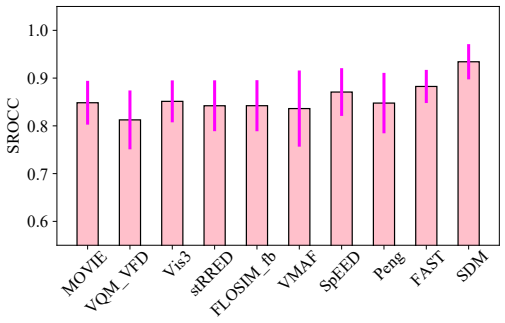


(b) IP-based packet loss

Fig. 8: Performance (SROCC) comparison on distortions in LIVE database



(a) Compression distortion



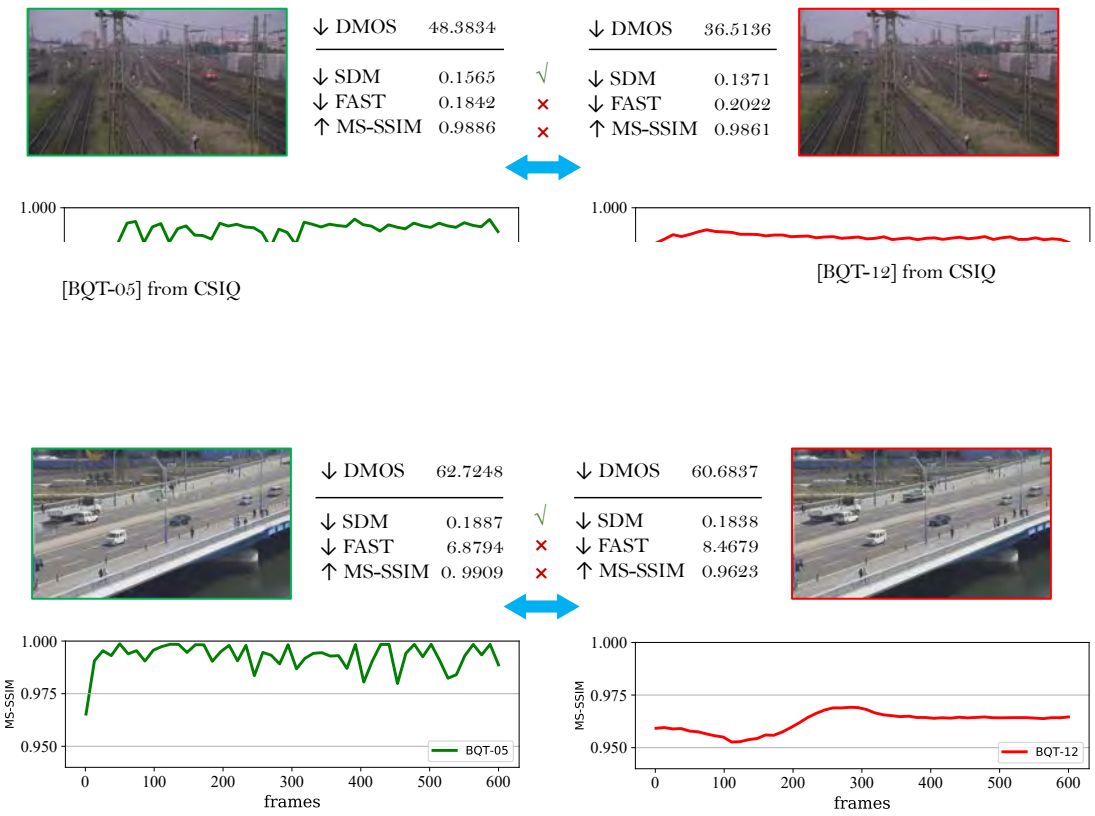
(b) Packet loss distortion

Fig. 9: Weighted average of performance (SROCC) on distortion types among databases

to IP-based packet loss, this advantage is more evident. VQA methods, that estimate quality in isolation and break up the relation from moment to moment, such as stRRED and FLOSIM, can achieve only limited consistency with the human perception. As shown in Fig. 8 (b), with the consideration of serial dependence, SDM shows an obvious improvement. The performance (SROCC) of SDM is 0.9429, which is much better than the best of the others (0.8286). And the consistency between SDM and the human perception is promoted by 13.8%. Besides, the lower variation also shows the superiority of SDM.

