IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 December 2022, accepted 12 March 2023, date of publication 20 March 2023, date of current version 23 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3259101

RESEARCH ARTICLE

Quality Feature Learning via Multi-Channel CNN and GRU for No-Reference Video Quality Assessment

NGAI-WING KWONG[1], YUI-LAM CHAN[1], (Member, IEEE), SIK-HO TSANG[2], AND DANIEL PAK-KONG LUN[1,2], (Senior Member, IEEE)

[1]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, SAR
[2]Centre for Advances in Reliability and Safety Limited (CAiRS), Hong Kong, SAR

Corresponding author: Yui-Lam Chan (enylchan@polyu.edu.hk)

This work was supported in part by the Hong Kong Research Grants Council under Research Grant PolyU 152069/18E, and in part by the Centre for Advances in Reliability and Safety (CAiRS) admitted under AIR@InnoHK Research Cluster.

ABSTRACT Nowadays, video quality assessment (VQA) plays a vital role in video-related industries to predict human perceived video quality to maintain the quality of service. Although many deep neural network-based VQA methods have been proposed, the robustness and performance are limited by small scale of available human-label data. Recently, some transfer learning-based methods and pre-trained models in other domains have been adopted in VQA to compensate for the lack of enormous training samples. However, they result in a domain gap between the source and target domains, which provides sub-optimal feature representation for VQA tasks and deteriorates the accuracy. Therefore, in the paper, we propose quality feature learning via a multi-channel convolutional neural network (CNN) with a gated recurrent unit (GRU), taking into account both the motion-aware information and human visual perception (HVP) characteristics to solve the above issue for no-reference VQA. First, inspired by self-supervised learning (SSL), the multi-channel CNN is pre-trained on the image quality assessment (IQA) domain without using human annotation labels. Then, semi-supervised learning is applied on top of the pre-trained multi-channel CNN to fine-tune the model to transfer the domain from IQA to VQA while considering motion-aware information for better frame-level quality feature representation. After that, several HVP features are extracted with frame-level quality feature representation as the input of the GRU model to obtain the final precise predicted video quality. Finally, the experimental results demonstrate the robustness and validity of our model, which is superior to the state-of-the-art approaches and is closely related to human perception.

INDEX TERMS Fine-tuning strategy, gated recurrent unit, human visual perception, motion-aware information, multi-channel convolutional neural network, no reference video quality assessment, self-supervised learning, semi-supervised learning.

I. INTRODUCTION

In the era of explosive information, video sharing has dramatic growth on social networks. As Cisco forecasts [1], by 2022, there will be approximately 400 exabytes of IP traffic per month, of which 82% of IP traffic will be video traffic. However, videos will inevitably be distorted after compression, processing, and transmission, thereby affecting the human visual experience (HVE) [2]. Consequently, to provide a better end-user experience, an accurate VQA approach is highly required to preserve the quality of service.

In considering the limited time and labor, although subjective VQA methods could estimate the most accurate perceived video quality, it is generally used to construct a benchmark video quality database only. In contrast, an objective VQA allows automatic video quality evaluation without enormous resources. Also, the ultimate goal of the objective VQA is to evaluate the perceptual quality highly

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

related to the subjective study. Therefore, it has recently become an attractive and challenging topic for researchers. There are three types of objective VQA methods according to their use of reference video [3]: Full-reference (FR) VQA [4], [5], [6], [7] requires complete information from the reference video; Reduced-reference (RR) VQA [8], [9] only takes part of the information from reference video; No-reference (NR) VQA [10], [11], [12], [13] does not require any information from the reference video. Since the reference video is not always available in real VQA applications, the NR-VQA approach is preferable to evaluate the video quality [14].

In the early stage, some traditional NR-VQA methods [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] have been developed by exploring different spatial and temporal features. For example, TLVQM [10] extracts 75 spatial and temporal features of frames and predicts the final video score using support vector regression (SVR). However, those hand-crafted features focus on the specific distortion only, limiting the performance and generalization of visual quality prediction.

Recently, many deep neural network (DNN) models have been proposed to learn the data representation, hidden features, and abstract features automatically. However, directly applying DNN in the VQA task faces two main challenges. The first challenge is that it requires high computational power and vast memory size. Usually, a raw video contains high spatial resolution and frame rate. Due to the limitation of graphics processing unit (GPU) memory size, it is hard to process a whole VQA database and train an end-to-end DNN model for the VQA task. Therefore, some existing NR-VQA methods use spatial resolution downscaling or/and temporally downsampling strategies to reduce the computational requirements and achieve end-to-end training. For example, the spatial resolution of the video in [22] is downsized to 448×448, and the video frame rate is reduced to one frame per second to achieve end-to-end model training. Also, RAPIQUE [23], HEKE [24], and RIRNet [25] also use temporally downsampling, which only processes one frame, around ten frames, and four frames per second, respectively, to reduce computational complexity. However, the effects of spatial resolution downsizing and temporally downsampling strategies lead to information loss that hinders the performance and accuracy of VQA models.

To relieve the above issue, most deep learning-based NR-VQA models separate the spatial and temporal learning process to avoid enormous high computational power at once and information loss. However, for the databases used in the subjective VQA study, each video only contains one mean opinion score (MOS) as ground truth to represent overall video quality. There is no human-annotated MOS label for each frame, i.e., the frame-level quality. While a video can contain thousands of frames or even more, it is impossible to annotate the frame-level quality for spatial feature learning, which is another challenge of DNN model training for the VQA task since it relies on large-scale data with robust labeling. Therefore, to ease the labeling burden of training a DNN from scratch, some pre-trained CNN models, such as ResNet [26] pre-trained on ImageNet [27], are used by the state-of-the-art NR-VQA methods, such as VSFA [12] and CNN-TLVQM [13]. These NR-VQA methods learn the features from the image classification task to the VQA target domain via transfer learning [28], [29]. However, the features learned from the image classification could only provide sub-optimal feature representation since the domain gap exists between the source image classification task and the target VQA domain.

Besides, some NR-VQA methods use a pre-trained DNN model in the IQA domain for the spatial feature learning process by assuming the spatial quality features of frames are close to the IQA domain. For example, the PHIQNet model in [30] is pre-trained on IQA to extract the perceptual quality features of video frames and then fed to the long short-term convolutional transformer (LSCT) model for temporal pooling. However, unlike still images, where continuous frames contain motion information that human visual attention is more attracted to regions with motion events than to the structural details or background [31]. Therefore, only using the IQA pre-trained model to extract frame perceptual quality and ignoring motion information could only provide the sub-optimal frame quality feature representation.

