CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/159299/

This is the author's version of a work that was submitted to / accepted for publication.

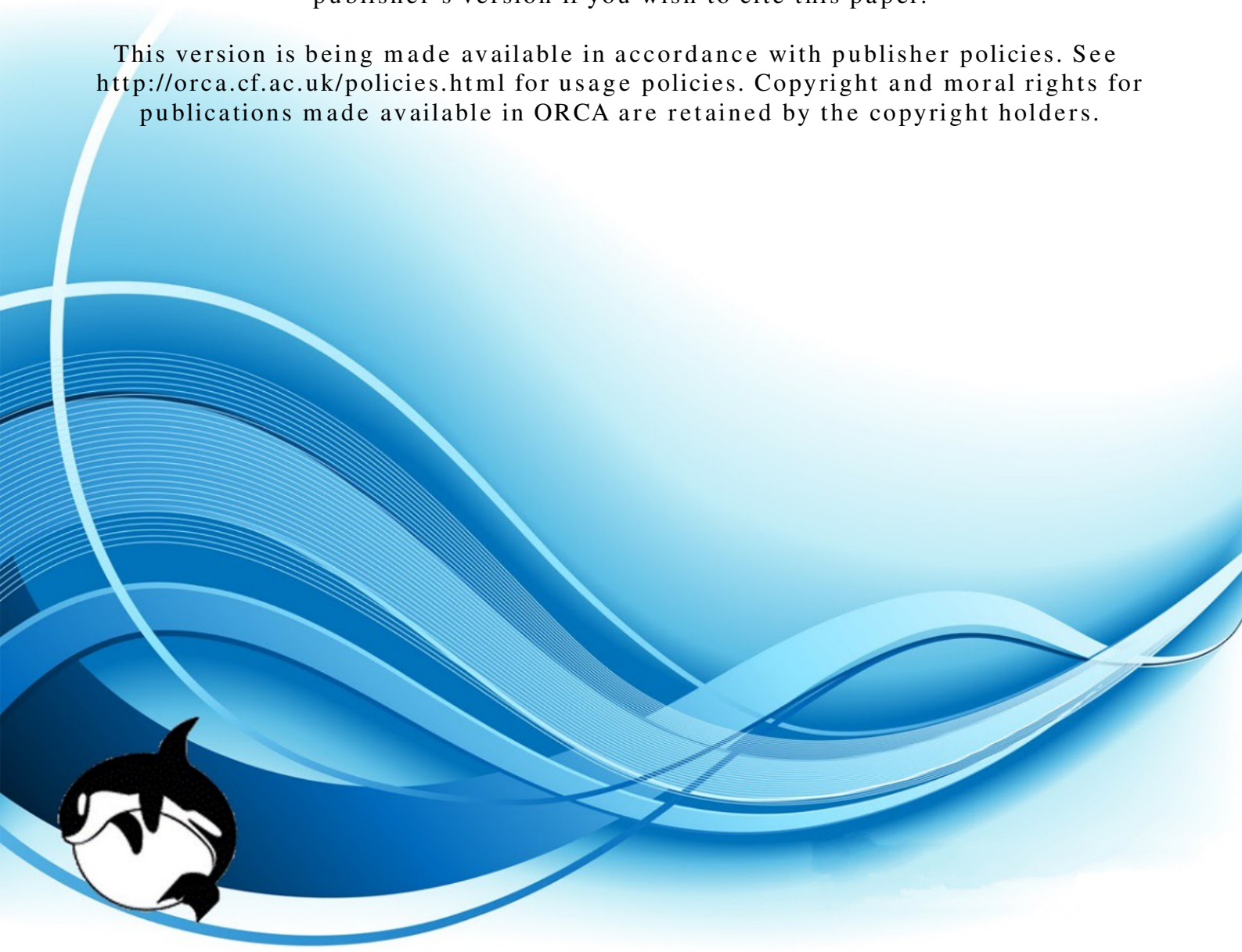
Citation for final published version:

Fang, Yuming, Li, Zhaoqian, Yan, Jiebin, Sui, Xiangjie and Liu, Hantao 2023. Study of spatio-temporal modeling in video quality assessment. IEEE Transactions on Image Processing 32 , pp. 2693-2702. 10.1109/TIP.2023.3272480 file

Publishers page: http://dx.doi.org/10.1109/TIP.2023.3272480

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



Study of Spatio-Temporal Modeling in Video Quality Assessment

Yuming Fang, *Senior Member, IEEE*, Zhaoqian Li, Jiebin Yan, Xiangjie Sui, Hantao Liu, *Member, IEEE*

***Abstract*—Video quality assessment (VQA) has received remarkable attention recently. Most of the popular VQA models employ recurrent neural networks (RNNs) to capture the temporal quality variation of videos. However, each long-term video sequence is commonly labeled with a single quality score, with which RNNs might not be able to learn long-term quality variation well. A natural question then arises: *What's the real role of RNNs in learning the visual quality of videos? Does it learn spatio-temporal representation as expected or just aggregating spatial features redundantly?* In this study, we conduct a comprehensive study by training a family of VQA models with carefully designed frame sampling strategies and spatio-temporal fusion methods. Our extensive experiments on four publicly available in-the-wild video quality datasets lead to two main findings. First, the plausible spatio-temporal modeling module (*i.e.*, RNNs) does not facilitate quality-aware spatio-temporal feature learning. Second, sparsely sampled video frames are capable of obtaining the competitive performance against using all video frames as the input. In other words, spatial features play a vital role in capturing video quality variation for VQA. To our best knowledge, this is the first work to explore the issue of spatio-temporal modeling in VQA.**

***Index Terms*—Video quality assessment, spatio-temporal modeling, recurrent neural network.**

I. Introduction

WITH the rapid development of digital devices and Internet, the amount of videos has tremendously increased in various areas, such as entertainment and video surveillance. At present, videos have become one of the important elements of entertainment in our daily life. However, videos may encounter quality degradation in the process of compression, storage, and transmission [1], which significantly degrades the quality of experience of the end users. Therefore, how to measure the quality of videos is critical in multimedia processing systems, since it can be used to optimize the video processing algorithms as well as performance monitoring of video processing systems. Video quality assessment (VQA) includes subjective assessment and objective assessment, where the former means quality evaluation by human beings, and the latter denotes designing objective VQA models for visual quality evaluation instead of human labors. Since VQA models can be easily embedded in the real applications, the research on VQA models has attracted much attention from both academia and industry communities.