A more general experimental result is shown in Fig. 9. In this experiment, the performance (SROCC) of compression artifacts and packet loss distortions included in three databases (i.e., LIVE, CSIQ and IVPL) is first evaluated, and then weighted according to the number of the samples in each database. That is, Fig. 9 reports the weighted average performance among the three databases, and shows a more general performance comparison. We ignore IVC-IC database since there's no packet loss distortion in this database. From the figure, a similar trend can be observed that the performance of SDM on both compression and packet loss is better comparing to the other VQA methods, and the performance improvement on packet loss is more evident than that on compression distortion. Compared with the second best method, SDM improves the performance by only 2.9% regrading to compression distortion, but 5.9% for packet loss distortion. The compression generally contributes to a stable degradation for videos, while packet loss usually results in a disrupted and unpredictable degradation. Therefore, from this experimental comparison, it is shown that SDM presents limited strength in distortion types with stable degradations, but a more remarkable improvement for those with disrupted and unpredictable degradations. This experimental result conforms to the illustration in Sec. II, and verifies the proposed method.

Besides the quantitative comparison, a qualitative comparison is given in Fig. 10 to better support the proposed method. In Fig. 10, examples are selected from both LIVE database and CSIQ database to verify the generalization. In each example, the video in the left with green is suffered from more interrupted and unpredictable degradation, comparing to the one in the right with red. Here, similar to Fig. 1, we still adopt frame-wise MS-SSIM scores to reflect the fluctuation of degradation, which is shown in the bottom of each example. In this comparison, SDM is compared with a robust method (FAST) and the classical MS-SSIM. A better video would result in lower DMOS, lower SDM score, lower FAST score and higher MS-SSIM score, which is represented by the arrow in the figure. As shown in the figure, the left video with more disrupted degradation obtains a worse perceptual quality (higher DMOS) comparing to the right one. And due to the consideration of the temporal influence, the proposed SDM performs consistently with DMOS. However, since the temporal fluctuation is not included by the other two methods (i.e., FAST and MS-SSIM), the simple average pooling which neglects the characteristic of serial dependence results in inconsistency.



(b) *BQT-05*/*BQT-12* from CSIQ database

Fig. 10: Qualitative comparison by examples from (a) LIVE database and (b) CSIQ database. For each example, the video in green is suffered from more interrupted and unpredictable degradation comparing to the one in red, which is shown with the frame-wise MS-SSIM scores in the bottom. Our proposed method (SDM) can effectively capture the temporal influence, thus perform consistently with human perception (DMOS). Note that a better video would result in a lower DMOS, lower SDM score, lower FAST score, and higher MS-SSIM score, which is depicted with the arrow.

C. Ablation experiments

To better cater to the VQA problem, we introduce a degradation recall into the LSTM, and propose the A-LSTM. Meanwhile, we also propose an attention-based quality decision procedure to simulate the subjective rating when assessing video quality. In this subsection, The effectiveness of the proposed A-LSTM and attention-based quality decision procedure is verified with ablation experiments. The ablation experiment is conducted on three databases (i.e., LIVE, CSIQ, and IVPL databases). IVC-IC is excluded since there are only compression artifacts.

In the first phase, we evaluate different temporal pooling methods over the instantaneous degradations, including the direct average (Direct Mean), Minkowski sum (Minkowski), percentile pooling (Percentile), VQPooling [28], and hysteresis pooling (Hysteresis) [29]. As for hysteresis pooling, we adopt the PyTorch implementation in [30], and A further summarization of these temporal pooling methods can be found in [32]. In the second phase, we construct a model consisting of a two-layer LSTM following the instantaneous degradations, and obtain the video quality by direct average of the outputs (LSTM+Mean), which serializes the instantaneous degradations but doesn't consider the effect of the perceptually worst moment and attention. Further, we explore the model with A-LSTM layer but without attention-based quality decision (A-LSTM+Mean). And the last model is the proposed SDM with both the A-LSTM and an attention-based quality decision (A-LSTM+Attention). These models are trained and evaluated through the same procedure as depicted above, and the experimental result is given in Tab. III.