Inspired by SSL, this paper proposes a multi-channel CNN model using non-human annotated supervision signals for frame-level quality feature learning, with a GRU model to take HVP characteristics into account for NR-VQA. First, the multi-channel CNN with a channel attention mechanism is pre-trained on the IQA domain with the distorted images and their corresponding structure-aware maps and saliency maps for learning the image quality feature representation guided by non-human annotated supervision signals, which is motivated by SSL using pretext tasks [32], [33], [34], [35], [36], [37], [38], [39]. For example, RotNet [32] predicts the image rotation as a pretext task to learn the image representation prior to the fine-tuning for image classification. There are also other pretext tasks such as image or video colorization [33], [34], [35], jigsaw puzzle [36], relative position [37], pixel generation (iGPT) [38], and visual token reconstruction (BEiT) [39]. In addition, since human visual attention is attracted by the region with motion event more than the structural details, we perform the semi-supervised learning to fine-tune the pre-trained CNN to reduce the domain gap further. To incorporate the motion-aware information on the video frame, the unlabeled distorted frame and its corresponding structure-aware map and motion-aware map are fed into the pre-trained CNN to predict the pseudo label, which is treated as the label of the frame quality to solve the limitation of the lack of available human-annotated label data for video frames. Then, the data from IQA and the data from VQA are combined

and fine-tuned to transfer the feature learning from IQA to VQA domain. It achieves a better frame-level quality feature representation while considering motion-aware information on a video frame. Besides, some temporal and color-aware features, such as motion intensity, video smoothing, and color description in HSV color space, that are highly related to HVP [40], [41], are also extracted and incorporated with the frame-level quality feature representation as the input of the GRU model to obtain the final precise predicted video quality. The contributions of this work are summarized as follows:

- To compensate for the shortage of human-annotated labels on video frames used for the VQA task, we are the first to adopt self-supervised learning (SSL)-based NR-VQA framework based on non-human annotated supervision signals for the frame-level quality feature learning. All the details of this contribution will be presented in Section III-A.
- On the top of SSL-based NR-VQA framework, we contrive semi-supervised learning to fine-tune the pre-trained CNN, that will be described in Section III-A.3. Our objective is to reduce the domain gap by taking motion-aware information into consideration, thereby providing the optimized frame-level quality feature representation for the VQA task.
- We also extract some HVP-related features to assist the perceived video quality prediction. All features are then fed into the GRU model with pre-padding and masking strategies to comprehensively evaluate the perceived quality of the whole video. This contribution will be described in Section III-B.

By evaluating our model on three UGC VQA databases and two traditional distortion VQA databases, we verify that our model can provide better frame-level quality feature representation for various distortions and contents and can predict the video quality precisely close to HVP compared with other state-of-the-art transfer learning/pre-trained model-based VQA methods.

The rest of this paper is organized as follows. In Section II, we present the relevant research work. In Section III, the details of our proposed model are described. Then, the experimental results and related analysis are presented in Section IV. Finally, Section V concludes the paper.

II. RELATED WORK IN NR-VQA

A. TRADITIONAL METHODS

The general model of the NR-VQA method contains two key points: discriminative feature extraction and accurate quality prediction. For the spatial information as a vital feature of HVE, some successful and efficient IQA [42], [43], [44] methods were exploited to develop the spatial feature extraction algorithms in some NR-VQA approaches. For example, some NR-VQA methods [15], [16], [17], [18] were proposed that uses an NR-IQA method with the help of the natural scene statistics (NSS) model to estimate the quality of the frame based on the statistical properties of the spatial information, and then weight the distorted videos frame by frame using average pooling or regression. However, videos with 3D information are different from images. The characteristics of video contain not only spatial information but also temporal information. Therefore, several methods take the temporal features into account for their NR-VQA methods. Manasa and Channappayya [19] proposed an optical flow-based NR-VQA algorithm by measuring irregularities at the patch and frame levels. Video intrinsic integrity and distortion evaluation oracle (VIIDEO) [20] observed the intrinsic statistical regularities in natural videos and used it to quantify disturbances introduced by the distortions. Saad et al. [21] proposed a blind VQA method, V-BLIINDS, that assesses the frame quality using the spatiotemporal NSS model in the discrete cosine transform (DCT) domain and quantifies the motion coherency to predict the video quality.

B. DEEP LEARNING-BASED METHODS

It is well-known that neural networks can automatically learn the data representation, hidden features, and abstract features. CNN is a typical type of DNN that can extract discriminative, semantic, and comprehensive features of image/video. Therefore, many deep learning-based methods have been adopted on NR-VQA. For instance, in [45], 3D-DCT is used to represent the spatiotemporal features of video blocks and form the deformation of AC coefficients to capture the temporal features. The CNN model and the frequency histogram mapping function are then employed to explore the spatiotemporal regularities and obtain the final video quality score. SACONVA [46] uses 3D shearlet transform to extract the primary spatiotemporal features, which can also capture the NSS properties of video blocks. Afterward, the CNN and regression are applied to expand these features further and predict the video quality. DeepBVQA [47] uses a CNN model to extract various spatial features. The sharpness variation is then handcrafted as the temporal features. Lastly, features are aggregated and regressed to obtain the final quality score.

Moreover, Tran [48] proposed a 3D CNN model to extract spatiotemporal features to further address the problem in which a 2D CNN is unable to extract the temporal information directly in videos. You and Korhonen [49] proposed an NR-VQA model based on the 3D CNN and the long short-term memory (LSTM) [50] model to extract spatiotemporal features from video blocks and resolve the time series processing of video blocks. Wu et al. [51] also proposed an NR-VQA model based on the 3D CNN and LSTM model to construct the spatial attention map of video blocks and combined with the corresponding predicted similarity map to further extract the spatial quality information by applying the average pooling and standard deviation pooling. These features are then fed into the LSTM model to predict the overall video quality. Besides, Yi et al. [22] proposed an end-to-end training model for the VQA task. First, the VGG16 model is used to extract the spatial features, while an attention module is added to calculate the dependency between local spatial features. Then, the GPU