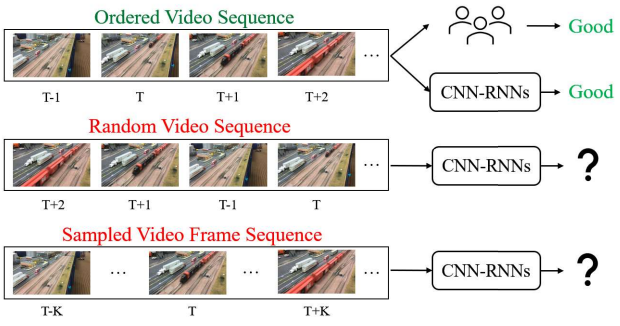


Fig. 1. Generally, the videos are viewed in order by human beings, thus it is straightforward to feed the video frames into the VQA model (*e.g.*, the popular framework: CNN-RNN) in sequence. To explore the plausible answer to whether it is necessary to follow the popular VQA model framework, we carefully design several experiments, including disrupting the order of input video frames and discarding partial video frames via sparse sampling.

According to the availability of reference video, objective VQA models can be classified into three types: full-reference VQA (FR-VQA) models [2]–[8], reduced-reference VQA (RR-VQA) models [9], and no-reference VQA (NR-VQA) models or blind-VQA models. The first two types of VQA models require full or partial reference information when being used to evaluate video quality, while the last type of VQA models do not need any reference information at hand and they have been widely used in the real applications. Traditional VQA models [10]–[12] mainly rely on extracting perceptually relevant hand-crafted features to quantify the visual quality of videos. With the prevalence of deep learning, there have been many deep neural network (DNN) based VQA models proposed recently [13]–[15]. Most of the DNN based VQA models follow the similar paradigm, in which a convolutional neural network (CNN) is used to capture spatial features, and the recurrent neural network (RNNs) are subsequently employed to learn the long-term quality variation of videos. However, a long-term video sequence is commonly labeled with a single quality score, with which RNNs might not be able to learn long-term quality variation well. As revealed in these studies [16], [17], feeding video frames disorderly in the associated models has little influence on prediction accuracy.

This work was supported in part by the National Natural Science Foundation of China under Grant 62132006, in part by the Natural Science Foundation of Jiangxi Province of China under Grants 20223AEI91002 and 20224BAB212012, in part by the project funded by China Postdoctoral Science Foundation under Grant 2022M721417, and in part by the Project of the Education Department of Jiangxi Province under Grant GJJ2200524. (Corresponding author: Jiebin Yan).

Yuming Fang, Zhaoqian Li, Jiebin Yan and Xiangjie Sui are with the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, Jiangxi, China (e-mail: fa0001ng@e.ntu.edu.sg; zhaoqianli.dev@foxmail.com; jiebinyan@foxmail.com; suixiangjie2017@163.com).

Hantao Liu is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF243AA, U.K (e-mail: liuh35@cardiff.ac.uk).

As shown in Figure 1, these facts naturally prompt us to arise a question: *Does the RNNs module contribute to learn quality-aware spatio-temporal features of videos?*

In this paper, we carefully design three types of experiments, where we choose a VQA model, *i.e.*, VSFA [13], as the backbone to construct various variants to testify our idea, since VSFA shows excellent performance in VQA. The first type of experiments is called *Frame-Level*, in which we shuffle the order of video frames in both training and testing stages, while other settings are the same with those in the original paper [13]. The second type of experiments is called *Cube-Level*, whose difference from the first type is that we replace the default input, *i.e.*, feature vectors extracted from ResNet-50 [18], with the spatio-temporal features using a 3D CNN [19]. The last type of experiments is named *Sample-Level*, where we attempt to take a close look at how sparse frame samplings affect the performance of VSFA. Our extensive experiments are conducted on four publicly available in-the-wild video quality datasets, including Konvid-1K [20], CVD2014 [21], LIVE-Qualcomm [22] and LIVE-VQC [23].

In sum, the main contributions of this work are shown as follows:

- We present the first study to deeply explore the relationship between spatio-temporal feature learning and VQA, aiming to investigate the core part in capturing video quality degradation by learned spatio-temporal features. All experiments are conducted based on the modifications of an effective VQA model.
- We carefully design three types of experiments, namely *Frame-Level*, *Cube-Level* and *Sample-Level*. The former two types of experiments are used to investigate the influence of disordered input on the VQA model, while the third type of experiments is employed to verify the redundancy of video frames for the VQA task.
- We conduct comprehensive experiments oriented by the input format and spatio-temporal information fusion, and conclude that spatial information of video plays a more important role in capturing video quality variation than the learned temporal features for VQA.

II. Related Work

In this section, we first review hand-crafted and deep learning-based spatio-temporal feature modeling methods that are commonly used in video processing tasks. We then briefly describe the NR-VQA models and the studies about evaluation of models' pros and cons.

A. Spatio-temporal Feature Modeling

Generally, a video contains rich static and dynamic information. Modeling the spatio-temporal features of the videos can automatically analyze the task-dependent information, it is therefore an essential module in many video processing tasks. Almost all of the early video spatio-temporal feature modeling algorithms are based on hand-crafted features. Specifically, according to specific methods of feature extraction, these algorithms can be divided into: optical flow algorithm based algorithms [6], [8], [24], [25], frame difference based algorithms [26], [27], background subtraction based algorithms [28]–[30], histogram of oriented gradients based algorithms (HOG) [31], and motion boundary histogram (MBH) based algorithms [32]. The performance of the deep learning based algorithms gradually surpasses the traditional algorithms, and hand-crafted algorithms gradually fade out of our attention.

Deep learning provides an automatic way to model the spatio-temporal features of videos. In particular, some excellent network architectures, *e.g.*, CNNs, long short-term memory (LSTM) network [33], and gate recurrent unit (GRU) [34], make the performance of the deep learning-based algorithms gradually surpass traditional algorithms. The deep learning-based spatio-temporal feature modeling algorithms can be roughly divided into three types, including the two-stream based networks [16], [35]–[37], 3D CNNs based networks [38]–[41], and CNN-RNNs based networks [13], [14], [42], [43]. The two-stream based networks [16], [35]–[37] mainly include two modules, one of which extracts spatial features and the other one extracts temporal features. The 3D CNNs based networks [38]–[41] employ 3D CNNs to extract spatio-temporal features. The CNN-RNNs based networks [13], [14], [42], [43] employ the pre-trained CNNs to extract the spatial features, and then use the GRU/LSTM module to extract temporal information.

B. NR-VQA Models

The NR-VQA models could be divided into two types, including distortion-specific models and general-purpose models. In [44], Brandão *et al.* proposed a NR-VQA model for compressed videos. The model first estimates errors and then weights the errors to obtain the final objective score. Wu *et al.* [45] proposed a NR-VQA model to evaluate the user experience of streaming videos. It utilizes the decoded video to extract image features, then uses a linear model to obtain a quantitative quality score. Liu *et al.* [46] proposed a deep learning-based multi-task model, called V-MEON, which utilizes the 3D CNNs to capture spatio-temporal information for video codec classification and quality assessment.