TABLE III: Ablation experiment of the proposed SDM

DB	LIVE	CSIQ	IVPL
Direct Mean	0.9091	0.9024	0.9409
Minkowski	0.8914	0.9066	0..9475
Percentile	0.8941	0.9036	0.9462
VQPooling	0.8576	0.8996	0.9357
Hysteresis	0.8891	0.9124	0.9338
LSTM+Mean	0.9070	0.9212	0.9487
A-LSTM+Mean	0.9110	0.9221	0.9484
A-LSTM+Attention	0.9284	0.9260	0.9487

Due to the adequate consideration of motion information, the direct mean of the instantaneous degradations (Direct Mean) can result in a quite good performance even without any modeling of the SDP. The other temporal pooling methods in the first phase may allow for some part of the properties during

TABLE IV: Performance of cross dataset experiment on CSIQ video database (trained on LIVE)

		SDM	DeepVQA	VMAF	stRRED	Vis3	MOVIE
H.264	SROCC	0.9156	0.8793	0.9284	0.9768	0.9194	0.8960
	PLCC	0.9394	0.8985	0.9443	0.9770	0.9146	0.9086
PL	SROCC	0.8054	0.7279	0.7701	0.8546	0.8533	0.8677
	PLCC	0.7995	0.7631	0.7900	0.8701	0.8542	0.8674
MJPEG	SROCC	0.8535	0.6435	0.8871	0.7601	0.7349	0.8855
	PLCC	0.8534	0.6048	0.8125	0.7290	0.7566	0.8937
Wavelet	SROCC	0.9024	0.8252	0.8965	0.9459	0.8999	0.9012
	PLCC	0.9301	0.8580	0.9026	0.9560	0.9286	0.9134
WN	SROCC	0.8340	0.9236	0.8831	0.9305	0.9179	0.8245
	PLCC	0.9108	0.9656	0.9165	0.9520	0.9562	0.8752
HEVC	SROCC	0.9158	0.7686	0.9287	0.9099	0.9174	0.9349
	PLCC	0.9689	0.7765	0.9391	0.9486	0.9326	0.9453
ALL	SROCC	0.8690	0.7374	0.6151	0.8129	0.8326	0.8083
	PLCC	0.8401	0.7186	0.6243	0.7933	0.8223	0.7924

TABLE V: Performance of cross dataset performance on IVPL video database (trained on LIVE)

		SDM	DeepVQA	VMAF	stRRED	Vis3	MOVIE
Dirac	SROCC	0.8914	0.8759	0.9017	0.8527	0.9155	0.8879
	PLCC	0.9101	0.9096	0.9160	0.8676	0.9136	0.8850
H.264	SROCC	0.8846	0.8323	0.8653	0.8655	0.8426	0.8482
	PLCC	0.9033	0.8955	0.9114	0.8784	0.8904	0.8617
MPEG	SROCC	0.8358	0.6605	0.7922	0.6912	0.7940	0.8162
	PLCC	0.8951	0.8464	0.8751	0.7440	0.8698	0.8362
IP	SROCC	0.8396	0.6968	0.3103	0.6672	0.7438	0.8582
	PLCC	0.8650	0.6905	0.3448	0.6235	0.7118	0.8417
ALL	SROCC	0.8850	0.7748	0.5799	0.7374	0.7948	0.8844
	PLCC	0.8888	0.7841	0.5919	0.7287	0.7959	0.8816