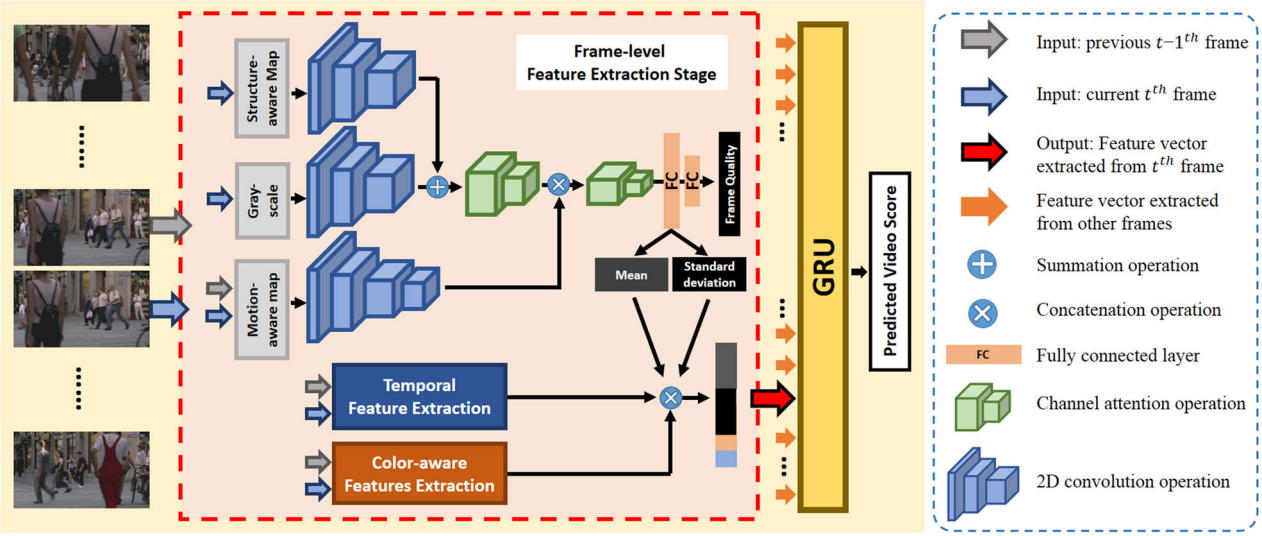


FIGURE 1. The framework of our proposed NR-VQA model.

and memory function are used to obtain the final video quality score.

C. TRANSFER LEARNING AND PRE-TRAINING BASED METHODS

To compensate for the lack of enormous training samples to train the robust deep CNN model, some NR-VQA methods learn the features from other domains and then transfer them to the VQA target domain via transfer learning. VSFA [12] extracts content-aware features from a CNN model pre-trained on an image classification task and then predicts the video quality using a GRU temporal-memory model. The authors also improved this method by training on mixed datasets in [52]. CNN-TLVQM [13] combines the handcrafted human vision system (HVS) features extracted from TLVQM [10] and the spatial features obtained from a pre-trained CNN via transfer learning, and then uses an SVR model to evaluate the predicted quality score. Chu et al. [53] also uses a CNN pre-trained on an image classification task to extract spatial features, and horizontal and vertical spatiotemporal slice features of frames. These features are then learned by multi-layer perceptron (MLP) to predict the frame-level quality, and SVR is adopted to fuse the scores of MLP into a final score. LSCT-PHIQNet [30] pre-trains the PHIQNet on an IQA task and then uses the features extracted from it with an LSCT model as a temporal regression model to predict the final video quality. PVQ [54] uses a pre-trained IQA model to extract spatial features of frames and a pre-trained model on a video classification task to extract spatiotemporal features of a 3D clip. Then, the final video quality can be predicted after the spatiotemporal pooling and time series regression with an inception time model. HEKE [24] creates a large-scale video dataset with weak labels to pre-train a feature encoder to extract the spatiotemporal representation of video and then uses the pre-trained feature encoder and hierarchical features regression to predict the video quality. RIRNet [25] extracts spatial quality features from a pre-trained model on an image classification task and then predicts the video quality by the motion effect modeling. However, since the features learned from other tasks are not closely related to the VQA target domain, we believe that there is still room for improvement in transfer learning/pre-trained model-based NR-VQA approaches by reducing the features gap between the source domain and VQA domain.

III. PROPOSED METHOD

In this section, we introduce a novel NR-VQA method that adopts a new multi-channel CNN model with GRU, incorporating motion-aware information and HVP characteristics. The framework of our proposed model is shown in Fig. 1. First, the multi-channel CNN is pre-trained on the IQA database to predict the image quality feature focusing on structure-aware features and saliency region, which can be regarded as a sort of SSL-based method using a pretext task. Then, with the semi-supervised learning and fine-tuning strategies, the features learned from the pre-trained CNN is fine-tuned to predict the frame-level quality feature representation focusing on structure-aware features and motion-aware regions to transfer the feature learning from IQA to VQA domain to reduce the domain gap for better feature representation. In the meantime, HVP-related temporal and color-aware features are also extracted. Lastly, all features are fed into the GRU model to explore spatiotemporal features and the gradient of temporal features to comprehensively evaluate the quality of the whole video. We will detail each part in the following sections.

A. SSL-BASED MULTI-CHANNEL CNN MODEL FOR VQA

The HVS is known to be sensitive to moving objects [31]. Hence, visual attention should be more attracted by the