Moreover, with the popularity of the network and the development of multimedia techniques, a growing number of videos are generated by mobile phones and digital cameras. These videos are usually captured in the wild, and they may suffer from authentic distortions. Traditional general-purpose VQA methods commonly extract low-level features to predict the visual quality. Saad *et al.* [11] proposed a NR-VQA model namely V-BLINDS, which extracts spatio-temporal features in the discrete cosing transform (DCT) domain and quantifies motion coherency with a motion model. Mittal *et al.* [12] proposed to use the intrinsic statistical regularities to predict video quality scores. Tu *et al.* [47] proposed a fusion-based model called VIDEVAL, which employs a feature selection strategy and a support vector regressor (SVR) as the regression model to learn the feature-to-score mapping. Compared with traditional NR-VQA methods, the deep learning-based models can effectively capture high-level semantic features of videos, and show the better performance. Li *et al.* [48]

proposed a NR-VQA metric based on the spatio-temporal natural video statistics in the 3D discrete cosine transform (3D-DCT) domain. Zhang *et al.* [49] proposed a weakly supervised learning-based model with an eight-layer CNNs module and a resampling strategy. VSFA [13] and RIRNet [14] are CNN-RNNs based networks, which first employ CNNs to explicitly capture the spatial semantic features of the video frames and then use RNNs to focus on learning the temporal quality-aware features.

C. Evaluation of Models' Pros and Cons

In recent years, with the development of deep learning theory and hardware technology, computer vision has made breakthrough progress. More and more models have been proposed to solve problems encountered in the real world applications. Although deep learning models have shown excellent performance in many fields, models' robustness is still a big challenge due to dataset scale [20], [23], [50], data leakage [51], adversarial attack [52]–[54], and model architecture [16], [17], [55]. For these problems encountered in computer vision community, researchers in different research fields have proposed analytical methods from different perspectives. Götz-Hahn *et al.* [51] pointed out that some models achieved state-of-the-art performance due to data leaks in fine-tuning and quality prediction stages. Xie *et al.* [55] built an effective and efficient video classification system by exploring the relationship between 2D CNN and 3D CNN, taking into account the relationship between the model's speed and accuracy. In addition, a growing number of works have explored the relationship between the labels in the dataset and the prediction accuracy of models. Hoiem *et al.* [56] proposed a method that studies the relationship between object features and detection performance, as well as the frequency and impact of different types of false positives. In order to explore how to quantify a semantic segmentation model, Csurka *et al.* [57] proposed a new evaluation criterion based on contours, which is suitable for unsupervised semantic segmentation models. Besides, some researchers introduced the concept of MAximum Discrepancy (MAD) [58] in evaluation IQA [59] and semantic segmentation [60] models.

III. The Proposed Framework for Diagnosing VQA

A. Motivation

Spatio-temporal feature modeling has been a long-standing problem in video understanding field [19], [36], [41], [55], [61], and it is one key element of the video understanding models. In the video understanding tasks which pursue for the balance between effectiveness and efficiency, some researchers tried to reduce the number of input frames by sparsely sampling with the consideration of the fact that there exists much redundant information in consecutive frames, and achieved promising performance [62]–[64]. In addition, some studies [55], [65] found that the validity of video understanding models may depend on particular databases, *e.g.*, the videos in the something-something database [66] are temporally relevant while that in some databases [67], [68] are not. Motivated by these video understanding studies [55], [62]–[65], in this paper, we focus on studying two problems that have been ignored in VQA: (1) Does the spatio-temporal module contribute to learn quality-aware spatio-temporal features? (2) Does the VQA model require all video frames as the input?

To answer these two questions, we carefully design three types of experiments, named *Frame-Level*, *Cube-Level*, and *Sample-Level*. All experiments are conducted on four datasets, including KoNViD-1k [20], CVD2014 [21], LIVE-Qualcomm [22], and LIVE-VQC [23]. For these three types of experiments, VSFA [13] is selected as the backbone with the corresponding modifications. The first two sets of experiments, *i.e.*, *Frame-Level* and *Cube-Level*, are used to explore to what extent the popular spatio-temporal feature modeling module contributes to capture video quality variation. To be more specific, we input frames in temporal order or disorderly when training and testing on the four datasets. The last type of experiment aims to explore whether it is possible to use only a few frames to capture video quality variation, where we adopt different sampling strategies in the training stage and investigate whether there is a significant difference among the results of different sampling strategies.

B. The Detailed Framework

In this section, we explain three variants of VSFA designed for the aforementioned three types of experiments. Each variant uses different frame input modes and sampling strategies. The detailed framework is summarized in Figure 2.

In the *Frame-Level* experiments, we use the original VSFA [13] model for training and testing. The original VSFA uses the ResNet-50 [18] as spatial feature extractor and a Gated Recurrent Unit (GRU) as temporal feature extractor. During training, validation, and testing stages, we feed all frames of each video into VSFA.

In order to explore the impact of short-term video sequences on the VQA model, we design the *Cube-Level* experiments by using 3D CNNs. We replace ResNet-50 in the VSFA model with 3D ResNet-101 [19] in feature extraction stage. Each video sequence is divided into K cubes, while each cube contains 16 video frames. 3D ResNet-101 is used to extract the features of 16 video frames, and the extracted features are then fed into the following module.

The model in the *Sample-Level* experiments is the same as that in the *Frame-Level* experiments. The difference is that we randomly extract video frames with different sampling strategies (see Section IV-C for details) as the input. The remaining experimental configurations are the same as that in the *Frame-Level* experiments.

IV. Experimental Methodology

A. Datasets

We conduct experiments on four in-the-wild datasets: KoNViD-1k [20], CVD2014 [21], LIVE-VQC [23], LIVE-Qualcomm [22]. Table I summarizes these four datasets.

- CVD2014 [21] video quality database uses complex distortions of real cameras introduced during video acquisition rather than introducing distortions via post-processing. The quality dimensions of videos include

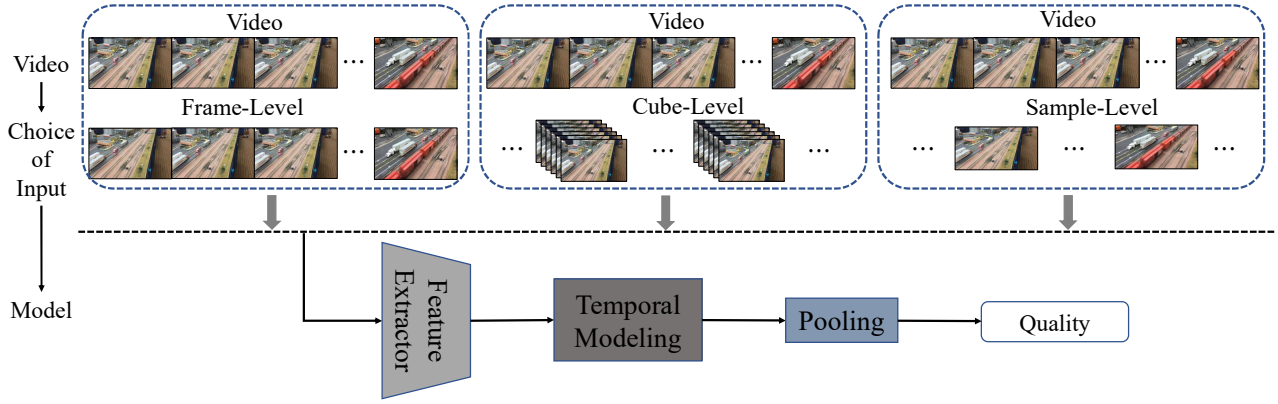


Fig. 2. The proposed framework for diagnosing VQA. Three sets of experiments provide three models, respectively: Frame-Level model, Cube-Level model and Sample-Level model. Each model uses different frame input modes and sampling strategies.