perception, but presents less tolerance on various validation datasets comparing to the direct average. It suggests that average pooling is simple enough but robust in most cases. By taking both the serial dependence and perceptual recall into consideration, the proposed A-LSTM is generally superior to the direct average. In CSIQ database, the SROCC of the third model (introduced by A-LSTM) elevates more than 2%, which demonstrates the effectiveness of A-LSTM and verifies that the serial dependence is important during visual perception. And comparing the performance between LSTM and A-LSTM, it shows that the performance is generally improved with the consideration of perceptual recall. The performance of A-LSTM is better than that of LSTM on LIVE and CSIQ. Even for IVPL, the performance of A-LSTM is almost the same as that of LSTM. Further experimental result shows that the attention-based quality decision procedure can make an improvement with the consideration of the attention mechanism. Through the ablation experiment, it is confirmed that the serial dependence should be well-considered in VQA modeling. And it is also verified that the perceptually worst moment can affect not only the serially dependent quality, but also the final quality decision procedure. Allowing for both the SDP during visual perception and the attention mechanism during subjective rating, the proposed SDM is well perception-inspired and accordant with subjective opinions.

D. Cross database evaluations

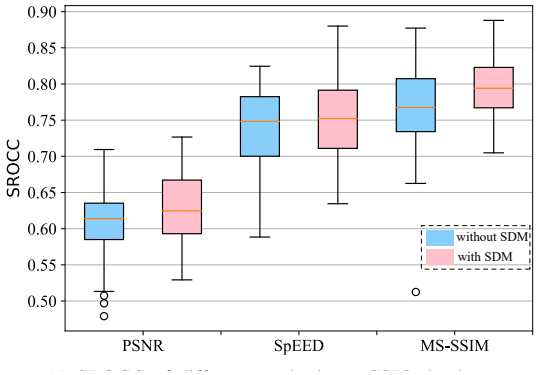
As learning-based methods are much dependent on the training data, in this subsection, cross dataset validation experiments are conducted to examine the generalization capability of SDM. For a more convincing comparison, another learning-based method (i.e., DeepVQA) and four classical VQA methods (i.e., VMAF, stRRED, Vis3 and MOVIE) are also introduced. In the first experiment, both SDM and DeepVQA are trained on the whole videos in LIVE video database and validated on CSIQ for all distortion types. Since all these methods are evaluated directly on the integral validation database, thus the performance is a bit different from those in Tab. II. Compared to LIVE video database, CSIQ contains more distortion types and video contents, thus the validation on CSIQ is much challenging for learning-based methods. As shown in Tab. IV, the performance of DeepVQA trained on LIVE is very limited (0.7374 on SROCC). Due to a good capability of motion representation and the consideration of serial dependence, SDM makes a considerable improvement, and shows good consistency with the human perception. As given in the table, SDM achieves 0.8690 on SROCC, elevating 17% for DeepVQA and Vis3, and 7% for MOVIE. Specified to distortion types, SDM can still obtain great advantages over DeepVQA for H.264, PL, MJPEG, Wavelet and HEVC. Especially for MJPEG, SDM gains more than 30% over DeepVQA on both criteria. The only less satisfying distortion type is WN, but the performance of SDM on this distortion type is still better than that of MOVIE.

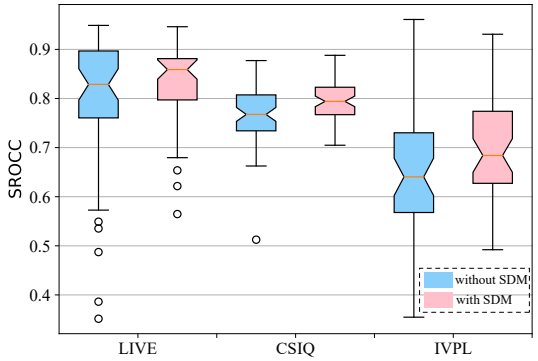
A similar cross dataset experiment is performed on IVPL. Same as the previous experiment, the model is trained on all the videos in LIVE and validated on IVPL for all distortion types. The detailed result is shown in Tab. V. As IVPL contains less distortion types, the differences between IVPL and LIVE seems less than those between CSIQ and LIVE. Hence, as we can see, both SDM and DeepVQA get an improvement for the overall performance compared to the validation experiment on CSIQ in Tab. IV. Moreover, though there are some differences between LIVE and IVPL, SDM trained on LIVE still obtains a good performance on IVPL, achieving over 0.88 on the both correlation coefficients. Based on the same training and validation data, SDM outperforms DeepVQA evidently, increasing by 14% on SROCC. And concerning each distortion type, SDM is more reliable than DeepVQA on each distortion, especially on MPEG and IP.