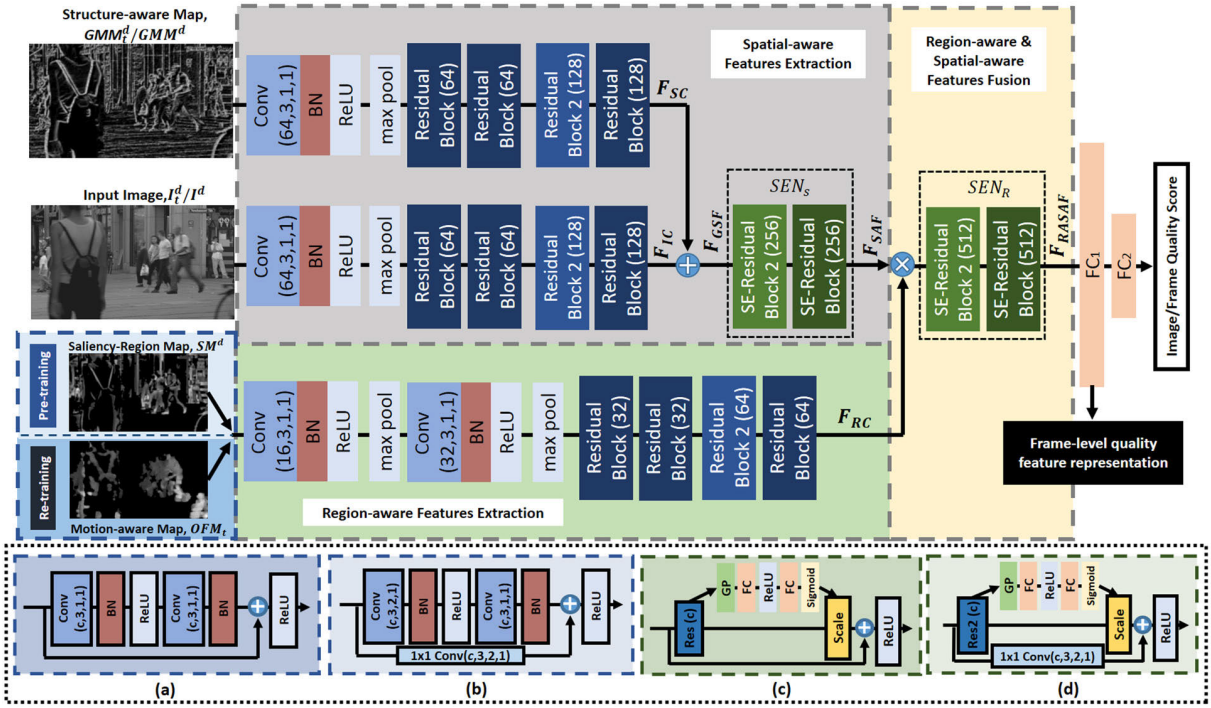


FIGURE 2. The network architecture of our proposed multi-channel CNN model. (a) The structure of Residual Block (ch); (b) The structure of Residual Block 2 (ch); (c) The structure of SE-Residual Block (ch); (d) The structure of SE-Residual Block 2 (ch). Conv(ch, kn, st, pd) represents the 2D convolution operation where ch is the output channel, kn × kn is the kernel size, st represent the size of stride and pd is the padding size. BN, FC and GP represent the batch normalization operation, fully connected layer and global pooling. I^d, $\mathbf{GMM}^d$, $\mathbf{SM}^d$ is the data from IQA dataset and I_t^d, $\mathbf{GMM}_t^d$, $\mathbf{OFM}_t$ is the data from VQA dataset.

motion event regions rather than the structural details of the video. Therefore, the distortion occurring in moving objects should affect the human perceptual quality more than those occurring in the background or spatial structures. However, most of the existing transfer learning/pre-trained CNN models in NR-VQA are used to extract the spatial features or content-aware features of the whole frame to represent the frame quality without considering motion information and the motion-aware region, which extends the domain gap between the source domain and target VQA domain and could only provide the sub-optimal feature representation for the VQA task. Besides, human-annotated labels for frame quality are not available in the VQA databases and the subjective quality scores of videos cannot represent the frame quality due to varying distortions over time and frames. In other words, there is no human annotated MOS label to represent the frame-level quality with motion. To address this issue, SSL, which is a form of unsupervised learning that can let network learn the critical feature from unlabeled data by providing the non-human annotated supervision signal, introduced in the proposed VQA framework is to guide and pre-train the multi-channel CNN model in both IQA and VQA databases to learn the frame-level quality feature representation by the non-human annotated supervision signal.

Based on the concept of the region of interest (ROI), we hypothesize that SM can guide the image quality prediction by focusing on the vital stillness region, while the motion-aware region map can guide the frame quality prediction by focusing on the motion-aware region. Therefore, we use the concept of semi-supervised learning on top of our SSL-based multi-channel CNN model and combine the data from IQA and the data from VQA to fine-tune our multi-channel CNN model, as shown in Fig. 2, to process the distorted frame, the structure-aware map, and the motion-aware region map to estimate the optimized frame-level quality feature representation by considering both spatial and motion-aware information at the frame level.

1) PRE-PROCESSING STAGE

Before the training process of the multi-channel CNN model, we first compute the gradient magnitude map (GMM) as the structure-aware map, because the GMM of the image is responsive to image distortions, such as compression, blur, and noise, and can effectively capture image local structures, to which the HVS is highly sensitive. Therefore, the GMM can reflect the structural information of images proved in a series of literature on image processing [55]. As shown in Fig. 3(b), the GMM can show the rich structural information of Fig. 3(a). To obtain the GMM of the input distorted image I^d, we convolve I^d using the Prewitt filters along the horizontal and vertical directions to compute the image directional gradients. The GMM of I^d, GMM^d, is then constructed by estimating the root mean square of horizontal

FIGURE 3. Results from the pre-processing stage. (a) Original frame; (b) Gradient magnitude Map; (c) Saliency Map; (d) Optical flow Map.

and vertical directional gradients as follows:

$$\mathrm{GMM}^d = \sqrt{(I^d * g_h)^2 + (I^d * g_v)^2} \tag{1}$$

where symbol $*$ denotes the convolution operation, and g_h and g_v are the Prewitt filters along horizontal and vertical directions defined by

$$g_h = \begin{bmatrix} 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \end{bmatrix}, g_v = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 \\ -1/3 & -1/3 & -1/3 \end{bmatrix} \tag{2}$$

Moreover, to compute the SM of the image, we first implement the method in [56] to determine the saliency residuals on the spectrum domain. This is because the log-spectrum can be sensitive to NSS that indicates the salient region of the image. The saliency residual $\mathcal{R}(f)$ defined as:

$$\mathcal{R}(f) = \mathcal{L}(f) - \mathcal{A}(f) \tag{3}$$

where $\mathcal{A}(f)$ is the real part of the image f after Fourier transform, which represent the shape information of the image

in the spectral domain and $\mathcal{L}(f)$ is the log spectrum of $\mathcal{A}(f)$. Then, the preliminary saliency map (PSM) is computed by inverting the saliency residuals from the spectral domain back to the spatial domain.

$$\mathrm{PSM}(f) = \mathcal{F}^{-1}(\mathcal{R}(f)) \tag{4}$$

where $\mathcal{F}^{-1}(\cdot)$ indicates the inversion process of saliency residuals from the spectral domain to the spatial domain. We also apply the visual saliency feature (VSF) method in [57] to calculate the center-surround differences on the salient region in the PSM. It further extracts the fine-grained features and defines borders for the PSM to compute our final SM, as shown in Fig. 3(c), as follows:

$$\mathrm{SM}^d = \mathrm{VSF}\left((\mathrm{PSM}\left(I^d\right)\right) \tag{5}$$

For the motion-aware map, according to HSV, human perception is caught by moving objects. Since the optical flow map (OFM) can determine the inter-frame motion variation, it is treated as the motion-aware map to represent the motion-aware region. In this paper, we applied the algorithm in [58] to compute the OFM, as shown in Fig. 3(d). First, two neighboring frames are transformed into polynomial expansion. Then, the polynomial expansion coefficients can be used to estimate the displacement field. Therefore, we use this algorithm as the motion estimation method to compute the OFM of two inter-frames as follows:

$$\mathrm{OFM}_t = \mathrm{ME}\left(\mathrm{PET}\left(I^d_{t-1}\right), \mathrm{PET}\left(I^d_t\right)\right) \tag{6}$$

where I^d_t and I^d_{t-1} represents the current t^{th} frame and t-1^{th} frame, respectively, PET $(\cdot)$ is the polynomial expansion transform and ME $(\cdot)$ is the motion estimation mechanism.