TABLE I
SUMMARY OF DATASETS

Name	Year	Resolution	Number	Distrotion Type	Format	MOS
KoNViD-1k [20]	2017	540p	1200	In-the-wild	MP4	1.220-4.640
LIVE-Qualcomm [22]	2017	1080p	208	In-the-wild	YUV420	16.562-73.643
CVD2014 [21]	2016	720p, 480p	234	In-the-wild	AVI	-6.500-93.380
LIVE-VQC [23]	2019	1080p-240p	585	In-the-wild	YUV420	6.224-94.287

sharpness, graininess, color balance, darkness, and jerkiness. The database consists of 234 videos captured by 87 cameras. The videos last $10s$-$25s$ with 11-31 frames per second (FPS). The MOSs range from -6.500 to 93.380.

- LIVE-Qualcomm [22] video quality database consists of 208 videos with the resolution of 1920×1080. The videos are captured by 8 different portable devices. The database includes six types of in-capture distortions, including artifacts, color, exposure, focus, sharpness, and stabilization. Each video lasts $15s$ with 30 FPS. The MOSs range from 16.562 to 73.643.
- KoNViD-1k [20] consists of 1200 public-domain videos with the resolution of 960×540. The videos come from a sizeable public video dataset, *i.e.*, YFCC100m [69]. The videos suffer from authentic distortions, *e.g.,* motion blur and color distortion, and each video lasts 8s with $24/25/30$ FPS. The MOSs range from 1.220 to 4.640.
- LIVE-VQC [23] video quality database consists of 585 videos, which are captured using 101 different devices. The videos are distorted by authentic distortions, such as poor exposures, motion blur, haziness, imperfect color representations, compression artifacts, *etc*. The MOSs range from 6.224 to 94.287.

B. Backbone

Here, we describe the backbone, *i.e.*, VSFA [13], used in our framework in detail. It consists of two sub-modules: a feature extraction module and a temporal modeling module. In the feature extraction module, VSFA uses ResNet-50 [18] to extract the content-aware feature maps from its top convolutional layer for each video frame. After feature extraction, VSFA uses a temporal modeling module for long-term dependency modeling. The detailed framework can refer to the original paper [13]. Specifically, for the feature maps extracted from each video frame, the 4096-dimensional content-aware feature vector is obtained by two spatial global pooling (GP) operations, which include a spatial global average pooling ($\mathrm{GP}_{\mathrm{mean}}$) and a spatial global standard deviation pooling ($\mathrm{GP}_{\mathrm{std}}$). Then, considering the high dimension of content-aware features, a dimension reduction operation is added before the GRU. A single fully connected layer is used as the dimension reduction module, so as to reduce the 4096-dimensional feature vector to a 128-dimensional feature vector. After that, the reduced feature vectors are fed into the GRU. The output of the GRU is used to predict the frame-level quality score. Finally, by using the subjectively-inspired temporal pooling method to fuse frame-level quality scores, the network can get the final video-level quality score.

C. Implementation Details

In this section, we explain the experimental details, including training/validation/testing splits, frame input modes, sampling strategies, model structures, training details, and evaluation criteria.

Datasets For each dataset, we use 60% videos for training, 20% videos for validation, and 20% videos for testing. The videos in these three sub-sets are non-overlapping. Each experiment repeats ten times, and the mean and standard deviation (std) of the results of the ten runs are reported.

Frame Input Modes. In training and testing stages, we design two input modes. The first is to feed the frames into the network in the temporal order, while the second is to input the frames into the network disorderly. Since the model includes the training, validation, and testing stages, there are two input

modes in the training stage and two input modes in the testing stage while keeping the same frame input mode in the training and validation stages. Therefore, four frame input modes and four types of experimental results can be obtained on each dataset (see Tables II, III and IV for detailed experimental results):

1. Input the video frames in temporal order during training and testing stages. This mode is denoted by *OO*.
2. Input the video frames in shuffled order during training stage and in temporal order during testing stage. This mode is denoted by *SO*.
3. Input the video frames in order during training stage and shuffle the video frames randomly during testing stage. This mode is denoted by *OS*.
4. Input the video frames in the shuffled order during training and testing stages. This mode is denoted by *SS*.

We choose four frame input modes for *Frame-Level* and *Cube-Level* experiments. In *Sample-Level* experiment, we just test the first frame input mode.

Sampling Strategies. We design four sampling strategies: sample-0, sample-4, sample-8, sample-16. Sample-0 denotes that the entire video sequence is used as the input of the network. Sample-K (K=4, 8, or 16) means that a video is divided into N cubes and each cube contains K frames. The value of N depends on frame number and K. If the last cube's frame number is less than K frames, we let the last cube to contain the last K frames of a video. Then, we randomly select a frame from each cube, and these frames are fed into the network as a video sequence. We choose sample-0 as the sampling strategy for *Frame-Level* and *Cube-Level* experiments.

Training. We use PyTorch [70] to implement the three variants. In *Frame-Level* and *Sample-Level* experiments, the ResNet-50 [18] pre-trained on ImageNet [71] is chosen for feature extraction. The $L1$ loss and Adam [72] optimizer with an initial learning rate of 0.00001 are used for training models in the *Frame-Level* and *Sample-Level* experiments. The batch size is set to 128 for the *Frame-Level* and 256 for *Sample-Level* experiments, respectively.

In the *Cube-Level* experiments, the 3D ResNet-101 [19] pre-trained on Kinetics [67] is chosen for feature extraction. In order to match the dimension of the temporal modelling module, we use 3D ResNet-101 to extract 2048 feature maps for each cube. Then, through spatial global pooling (GP), 2048 feature maps will get a 4096-dimensional feature vector for each cube. We use $L1$ loss and Adam [72] optimizer with an initial learning rate of 0.00001 and set the training batch size as 4096. During the training stage, we freeze the parameters in pre-trained ResNet-50 [18] and 3D ResNet-101 [19]. For the temporal modeling stage, we select the GRU [34] and the same pooling strategy adopted in VSFA.