E. Effectiveness of SDP

To fully illustrate the importance of SDP in VQA, we further transplant the SDM to other VQA algorithms to test its effectiveness. For a simple and fast calculation, PSNR, SpEED, and MS-SSIM are adopted in this experiment. The frame-by-frame quality values of these methods are first calculated as the instantaneous degradations. And the rest parts of the proposed method is followed to obtain the predicted video quality. To comprehensively analyze the effectiveness, the experiment is conducted from two aspects. Firstly we test the performance of different QA algorithms on the same database (i.e., CSIQ). And then the performance of the same QA algorithm (i.e., MS-SSIM) on different databases is examined.

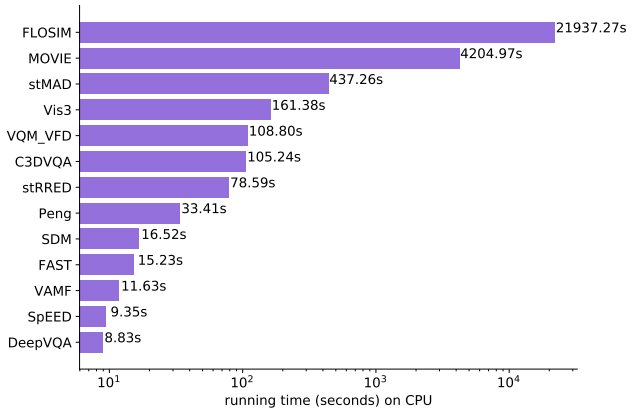
The experimental result is shown in Fig. 11. In Fig. 11(a), the performance (SROCC) of different methods on CSIQ



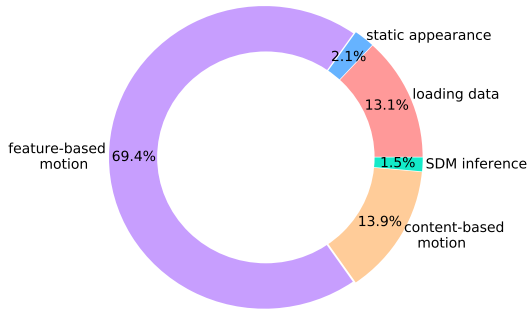


(b) SROCC of MS-SSIM on different databases

Fig. 11: Performance comparison between with or without SDM component. The blue is for the original performance (SROCC) without SDM, while the red for the one with SDM.



(a) Runtime comparison over FR-VQA methods



(b) Runtime analysis on each component

Fig. 12: Runtime analysis. All the methods are tested with CPU only.

database is first evaluated. For each method, the blue one stands for the original performance (without SDM), while the red one for that with SDM. As can be seen, taking the SDP into consideration, both the median SROCC and the worst performance are elevated for those algorithms. The improvements on these algorithms demonstrate that the SDP plays an important role during visual perception and the proposed SDM is effective. Further, it is found that, the improvement of MS-SSIM is much higher than those of PSNR and SpEED. The possible reason is that, MS-SSIM can provide a more reliable quality reference (The correlation between MS-SSIM and the human perception is higher than those of PSNR and SpEED as given in Tab. II). Since SDP suggests that the current perception is highly correlated with the previous, and the previous information can be a good indicator to predict the current status. Hence, the more convincing instantaneous degradations it provides, the more accurate the previous information is, thus the better result it can acquire with SDM. In this way, as MS-SSIM can generate relatively credible instantaneous degradations, the MS-SSIM+SDM can obtain a good indicator, thus performing much better. On the contrary, since the PSNR+SDM can be misguided by the unreliable instantaneous degradations from PSNR, the improvement of PSNR+SDM is limited.