After the pre-processing state, the GMM and SM of images are computed for the IQA dataset, and the GMM and OFM of frames are extracted for the VQA dataset for the following training process.

2) SELF-SUPERVISED LEARNING-BASED CNN MODEL PRE-TRANING

To train our multi-channel CNN model, inspired by [59] and [60] using distortion intensity as the self-supervised signal for the regression task in SSL, we use Gradient Magnitude Similarity Deviation (GMSD) [55] as the non-human annotated supervision signal, or so-called pseudo label (PL), PL^d_I to label the unlabeled data from IQA $U^d_I = \{I^d, \mathrm{GMM}^d, \mathrm{SM}^d$ first, where U^d_I includes the distorted image, and the corresponding GMM and SM. It is noted that GMSD has been proven to be effective in representing the distortion intensity of an image and image quality by comparing the similarity between its reference version and distorted version [55]. Therefore, GMSD as the non-human annotated supervision signal, $\mathrm{PL}^d_I = \mathrm{GMSD}$, can be used to guide the model for effective learning of valuable information and image quality. This pre-training method is inspired by SSL using pretext tasks. By learning GMSD as a pretext task, our multi-channel CNN model can learn the image quality feature representation.

To compute the structure-based quality features, we extract the global spatial features by the summation of the features extracted from the channel of GMM^d and the channel of I^d as shown in Fig. 2. Features extracted from channel I^d represent the basic image quality features of the distorted image. Additionally, features extracted from channel GMM^d are the structure-aware quality features highlighted in the structural information of the image. By summing both features, we can capture the global spatial features, F_{GSF}, of the image accentuated by the structure-aware features, which is given by:

$$F_{\mathrm{GSF}} = F_{\mathrm{IC}}(I^d) \oplus F_{\mathrm{SC}}(\mathrm{GMM}^d) \tag{7}$$

where symbol $\oplus$ is the element-wise summation operation, $F_{\mathrm{IC}}(\cdot)$ and $F_{\mathrm{SC}}(\cdot)$ represent the feature extraction processes on I^d and GMM^d, respectively.

Also, we incorporate the squeeze-and-excitation block [61] with residual block, which can squeeze the features to be one

dimensional data as global information. It can then reinforce the critical features and weaken the inconsequence features by the channel-wise multiplication. Therefore, by performing the channel attention mechanism on the global spatial features, F_{GSF}, we can extract the high-level spatial quality features, which are highly related to image quality, named as spatial-aware features F_{SAF} and defined as:

$$F_{\mathrm{SAF}} = \mathrm{SEN}_S\left(F_{\mathrm{GSF}}\right) \tag{8}$$

where $\mathrm{SEN}_S(\cdot)$ is the channel attention mechanism for F_{GSF}.

In the meantime, features extracted from the SM channel are concatenated with the spatial-aware features F_{SAF}. Thus, features extracted from channel SM^d can be used as the region-aware based side information to guide the spatial-aware features for better image quality prediction by focusing on the vital region, which is sensitive to human visual attention. Therefore, via the second channel attention mechanism, spatial-aware features can be weighted by the vital region of the image with the guidance of the region-aware features. The fusion of the region-aware features and spatial-aware features, F_{RASAF}, is defined as:

$$F_{\mathrm{RASAF}} = \mathrm{SEN}_R\left(F_{\mathrm{SAF}} \otimes F_{\mathrm{RC}}\left(\mathrm{SM}^d\right)\right) \tag{9}$$

where symbol $\otimes$ is the concatenation operation, $F_{\mathrm{RC}}(\cdot)$ represents the region-aware features extraction process on SM channel and $\mathrm{SEN}_R(\cdot)$ is the channel attention mechanism for region attention fusion. Finally, two fully connected layer are appended to F_{RASAF} to predict the image quality given by:

$$\hat{p}^L = \mathrm{MultiCNN}_{\mathrm{IQA}}(U_I^d) = \mathrm{FC}_2\left(\mathrm{FC}_1\left(F_{\mathrm{RASAF}}\right)\right) \tag{10}$$

where $\hat{p}^L$ denotes the predicted GMSD of the multi-channel CNN model, $\mathrm{MultiCNN}_{\mathrm{IQA}}(\cdot)$. The loss function of our $\mathrm{MultiCNN}_{\mathrm{IQA}}(\cdot)$ pre-trained in IQA data is defined as:

$$\mathcal{L}_{self}(\hat{p}^L, PL_I^d) = \frac{1}{M}\sum_{i=0}^{M-1}\left(\hat{p}_i^L - PL_{I,i}^d\right)^2 \tag{11}$$

where M is the batch size, and PL_I^d represent the corresponding supervision signal, GMSD, of the IQA data, U_I^d. With this pretraining, our multi-channel CNN model is able to learn image quality feature representation first. Then we will fine-tune the model with video frames for frame quality feature learning.

3) SEMI-SUPERVISED LEARNING FOR FINE-TUNING

As mentioned before, frame quality feature extraction without considering motion information and the motion-aware region generates the domain gap between the source IQA and target VQA tasks. To transfer the feature representation of our multi-channel CNN models from IQA to VQA domain, we incorporate the motion-aware information into our model to compute the frame-level quality feature representation by using the semi-supervised learning and fine-tuning strategies from our pre-trained multi-channel CNN model, $\mathrm{MultiCNN}_{\mathrm{IQA}}(\cdot)$ in Section III-A.2.

With the same concept of ROI, we assume that the OFM can be used to guide the video frame quality prediction by focusing on the motion-aware region as same as the SM guide the image quality prediction by focusing on the stillness salient structure region. Therefore, we replace SM with OFM as the region-aware map for VQA data $U_V^d = \{I_t^d, \mathrm{GMM}_t^d, \mathrm{OFM}_t$ to fine-tune the model, where U_V^d includes the distorted t^{th} frame, and the corresponding GMM and OFM. By doing so, after extracting the spatial-aware features of frames F_{SAF} using I_t^d and GMM_t^d via (7) and (8), the spatial-aware features of frames F_{SAF} will be weighted by the motion-aware region OFM_t instead of SM in (9), which can optimize the frame quality feature representation extraction while considering the motion-aware region and motion information.