Evaluation Criteria. In the modeling of the objective VQA method, Spearman's rank-order correlation coefficient (SROCC), Kendall's rank-order correlation coefficient (KROCC), Pearson's linear correlation coefficient (PLCC), and root mean square error (RMSE) are used as performance evaluation criterion. Among them, SROCC and KROCC are used to evaluate the monotonicity of model predictions. PLCC and RMSE are used to evaluate the accuracy of the model prediction. The larger the absolute value of SROCC, KROCC, and PLCC, the better, while the smaller the value of RMSE, the better. As suggested by Video Quality Experts Group (VQEG) [73], before calculating PLCC and RMSE values, we adopt a four-parameter logistic function for mapping the objective score to the subjective score s:

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

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

V. Diagnostic Analysis

A. Analysis of Temporal Information

In the *Frame-Level* experiments, we train the original VSFA [13] model on each dataset to verify whether the temporal module can effectively extract the quality-aware temporal information among video frames. The experimental results are shown in Tables II, III and IV. To intuitively observe the results showed in Tables II, III and IV, we show the difference between the SROCC results of the variant with default input mode (*i.e.*, input video frames into the network in order in training and testing stages) of that of the variant with other input modes in Figure 3(a).

As shown in Figure 3, *OS*, *SO* and *SS* represent the difference between the SROCC results of this mode and the *OO* mode. The positive part of the Y-axis denotes that experiments' performance of the variants with other modes is better than that of the variant with default input mode and vice versa. In the KoNViD dataset, the performance of the variant with the third input mode improves by 0.38% compared with that of the variant with default input mode. In the CVD2014 dataset, the performance of the variant with the second input mode improves by 0.96% compared with that of the variant with default input mode. In the LIVE-Qualcomm dataset, the performance of the variant with the second input mode improves by 1.53% compared with that of the variant with default input mode. In the LIVE-VQC dataset, the performance of the variant with the second input mode improves by 1.90% compared with that of the variant with default input mode. As this regard, we can find that the temporal module can not effectively extract the quality-aware temporal information between video frames. This may result from the possibility that the spatial features contain hints on temporal features, *e.g.,* the motion blur can be captured by the spatial features extracted from a single frame.

In the *Cube-Level* experiments, we explore the relationship between short-term temporal information and quality-aware video features. The existing video understanding tasks show that 3D CNNs can extract short-term temporal and spatial information, thus, we use 3D ResNet-101 instead of ResNet-50 to extract video frame features in the *Cube-Level* experiments. The results are shown in Table III and Figure 3(b). It is worth noting that, with the experimental results of the *Frame-Level* and *Cube-Level* experiments, we find that the experimental results of the *Cube-Level*-based model are worse than those

TABLE II
PERFORMANCE COMPARISON OF THE FRAME-LEVEL EXPERIMENTS. MEAN AND STANDARD DEVIATION (STD) OF THE PERFORMANCE VALUES IN 10 RUNS ARE REPORTED.

	Mode	Frame-Level			
		SROCC	PLCC	KROCC	RMSE
KoNViD-1k [20]	OO	0.779 (0.030)	0.798 (0.027)	0.586 (0.029)	**0.385 (0.021)**
	OS	0.777 (0.033)	0.796 (0.029)	0.584 (0.031)	0.387 (0.022)
	SO	**0.782 (0.030)**	**0.799 (0.024)**	**0.588 (0.028)**	**0.385 (0.019)**
	SS	0.780 (0.031)	0.798 (0.026)	0.587 (0.030)	0.386 (0.020)
CVD2014 [21]	OO	0.832 (0.031)	0.847 (0.028)	0.645 (0.034)	11.202 (1.130)
	OS	**0.840 (0.031)**	**0.851 (0.028)**	**0.652 (0.033)**	**11.066 (1.128)**
	SO	0.825 (0.032)	0.841 (0.025)	0.638 (0.035)	11.428 (1.187)
	SS	0.833 (0.027)	0.848 (0.026)	0.648 (0.030)	11.183 (1.110)
LIVE-Qualcomm [22]	OO	0.707 (0.073)	0.742 (0.066)	0.516 (0.066)	8.377 (0.934)
	OS	**0.718 (0.073)**	**0.753 (0.063)**	**0.526 (0.067)**	**8.217 (0.940)**
	SO	0.699 (0.074)	0.723 (0.081)	0.507 (0.069)	8.601 (1.094)
	SS	0.713 (0.075)	0.742 (0.070)	0.522 (0.071)	8.354 (1.029)
LIVE-VQC [23]	OO	0.686 (0.038)	0.744 (0.039)	0.500 (0.034)	11.324 (0.536)
	OS	0.697 (0.036)	0.749 (0.035)	0.509 (0.032)	11.236 (0.502)
	SO	0.686 (0.040)	0.745 (0.041)	0.501 (0.036)	11.293 (0.546)
	SS	**0.699 (0.035)**	**0.750 (0.038)**	**0.512 (0.032)**	**11.212 (0.507)**

TABLE III
PERFORMANCE COMPARISON OF THE CUBE-LEVEL EXPERIMENTS. MEAN AND STANDARD DEVIATION (STD) OF THE PERFORMANCE VALUES IN 10 RUNS ARE REPORTED.

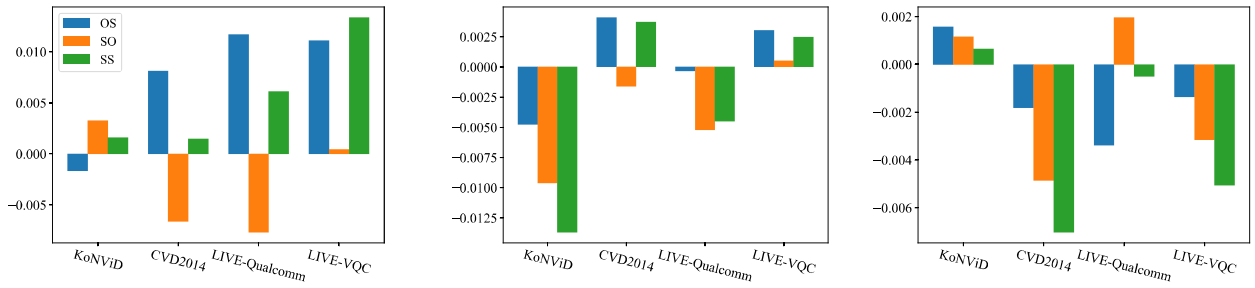
	Mode	Cube-Level			
		SROCC	PLCC	KROCC	RMSE
KoNViD-1k [20]	OO	**0.589 (0.040)**	**0.604 (0.038)**	**0.415 (0.032)**	**0.511 (0.025)**
	OS	0.585 (0.037)	0.599 (0.035)	0.411 (0.029)	0.513 (0.023)
	SO	0.580 (0.040)	0.596 (0.037)	0.408 (0.031)	0.515 (0.025)
	SS	0.576 (0.037)	0.592 (0.034)	0.404 (0.029)	0.517 (0.023)
CVD2014 [21]	OO	0.750 (0.066)	0.762 (0.069)	0.549 (0.063)	13.479 (1.587)
	OS	**0.754 (0.066)**	**0.766 (0.070)**	0.552 (0.063)	**13.385 (1.584)**
	SO	0.749 (0.067)	0.758 (0.067)	0.548 (0.063)	13.584 (1.498)
	SS	**0.754 (0.065)**	0.763 (0.072)	**0.554 (0.061)**	13.435 (1.582)
LIVE-Qualcomm [22]	OO	**0.524 (0.135)**	0.575 (0.115)	0.371 (0.100)	10.215 (1.417)
	OS	**0.524 (0.135)**	0.581 (0.121)	**0.374 (0.104)**	**10.133 (1.410)**
	SO	0.519 (0.139)	0.569 (0.106)	0.369 (0.104)	10.286 (1.358)
	SS	0.520 (0.134)	**0.583 (0.112)**	0.369 (0.102)	10.139 (1.347)
LIVE-VQC [23]	OO	0.544 (0.054)	0.619 (0.055)	0.387 (0.044)	**13.303 (0.473)**
	OS	**0.547 (0.052)**	0.615 (0.053)	0.390 (0.043)	13.359 (0.508)
	SO	0.544 (0.048)	0.618 (0.045)	0.389 (0.041)	13.339 (0.423)
	SS	0.546 (0.049)	**0.620 (0.046)**	**0.390 (0.041)**	13.317 (0.489)