In addition to the experiment of different methods on the same database, the performance of the same method on different databases is also examined. In this work, MS-SSIM is selected as the candidate, and the performance with or without SDM on the three databases (i.e., LIVE, CSIQ, and IVPL) is validated. As shown in Fig. 11(b), the performance with SDM has a notable advantage over the original ones, which shows again the importance of SDP in subjective perception and quality assessment tasks. It is also validated that SDM has a good capability to capture the SDP, and improves the performance of VQA methods on different databases. It is also noted that the improvements on LIVE and CSIQ are larger than that on IVPL. As explained above, the possible explanation is that the performance of MS-SSIM on IVPL is slightly worse than those on LIVE and CSIQ, and the prediction accuracy possibly affects this promotion. This is highly reasonable since SDM is surrounded by the correlation and prediction between the previous and the current. If the given indicator is not reliable, the performance is surely influenced. In short, in consideration of the SDP, the performance can be further promoted for various VQA methods and databases. Experimental results show the effectiveness of both the SDP during visual perception and the proposed SDM in the VQA task.

F. Runtime Analysis

Besides performance comparison, in this subsection, the runtime analysis is also conducted. All the methods are evaluated on an operation system of Windows 10 Enterprise (x64), with a 3.60 GHz Inter Core i7-4790 CPU and 16GB RAM. To maintain a fair comparison, all the methods are running with CPU only, including DeepVQA and C3DVQA, which are deep learning-based methods beneficial from GPU acceleration. The test samples are from LIVE video databases, with 250 frames and a resolution of 768×432. The time of data loading is also included in each algorithms, and the comparison over these FR-VQA methods is shown in Fig. 12.

As shown in Fig. 12(a), the proposed SDM achieves a better performance with acceptable time cost comparing to most of the methods. C3DVQA takes all the video data as input, and involves 3D convolution, which is much costly in both memory and time. DeepVQA only takes 10 frames only from each video thus performing faster than all the others. However, in training phase, both C3DVQA and DeepVQA would train and update their convolution kernels, which costs more time. The proposed SDM extracts related features once only, and the following phase of model training is much faster. As shown in Fig. 12(b), the actually SDM inference only takes up 1.5% (about 0.24s) of all the time consumption. Due to the calculation of optical flow maps, the most time-consuming bottleneck is feature-based motion characterization, which would be partially solved by hardware accelerating and algorithm updating in motion estimation. Also, As we have shown in Sec. IV-E, the instantaneous degradations can be totally generated by other more fast and advanced methods in the future, which still suits the framework of SDM.

V. Conclusion

Previous VQA work mainly concentrates on the degradation representation of motion information in videos, and in this work, we suggest that the temporal influence is still important for VQA. In this work, enlightened by the mechanism of serial dependence during human visual perception, we have suggested that the HVS prefers a sequentially dependent and predictable degradation than an interrupted and unpredictable one. Thus, we have tackled the VQA problem with serially dependent modeling, and proposed a novel FR-VQA framework. The instantaneous degradation has been first charactered through both explicit content-based and implicit feature-based motion representations. Then, we have proposed an A-LSTM for the serially dependent degradation modeling. Through an attention-based quality decision procedure, the video quality has been formulated. Experimental results have demonstrated that the proposed method achieves good consistency with the human perception.

References

[1] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 1, pp. 154–166, 2016.

[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conf. Record of the 37th Asilomar Conf. on*, vol. 2, 2003, pp. 1398–1402.

[3] W. Zhang and H. Liu, "Study of saliency in objective video quality assessment," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1275–1288, 2017.

[4] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, 2013.

[5] Y. Liu, G. Zhai, K. Gu, X. Liu, D. Zhao, and W. Gao, "Reduced-reference image quality assessment in free-energy principle and sparse representation," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 379–391, 2018.

[6] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, March 2014.

[7] Y. Liu, K. Gu, Y. Zhang, X. Li, G. Zhai, D. Zhao, and W. Gao, "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 929–943, 2020.

[8] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Am. A-Opt. Image Sci. Vis.*, vol. 24, no. 12, pp. 61–69, Dec. 2007.

[9] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, 2017.

[10] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[11] P. Peng, D. Liao, and Z. Li, "An efficient temporal distortion measure of videos based on spacetime texture," *Pattern Recognit.*, vol. 70, pp. 1–11, Oct. 2017.