Besides, as aforementioned, there is no human-annotated label for each video frame. Therefore, for each video frame of the training VQA dataset, $\mathrm{MultiCNN}_{\mathrm{IQA}}(\cdot)$ is initially used to generate the pseudo labels PL_V^d of data from VQA, U_V^d. During the fine-tuning process, the re-trained multi-channel CNN model is then used to predict the new PL_V^d of data from U_V^d for the next training process, similar to the semi-supervised image classification in [62]. Thus, assuming $\mathrm{MultiCNN}_{\mathrm{trans}}(\cdot)$ is the multi-channel CNN model that is being transferred from IQA to VQA, the PL_V^d of data from U_V^d is then generated as:

$$\mathrm{PL}_V^d = \mathrm{MultiCNN}_{\mathrm{trans}}(U_V^d) \tag{12}$$

After preparing the VQA dataset $\{U_V^d, \mathrm{PL}_V^d\}$, it will be combined with the IQA dataset $\{U_I^d, \mathrm{PL}_I^d\}$ for semi-supervised learning to fine-tune the multi-channel CNN model for domain adaptation, using the same features learning process, (7)-(9), as shown in algorithm 1. Hence, the loss function of the entire training process of semi-supervised learning including both IQA dataset, $\{U_I^d, \mathrm{PL}_I^d\}$, and VQA dataset, $\{U_V^d, \mathrm{PL}_V^d\}$, is defined as:

$$\mathcal{L}_{\mathrm{semi}} = \mathcal{L}_{\mathrm{self}}(\hat{p}^L, \mathrm{PL}_I^d) + a(k)\mathcal{L}_{\mathrm{self}}(\hat{u}^L, \mathrm{PL}_V^d) \tag{13}$$

where

$$a(k) = \begin{cases} 0 & k < K_1 \\ \left[(k-K_1)/(K_2-K_1)\right] a_f & K_1 \le k < K_2 \\ a_f & k \ge K_2 \end{cases} \tag{14}$$

$\hat{u}^L$ denote the predicted result of VQA data, U_V^d, k is the current epoch, and K_1, K_2, and a_f are the parameters for tuning $a(k)$ at different epochs.

As we can see in (13) and (14), only $\mathcal{L}_{\mathrm{self}}(\hat{p}^L, \mathrm{PL}_I^d)$ is performed when $k < K_1$ since the network is at the pre-training stage mentioned in Section III-A.2. When $k \ge K_1$, $a(k)$ is progressively increased by the epoch, which also increase the influence of VQA dataset $\{U_V^d, \mathrm{PL}_V^d\}$ in the training process. Thus, the model is gradually fine-tuned with VQA data, U_V^d, so that the domain gap between IQA and target VQA domain is reduced with the consideration of motion-aware information. Finally, after

Algorithm 1 Training Process of Semi-Supervised Learning Performed on Our SSL-Based Multi-Channel CNN Mode

Input: IQA data $U_I^d = \{I^d, \text{GMM}^d, \text{SM}^d$, and VQA data $U_V^d = \{I_t^d, \text{GMM}_t^d, \text{OFM}_t$

Output: A well-trained multi-channel CNN model, $\text{MultiCNN}_{\text{VQA}}(\cdot)$

Let k be the current epoch, K_1 is the number of epochs to start the fine-tuning process, and the total number of training epoch is 1000;
Compute the pseudo label of IQA data $\text{PL}_I^d = \text{GMSD}$;
while $k < 1000$ **do**
  Obtain global spatial features F_{GSF} of IQA data (U_I^d) using (7);
  Obtain spatial-aware features F_{SAF} of U_I^d using (8);
  Obtain region-aware and spatial-aware fusion features F_{RASAF} of U_I^d using (9);
  Predict the image quality $\hat{p}^L = \text{MultiCNN}_k(U_I^d)$;
  if $k \geq K_1$ **then**
    /* *insert VQA data and combine with IQA data to fine-tune the model* */
    Compute the pseudo label of VQA data U_V^d;
    Obtain F_{GSF} of VQA data U_V^d using (7);
    Obtain F_{SAF} of VQA data U_V^d using (8);
    Obtain F_{RASAF} of VQA data U_V^d using (9);
    Predict the frame quality $\hat{u}^L = \text{MultiCNN}_k(U_V^d)$;
  end
  Update $a\,(k)$ using (14);
  Update the model $\text{MultiCNN}_k(\cdot)$ using the loss function $\mathcal{L}_{\text{semi}}$ in (13);
end

Return: A well-trained multi-channel CNN model, $\text{MultiCNN}_{\text{VQA}}(\cdot)$, is contained to use for the frame quality feature representation extraction in (15) for VQA tasks

completing the training process at the last epoch, the features learned from the IQA domain is transferred to our target VQA domain. The well-trained multi-channel CNN model, named as $\text{MultiCNN}_{\text{VQA}}(\cdot)$, is obtained for extracting the optimized frame-level quality feature representation for the VQA task. The summary of the training process of semi-supervised learning performed on our SSL-based multi-channel CNN model, including the content in Section III-A.2 and Section III-A.3, is shown in algorithm 1.

4) FRAME-LEVEL QUALITY FEATURES EXTRACTION

Specifically, the frame-level quality features, FQFR, are extracted at the output of the first fully connected layer in $\text{MultiCNN}_{\text{VQA}}(\cdot)$, as shown in Fig. 2. In practice, we divide the frame into B non-overlapping frame-blocks and each frame block goes through $\text{MultiCNN}_{\text{VQA}}(\cdot)$ to obtain FQFR_b. At the end, we take the mean and standard deviation of all FQFR_b within the frame which as shown in Fig. 1:

$$\text{FQFR}_t = \left\{\mu\{\text{FQFR}_b\}_{b=1}^{b=B}, \sigma\{\text{FQFR}_b\}_{b=1}^{b=B}\right\} \tag{15}$$

where FQFP_b is the quality feature representation of a frame-block in t^{th} frame, B is the total number of frame-blocks in t^{th} frame, and $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation operation, respectively. With our proposed pre-training and fine-tuning strategies, the extracted frame quality feature representation FQFR_t (512 dimensions) includes both structure-aware features and motion-aware features for VQA task.

B. HVP-RELATED FEATURES EXTRACTION

Although the proposed multi-channel CNN considers the motion-aware information of the inter-frame to extract the optimized frame-level quality feature representation, some motion events, such as sudden screen changes, frame freeze, and new object appearances, and the large variation of color could also have a significant impact in HVP [40], [41]. The works in [10], [13], and [23] considered human perception and HVS features, we also adopt the HVP characteristic, which reflects the comprehensive perceived quality following the HVS, by extracting additional temporal features and color-aware features of frames and incorporating them with the frame-level quality features into the GRU model to assist in predicting precisely video quality that is close to human perception.