TABLE IV
PERFORMANCE COMPARISON OF THE SAMPLE-16 EXPERIMENTS. MEAN AND STANDARD DEVIATION (STD) OF THE PERFORMANCE VALUES IN 10 RUNS ARE REPORTED.

	Mode	Sample-16			
		SROCC	PLCC	KROCC	RMSE
KoNViD-1k [20]	OO	0.774 (0.022)	0.790 (0.019)	0.580 (0.020)	0.393 (0.014)
	OS	**0.775 (0.027)**	0.792 (0.023)	**0.581 (0.025)**	0.391 (0.018)
	SO	**0.775 (0.027)**	**0.793 (0.024)**	**0.581 (0.024)**	**0.390 (0.018)**
	SS	0.774 (0.027)	0.791 (0.023)	0.580 (0.025)	0.392 (0.018)
CVD2014 [21]	OO	**0.868 (0.033)**	0.871 (0.037)	**0.691 (0.043)**	10.226 (1.100)
	OS	0.866 (0.035)	0.872 (0.036)	0.688 (0.041)	10.227 (1.050)
	SO	0.863 (0.033)	**0.878 (0.023)**	0.685 (0.041)	**10.070 (0.970)**
	SS	0.861 (0.036)	**0.878 (0.024)**	0.681 (0.041)	10.083 (0.986)
LIVE-Qualcomm [22]	OO	0.707 (0.078)	**0.750 (0.063)**	**0.519 (0.069)**	**8.276 (0.947)**
	OS	0.703 (0.082)	0.736 (0.065)	0.513 (0.076)	8.476 (1.042)
	SO	**0.708 (0.080)**	0.734 (0.062)	0.518 (0.072)	8.516 (1.040)
	SS	0.706 (0.084)	0.731 (0.060)	0.515 (0.077)	8.558 (1.024)
LIVE-VQC [23]	OO	**0.693 (0.040)**	**0.749 (0.046)**	**0.505 (0.037)**	**11.206 (0.588)**
	OS	0.692 (0.039)	**0.749 (0.040)**	**0.505 (0.036)**	11.216 (0.534)
	SO	0.690 (0.037)	0.745 (0.045)	0.503 (0.036)	11.287 (0.585)
	SS	0.688 (0.035)	0.745 (0.039)	0.501 (0.033)	11.292 (0.520)

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT SAMPLING STRATEGIES ON THE FOUR VQA DATASETS. MEAN AND STANDARD DEVIATION (STD) OF THE PERFORMANCE VALUES IN 10 RUNS ARE REPORTED.

	Sample	SROCC	PLCC	KROCC	RMSE
KoNViD-1k [20]	0	**0.779 (0.030)**	**0.798 (0.027)**	0.586 (0.029)	**0.385 (0.021)**
	4	**0.779 (0.029)**	0.795 (0.026)	0.585 (0.027)	0.388 (0.020)
	8	0.781 (0.027)	0.796 (0.024)	**0.587 (0.026)**	0.388 (0.018)
	16	0.774 (0.022)	0.790 (0.019)	0.580 (0.020)	0.393 (0.014)
CVD2014 [21]	0	0.832 (0.031)	0.847 (0.028)	0.645 (0.034)	11.202 (1.130)
	4	0.834 (0.034)	0.848 (0.032)	0.646 (0.035)	11.177 (1.305)
	8	0.839 (0.034)	0.839 (0.042)	0.653 (0.034)	11.413 (1.384)
	16	**0.868 (0.033)**	**0.871 (0.037)**	**0.691 (0.043)**	**10.226 (1.100)**
LIVE-Qualcomm [22]	0	0.707 (0.073)	0.742 (0.066)	0.516 (0.066)	8.377 (0.934)
	4	0.713 (0.075)	0.750 (0.062)	0.522 (0.069)	8.270 (0.941)
	8	**0.714 (0.070)**	**0.756 (0.063)**	**0.524 (0.062)**	**8.186 (0.962)**
	16	0.707 (0.078)	0.750 (0.063)	0.519 (0.069)	8.276 (0.947)
LIVE-VQC [23]	0	0.686 (0.038)	0.744 (0.039)	0.500 (0.034)	11.324 (0.536)
	4	0.694 (0.035)	0.745 (0.038)	0.507 (0.031)	11.291 (0.456)
	8	**0.698 (0.039)**	**0.750 (0.038)**	**0.510 (0.035)**	**11.201 (0.464)**
	16	0.693 (0.040)	0.749 (0.046)	0.505 (0.037)	11.206 (0.588)



(a) Frame-Level (b) Cube-Level (c) Sample-16

Fig. 3. To verify whether the use of RNN network can effectively extract the temporal information of video quality, we conducted three types of experiments. For each type of experiments, we can get four results on each dataset. We plot their SROCC results relative to the first experiment.