[12] Y. Wang, T. Jiang, S. Ma, and W. Gao, "Novel spatio-temporal structural information based video quality metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 989–998, July 2012.

[13] L. He, W. Lu, C. Jia, and L. Hao, "Video quality assessment by compact representation of energy in 3D-DCT domain," *Neurocomputing*, vol. 269, pp. 108–116, Dec. 2017.

[14] P. V. Vu and D. M. Chandler, "ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, p. 013016, Feb. 2014.

[15] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment Algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.

[16] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin, "Quality assessment for video with degradation along salient trajectories," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2738–2749, Nov 2019.

[17] K. Gu, Z. Xia, J. Qiao, and W. Lin, "Deep dual-channel neural network for image-based smoke detection," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 311–323, 2020.

[18] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Objective video quality assessment combining transfer learning with CNN," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2716–2730, 2020.

[19] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: from spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 219–234.

[20] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3DVQA: full-reference video quality assessment with 3D convolutional neural network," in *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 4447–4451.

[21] J. Fischer and D. Whitney, "Serial dependence in visual perception," *Nat. Neurosci.*, vol. 17, no. 5, p. 738, 2014.

[22] A. Liberman, J. Fischer, and D. Whitney, "Serial dependence in the perception of faces," *Curr. Biol.*, vol. 24, no. 21, pp. 2569 – 2574, 2014.

[23] A. Kiyonaga, J. M. Scimeca, D. P. Bliss, and D. Whitney, "Serial dependence across perception, attention, and memory," *Trends Cogn. Sci.*, vol. 21, no. 7, pp. 493–497, 2017.

[24] C. Summerfield and F. P. de Lange, "Expectation in perceptual decision making: neural and computational mechanisms," *Nat. Rev. Neurosci.*, vol. 15, no. 12, pp. 816–816, 2014.

[25] L. Xu, W. Lin, L. Ma, Y. Zhang, Y. Fang, K. N. Ngan, S. Li, and Y. Yan, "Free-energy principle inspired video quality metric and its use in video coding," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 590–602, 2016.

[26] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 525–535, Jun. 2012.

[27] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 2, pp. 253–265, April 2009.

[28] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, 2013.

[29] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *2011 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, May 2011, pp. 1153–1156.

[30] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2351–2359.

[31] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *2019 IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 2349–2353.

[32] Z. Tu, C. J. Chen, L. H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," in *2020 IEEE Int. Conf. Image Process. (ICIP)*, 2020, pp. 141–145.

[33] D. Rahnev, A. Koizumi, L. Y. McCurdy, M. D'Esposito, and H. Lau, "Confidence leak in perceptual decision making," *Psychol. Sci.*, vol. 26, no. 11, pp. 1664–1680, 2015.

[34] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[35] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[36] G. Francis, "Cortical dynamics of visual persistence and temporal integration," *Perception & Psychophysics*, vol. 58, no. 8, pp. 1203–1212, 1996.

[37] D. Whitney, I. Murakami, and P. Cavanagh, "Illusory spatial offset of a flash relative to a moving stimulus is caused by differential latencies for moving and flashed stimuli," *Vision Res.*, vol. 40, no. 2, pp. 137–149, 2000.

[38] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: a highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, 2014.

[39] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Res.*, vol. 38, no. 5, pp. 743–761, 1998.

[40] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conf. Image Analysis*. Springer, 2003, pp. 363–370.

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

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[43] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan, "IVP subjective quality video database," *The Chinese University of Hong Kong, http://ivp. ee. cuhk. edu. hk/research/database/subjective*, 2011.

[44] Y. Pitrey, M. Barkowsky, R. Pépion, P. L. Callet, and H. Hlavacs, "Influence of the source content and encoding configuration on the perceived quality for scalable video coding," in *Human Vision and Electronic Imaging XVII*, vol. 8291, 2012, p. 82911K.

[45] S. Wolf and M. H. Pinson, "Video quality model for variable frame delay (VQM_VFD)," *US Dept. Commer., Nat. Telecommun. Inf. Admin., Boulder, CO, USA, Tech. Memo TM-11-482*, 2011.

[46] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, 2016.