For the temporal features, as mentioned, optical flow is used to determine the motion variation of the inter-frame that reflects the temporal attention. Therefore, based on the OFM in (6), we further calculate the global motion intensity (GMI), size of the motion event region (MER), and the mean (MV_μ) and standard deviation (MV_σ) value within the motion region of t^{th} frame to reflect the level of motion variation as follows:

$$\begin{aligned}
\text{GMI}_t &= \frac{1}{W \times H}\sum\nolimits_{x=0}^{W-1}\sum_{y=0}^{H-1}\text{OFM}_t(x,y)\\
\text{MER}_t &= \frac{1}{W \times H}\sum\nolimits_{x=0}^{W-1}\sum\nolimits_{y=0}^{H-1}\text{MR}_t(x,y)\\
\text{MV}_\mu^t &= \mu(\{\text{OFM}_t(x,y)\,|\,\text{OFM}_t(x,y) > 0\})\\
\text{MV}_\sigma^t &= \sigma(\{\text{OFM}_t(x,y)\,|\,\text{OFM}_t(x,y) > 0\})
\end{aligned} \tag{16}$$

with

$$\text{MR}_t(x,y) = \begin{cases} 1 & \text{OFM}_t(x,y) > 0 \\ 0 & \text{OFM}_t(x,y) = 0 \end{cases} \tag{17}$$

where W and H are the width and height of the OFM, $\text{OFM}_t(x,y)$ is the pixel value of OFM at the location (x,y), $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean operation and the standard deviation operation, respectively.

Also, since human eyes are sensitive to sudden scene changes, frame freeze, and new object appearance in videos, we also use the Structural Similarity Index Measure (SSIM) [63] to measure the structural similarity between

frames to indicate their structural difference and represent the video smoothing (VS) by:

$$\mathrm{VS}_t = \mathrm{SSIM}(I_{t-1}^d, I_t^d) \tag{18}$$

To extract the color-aware features, we first convert the frame into HSV color space since it is a color description model that is more consistent with human perception. In addition to the human attention that may be drawn to color variation, spatial distortions may also be reflected in the color domain. Thus, we compute the standard deviation of the frame and the mean-square error (MSE) between frames in hue (H) and saturation (S) to represent the color spatial distortion (CSD) and the level of color variation (CV) as follows:

$$\begin{aligned} \mathrm{CSD}_t &= \{\sigma\left(H\left(I_t^d\right)\right), \sigma\left(S\left(I_t^d\right)\right) \\ \mathrm{CV}_t &= \left\{\mathrm{MSE}\left(H\left(I_{t-1}^d\right), H\left(I_t^d\right)\right), \right. \\ &\quad \left. \mathrm{MSE}\left(S\left(I_{t-1}^d\right), S\left(I_t^d\right)\right)\right\} \end{aligned} \tag{19}$$

where $H\left(I_t^d\right)$ and $S\left(I_t^d\right)$ represent the t^{th} frame in H and S color spaces, respectively, and MSE $(\cdot)$ is the MSE operation. Consequently, we can use the above additional temporal and color-aware features of frames, which are highly related to HVS and consistent with human perception, to assist with FQFR_t for video quality prediction.

C. VIDEO QUALITY PREDICTION VIA GRU MODEL

A GRU model is a well-known recurrent neural network that has a recurrent nature to process input sequences in an iterative way. The recurrent nature means that the node output at the current timestamp acts as feedback and inputs into the node at the next timestamp. This makes GRU extract the temporal feature of data efficiently. Consequently, GRU can make predictions based on time series data and can explore the spatiotemporal regularities of distorted videos for our VQA task. Therefore, we take advantage of the GRU model for VQA to learn a temporal variation of frame-level quality feature representation and additional HVP-related features along with time series to represent the spatiotemporal features of the video. The GRU model can reveal the gradient of temporal features by analyzing the entire temporal data sequences, which can comprehensively reflect the whole video quality.

First, we concatenate the frame-level quality feature representation, and temporal and color-aware features as a feature vector $\mathbf{fv}_t = \{\mathrm{FQFR}_t, \mathrm{GMI}_t, \mathrm{MER}_t, \mathrm{MV}_\mu, \mathrm{MV}_\sigma, \mathrm{VS}_t, \mathrm{CSD}_t, \mathrm{CV}_t\}$ with 521 dimensions. For the c^{th} distorted video, a features vector, $\mathbf{FV_c} = [\mathbf{fv_1}, \mathbf{fv_2}, \ldots, \mathbf{fv_{T_c-1}}, \mathbf{fv_{T_c}}]$ is then generated, where $c = 1, 2, 3, \ldots, C$, C is the total number of videos in the VQA database, $\mathbf{fv_t}$ is the feature vector of t^{th} frame, and T_c is the total number of frames of the c^{th} distorted video. After that, we built a GRU model to process the input data, $\mathbf{FV_c}$, sequentially to explore spatiotemporal features and the gradient of temporal features to comprehensively evaluate the quality of the whole video. We also perform the pre-padding and masking strategy on $\mathbf{FV_c}$ to improve the performance since the memory function of GRU can reduce the influence of padding data placed in front of the actual data and it benefits the gradient descent and let the GRU model focus more on meaningful data when the meaningful data are placed at the back. The loss function of the supervised learning GRU model is defined as:

$$\mathcal{L}_{\mathrm{GRU}} = \frac{1}{N}\sum_{i=0}^{N-1}(\hat{v}_i^L - v_i^L)^2 \tag{20}$$

where N is the batch size, $\hat{v}^L$ represents the final predicted video quality score by the GRU model for the distorted video and v^L is the ground truth label of the corresponding distorted video collected from the subjective study.

Hence, during inference, we first extract the features in (15), (16)-(18) and (19) from all frames within a video to obtain $\mathbf{FV_c}$. Then, $\mathbf{FV_c}$ is input into GRU to get the final predicted video quality score $\hat{v}^L$.

TABLE 1. The summary of the video databases.

Database	Resolution	Number of Videos	Frame Rate (fps)	Duration (seconds)
KoNViD-1k	960×540	1200	24, 25, 30	8
LIVE-Qualcomm	1920×1080	208	30	15
LIVE-VQC	240P-1080P	585	19-30 (one 120)	10
LIVE	768×432	150	25, 50	8.68–10
CSIQ	832×480	216	24, 25, 30, 50, 60	10

IV. EXPERIMENTAL RESULTS

A. VIDEO QUALITY DATABASES AND EVALUATIONS

To demonstrate the validity and the robustness of our proposed model, three UGC VQA databases (KoNViD-1k [64], LIVE-Qualcomm [65], and LIVE-VQC [66]) and two traditional distortion VQA databases (LIVE [2] and CSIQ [67]) were tested on model. The summary of the above video databases is shown in Table 1.

1) *KoNViD-1k* [64] is an extensive database that contains 1200 real-world video sequences with frame rates of 24, 25, and 30 fps. The large number of video sequences in KoNViD-1k represents a wide variety of content and covers almost all kinds of distortions. The MOS ranges are from 1.22 to 4.64.