TABLE VI
QUALITATIVE AND QUANTITATIVE ANALYSIS OF TEMPORAL INFORMATION EXPERIMENTAL RESULTS USING THE PAIRED T-TEST

	Mode	KoNViD-1k	CVD2014	LIVE-Qualcomm	LIVE-VQC
Frame-Level	OO	-	-	-	-
	OS	0.1880	0.0377	0.0013	0.0006
	SO	0.0079	0.3951	0.3268	0.8614
	SS	0.0648	0.4907	0.4165	0.0001
Cube-Level	OO	-	-	-	-
	OS	0.1440	0.3654	0.9409	0.0006
	SO	0.0001	0.5931	0.2097	0.8614
	SS	0.0126	0.4814	0.3692	0.0001
Sample-16	OO	-	-	-	-
	OS	0.8893	0.5914	0.4535	0.5244
	SO	0.7269	0.1950	0.7006	0.1318
	SS	0.7871	0.1972	0.9383	0.1398

of the *Frame-Level* model. We think there are two factors that affect the final experimental results: first, spatial information plays a more important role in capturing video quality degradation than temporal information, as demonstrated by the experimental results; second, as shown in relevant studies, the features of ResNet-50 trained on ImageNet are effective in capturing quality variations [74], [75]. In the CVD2014 dataset, the performance of the variant with the second input mode improves by 0.53% compared with that of the variant with default input mode. In the LIVE-VQC dataset, the performance of the variant with the second input mode improves by 0.55% compared with that of the variant with default input mode. In KoNViD and LIVE-Qualcomm datasets, the results of the variants with other input modes are still competitive. In this experiment, we do not find any strong connection between short-term temporal information and quality-aware video features.

In the Sample-16 experiments, we explore the relationship between *Frame-Level* and *Sample-Level*, and choose sample-16 sampling strategy and training configuration of *Frame-Level* to get four results. Table IV and Figure 3(c) show the experimental results. The training results on all four datasets

TABLE VII
QUALITATIVE AND QUANTITATIVE ANALYSIS OF SAMPLING STRATEGIES EXPERIMENTAL RESULTS USING THE PAIRED T-TEST

	Sample	KoNViD-1k	CVD2014	LIVE-Qualcomm	LIVE-VQC
Sample-Level	0	-	-	-	-
	4	0.9858	0.6619	0.1568	0.0015
	8	0.5265	0.3027	0.2380	0.0010
	16	0.3556	0.0005	0.9986	0.0862

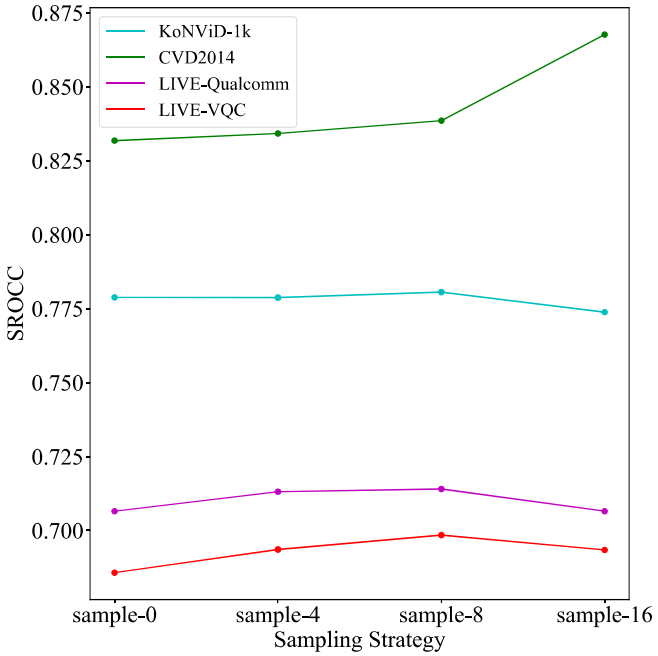


Fig. 4. The influence of different sampling strategies in different datasets on SROCC

are very close. Thus, we argue that RNNs could not make use of temporal information effectively in the context of VQA. In summary, we can draw a conclusion that the spatio-temporal module (*i.e.*, RNNs) contributes little to learn quality-aware spatio-temporal features.

It is worth noting that the conclusion may be only appropriate in the context of "in-the-wild" distorted videos, which suffer from authentic distortions, since the temporal distortion, in most cases (at least in the current context), can be regarded as the by-product of the discontinuity between successive frames. These test datasets don't consider the temporal artifacts, *e.g.,* frame freeze and transmission errors, which are often studied in the research on quality of experience of video streaming [76]–[78]. Thus, whether it can be generalized to other research filed like quality of experience of video streaming is still an open problem.

B. Analysis of Sampling Strategies

The above mentioned experiments show that the 3D CNNs and RNNs could not make use of temporal features of VQA datasets. In the third set of experiments, we verified whether partial frames could be used for video quality prediction.

Table V summarizes the performance of four different sampling strategies on the four datasets, which shows that training with partial frames does not significantly impact the experimental results, and the performance even improves in some cases. To visualize the results of Tabel V, Figure 4 shows SROCC results of different sampling strategies on the four datasets. In Figure 4, for the KoNViD-1k dataset, we can essentially see that there are some fluctuations in performance as the sampling strategy changes. We can also see that compared with sample-0, sample-8 achieves a higher SROCC value, and sample-16 has the lowest SROCC value. In contrast, the SROCC value of sample-4 is almost the same as that of sample-0. The top of Figure 4 shows the training results of CVD2014 dataset. Interestingly, as the number of video frames decreases, SROCC increases. More specifically, the sampling strategy of sample-16 achieves the best performance, while using the entire video as the input of a model achieves the worst performance. For the LIVE-Qualcomm dataset, we can find that sample-8 achieves the best performance, and sample-0 and sample-16's SROCC values are almost the same. As shown in the bottom of Figure 4, the results of the training in the LIVE-VQC are also surprising. The sample-8 achieves the best training effect, while sample-0 has the lowest SROCC, and the performance of sample-4 and sample-16 are almost the same. Through the above extensive experiments, we argue that sparse sampling strategy can be used to predict video quality.

C. Statistical Significance

In order to qualitatively and quantitatively analyze the effect of temporal information and sampling strategies on model performance, the paired t-test is used to test the statistical characteristics of the experiments on each dataset for significance. Specifically, a paired t-test was performed on the experimental results among SROCC values at the 5% significance level, and the results are shown in the Table VI and Table VII. When the paired t-test value is less than the significance level of 0.05, it indicates that there is a significant difference in the performance of the models, and vice versa, there is no significant difference. As shown in Table VI, we could argue that RNNs could not make use of temporal information effectively in the context of VQA. As shown in Table VII, the models trained using the sparse sampling strategy do not have a significant negative impact on the experimental results, and the performance of the models trained using the sparse sampling strategy is even significantly improved on some datasets, so we argue that sparse sampling strategy can be used to predict video quality.

VI. Conclusion

In this paper, we conduct a thorough study on the effectiveness of RNNs on learning quality-aware spatio-temporal features, aiming to explore the plausible answer to whether the current dominant design scheme of VQA model is necessary or not. Specifically, we test various spatio-temporal modeling strategies with the associated input data format, including *Frame-Level*, *Cube-Level*, and *Sample-Level*. Based on extensive experiments, some interesting findings can be clearly observed. First, the spatio-temporal network can not learn temporal information for video quality prediction effectively, at least on the tested four databases. Second, there are too many redundant frames in the video that we can use to predict the video quality without using the entire video data. For video quality prediction, we have to design some more effective models to extract useful spatio-temporal features for video quality prediction, and promote the development of other relevant video processing techniques.

References

[1] Z. Wang and A. Rehman, "Begin with the end in mind: A unified end-to-end quality-of-experience monitoring, optimization and management framework," in *SMPTE Motion Imaging Journal*, 2017, pp. 1–11.

[2] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, vol. 7, no. 2, 2005, pp. 2117–2128.

[3] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2009.

[4] P. V. Vu and D. M. Chandler, "ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, no. 1, pp. 013 016–013 016, 2014.

[5] S. Hu, L. Jin, H. Wang, Y. Zhang, S. Kwong, and C.-C. J. Kuo, "Objective video quality assessment based on perceptually weighted mean squared error," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 9, pp. 1844–1855, 2016.

[6] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480–2492, 2016.

[7] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2256–2270, 2018.

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

[9] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2012.

[10] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5923–5938, 2019.

[11] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[12] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2015.

[13] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *ACM International Conference on Multimedia*, 2019, pp. 2351–2359.

[14] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "RIRNet: Recurrent-in-recurrent network for video quality assessment," in *ACM International Conference on Multimedia*, 2020, pp. 834–842.

[15] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *International Journal of Computer Vision*, vol. 129, pp. 1238–1257, 2021.

[16] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *European Conference on Computer Vision*, 2018, pp. 803–818.

[17] L. Zhang, Z. Shi, J. T. Zhou, M.-M. Cheng, Y. Liu, J.-W. Bian, Z. Zeng, and C. Shen, "Ordered or orderless: A revisit for video based person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1460–1466, 2020.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[19] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[20] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (KoNViD-1k)," in *International Conference on Quality of Multimedia Experience*, 2017, pp. 1–6.

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

[22] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2061–2077, 2017.

[23] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

[24] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[25] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[26] C. Zhan, X. Duan, S. Xu, Z. Song, and M. Luo, "An improved moving object detection algorithm based on frame difference and edge detection," in *International Conference on Image and Graphics*, 2007, pp. 519–523.

[27] N. Singla, "Motion detection based on frame difference method," *International Journal of Information & Computation Technology*, vol. 4, no. 15, pp. 1559–1565, 2014.

[28] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-based Surveillance Systems*, 2002, pp. 135–144.

[29] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *International Conference on Pattern Recognition*, vol. 2, 2004, pp. 28–31.

[30] Z. Zivkovic and F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773–780, 2006.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[32] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*, 2006, pp. 428–441.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[35] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[36] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[37] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*, 2016, pp. 20–36.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[39] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[40] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[41] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[42] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4694–4702.

[43] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i Nieto, "RVOS: End-to-end recurrent network for video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5277–5286.

[44] T. Brandao and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1437–1447, 2010.

[45] Q. Wu, H. Li, F. Meng, and K. N. Ngan, "Toward a blind quality metric for temporally distorted streaming video," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 367–378, 2018.

[46] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks." in *ACM International Conference on Multimedia*, 2018, pp. 546–554.

[47] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, 2021.

[48] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3329–3342, 2016.

[49] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2244–2255, 2018.

[50] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *Multimedia Signal Processing*, 2019, pp. 1–5.

[51] F. Götz-Hahn, V. Hosu, and D. Saupe, "Critical analysis on the reproducibility of visual quality assessment using deep features," *arXiv preprint arXiv:2009.05369*, 2020.

[52] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations*, 2015.

[53] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient smt solver for verifying deep neural networks," in *International Conference on Computer Aided Verification*, 2017, pp. 97–117.

[54] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, "Safety verification of deep neural networks," in *International Conference on Computer Aided Verification*, 2017, pp. 3–29.

[55] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *European Conference on Computer Vision*, 2018, pp. 305–321.

[56] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *European Conference on Computer Vision*, 2012, pp. 340–353.

[57] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?." in *British Machine Vision Conference*, vol. 27, no. 2013, 2013, pp. 10–5244.

[58] Z. Wang and E. P. Simoncelli, "Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities," *Journal of Vision*, vol. 8, no. 12, pp. 8–8, 2008.

[59] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang, "Group maximum differentiation competition: Model comparison with few samples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 851–864, 2020.

[60] J. Yan, Y. Zhong, Y. Fang, Z. Wang, and K. Ma, "Exposing semantic segmentation failures via maximum discrepancy competition," *International Journal of Computer Vision*, vol. 129, p. 1768–1786, 2021.

[61] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[62] S. Bhardwaj, M. Srinivasan, and M. M. Khapra, "Efficient video classification using fewer frames," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 354–363.

[63] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[64] M. Zolfaghari, K. Singh, and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *European Conference on Computer Vision*, 2018, pp. 695–712.

[65] S. J. Miech A, Laptev I, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.069059*, 2017.

[66] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *IEEE International Conference on Computer Vision*, 2017, pp. 5842–5850.

[67] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[68] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[69] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[70] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Neural Information Processing Systems*, 2017, pp. 1–4.

[71] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

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

[74] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[75] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[76] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 154–166, 2016.

[77] Z. Duanmu, K. Ma, and Z. Wang, "Quality-of-experience for adaptive streaming videos: An expectation confirmation theory motivated approach," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 6135–6146, 2018.

[78] Z. Duanmu, A. Rehman, and Z. Wang, "A quality-of-experience database for adaptive video streaming," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 474–487, 2018.

Yuming Fang (M'13-SM'17) received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree from Nanyang Technological University, Singapore. He is currently a Professor with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include visual attention modeling, visual quality assessment, computer vision, and 3D image/video processing. He serves on the editorial board for IEEE Transactions on Multimedia and Signal Processing: Image Communication.

Zhaoqian Li eceived the Bachelor of Economics degree from the Jiangxi University of Finance and Economics, Nanchang, China, in 2018. and the M.A.Sc. degree with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China, in 2022. His research interests include visual quality assessment, and image/video processing.

Jiebin Yan received the Ph.D. degree from Jiangxi University of Finance and Economics, Nanchang, China. He was a computer vision engineer with MTlab, Meitu. Inc, and a research intern with MOKU Laboratory, Alibaba Group. From 2021 to 2022, he was a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is currently a Lecturer with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include visual quality assessment and computer vision.

Xiangjie Sui received the B.E. and M.A.Sc. degree from the Jiangxi University of Finance and Economics, Nanchang, China, in 2018 and 2021 respectively. He is currently pursuing the PhD degree. His research interests include visual quality assessment, and VR image/video processing.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology and the IEEE Signal Processing Letters.