2) *LIVE-Qualcomm* [65] contains 208 distorted videos. These videos are with six common in-capture distortions: artifacts (noise and blocking effect), color, exposure, focus, blurriness, and camera shaking. All videos have a duration of 15 seconds with a frame rate of 30 fps and with the MOS ranging from 16.56 to 73.64.

3) *LIVE-VQC* [66] contains 585 distorted videos, with the MOS ranging from 6.22 to 94.29. All videos have a duration of 10s with frame rates of 19-30 fps (one is 120fps). These videos contain 18 types of resolutions from 240P to 1080P, unique contents, and different combinations of distortions.

4) *LIVE* [2] includes 150 distorted videos. These videos are generated from each reference video with four different distortion types: wireless distortions, IP distortions, H.264 compression, and MPEG-2 compression. Also, each distorted video is an 8.68-10-second video with a frame rate of 25 fps or 50 fps. Besides, the average differential MOS is provided in the range of 30.94 to 81.16.

5) *CSIQ* [67] has 216 distorted videos. All videos have a duration of 10 seconds and span various frame rates: 24 to 60 fps. The 18 distorted videos produced by each reference video have three different levels, and each type has six distortions: Motion JPEG compression, H.264 compression, HEVC compression, wavelet compression using SNOW codec [68], packet-loss, and additive white Gaussian noise (AWGN). Besides, the average differential MOS is provided in the range of 14.48 to 82.80.

To evaluate the performance of our proposed model, we used the Pearson Linear Correlation Coefficient (PLCC) and root MSE (RMSE) to measure the accuracy between the objective prediction and subjective assessment, and the Spearman Rank Order Correlation Coefficient (SROCC) to measure the monotonic consistency between the objective prediction and subjective assessment. The closer the value of the correlation coefficient is to 1, the higher the performance of the VQA model. Also, a nonlinear regression process is performed to map the prediction result to the subjective scores with different value domains according to the video quality experts group (VQEG) [69] as follows:

$$\hat{Q} = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + \exp[\beta_2 (Q - \beta_3)]} \right] + \beta_4 Q + \beta_5 \quad (21)$$

where $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the parameters to be determined.

B. IMPLEMENTATION DETAILS

In our experiments, each video database was divided into five non-overlapping datasets. Four sets of the distorted videos were selected for training and validation, while the remaining set was used for testing. Then, five-fold cross-validation was conducted in our experiments. Also, of the distorted videos for training and validation of each cross-validation, 80% were used for training, and 20% were used for validating ablation performance. For the multi-channel CNN model training, the CSIQ IQA database [70] was used to pre-train the multi-channel CNN model using (11) since the CSIQ IQA database is a large-scale dataset containing many distorted images with common, general, and diverse distortion, which is suitable as the baseline for image quality feature representation learning. All images were split into 128 × 128 image blocks and labeled by GMSD as non-human annotated supervision signals. During the fine-tuning process, we randomly selected the video frames in the training set of the video database and split them into 128 × 128 frame-blocks. And the sample size of the VQA data U_V^d is made the same as the IQA data U_I^d to train the multi-channel CNN model. The parameters a_f, K_1, and K_2 in (14) were set as 3, 100, and 700, respectively, through the experiments. We trained the model for 1000 epochs with an initial learning rate of 0.0001 using (13) as the loss function and an Adam optimizer.

The mean and standard deviation of frame-level quality feature representation are then extracted using (15) and concatenated with temporal features and color-aware features in (16)-(18) and (19) to form the feature vector, $\mathbf{FV}_c$, as the input of the GRU model. For the training process of the GRU model, specifically, we built a GRU model with three layers and 75 cell units. We set the maximum length of the video in each database (Note that the video of 120 fps in LIVE-VQC was treated as an isolated case and was not used in our experiment) as the length of $\mathbf{FV_c}$ with pre-padding data. Similarly, we used (20) as a loss function and an Adam optimizer to train the model for 500 epochs with an initial learning rate of 0.0001.

C. PERFORMANCE EVALUATION ON UGC VQA DATABASES

To comprehensively evaluate the efficiency of our proposed method, we trained and tested our model on the three UGC databases individually and compared the performance with other state-of-the-art NR-VQA approaches. Twelve NR-VQA methods, BRISQUE [17], NIQE [15], V-BLIINDS [21], VIIDEO [20], TLVQM [10], VSFA [12], CNN-TLVQM [13], HEKE [24], RAPIQUE [23], VIDEVAL [71], MDTVSFA [52], and LSCT-PHIQNet [30] were included. In particular, VSFA, CNN-TLVQM, RAPIQUE, and MDTVSFA use the CNN model pre-trained on ImageNet classification task to extract the content-aware features via transfer learning. Also, LSCT-PHIQNet uses the pre-trained model on the IQA task to extract quality-aware features. The mean and standard deviation performances of PLCC, SROCC, and RMSE results for the mentioned competitors and the proposed model on the KoNViD-1k, LIVE-Qualcomm, and LIVE-VQC video databases are given in Table 2. Following the method in [12], Table 2 also includes the weighted average to weigh the results according to the number of videos to represent the overall performance. In Table 2, the best and the second-best performances of PLCC, SROCC, and RMSE are highlighted in bold and underlined, respectively.

As shown in Table 2, our proposed model achieves the best and second-best performance in terms of PLCC, SROCC, and RMSE in these three UGC databases and has small standard deviation values, representing that our model is more robust. Although our performance is the second-best performance in terms of PLCC in the KoNViD-1k database that is on par with LSCT-PHIQNet with a 0.003 slight difference, our proposed model outperforms LSCT-PHIQNet in LIVE-Qualcomm and LIVE-VQC databases with about 0.02 to 0.03 improvement in terms of PLCC and SROCC. Furthermore, in the LIVE-Qualcomm database, the correlation and accuracy (PLCC, SROCC, and RMSE) of our proposed model are superior to other NR-VQA methods. Ours also achieves the best performance

TABLE 2. Performance comparison of NR-VQA models on the three UGC databases. the boldfaced and underlined entries indicate the best and the second-best performers on each database for each performance metric.

Method	KoNViD-1k			LIVE-Qualcomm		
	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
BRISQUE [17]	0.603 (±0.041)	0.614 (±0.042)	0.516 (±0.024)	0.508 (±0.112)	0.501 (±0.094)	10.631 (±0.97)
NIQE [15]	0.574 (±0.038)	0.561 (±0.040)	0.522 (±0.031)	0.476 (±0.124)	0.479 (±0.115)	10.705 (±1.26)
V-BLIINDS [21]	0.694 (±0.031)	0.686 (±0.024)	0.461 (±0.032)	0.637 (±0.087)	0.647 (±0.082)	9.643 (±1.02)
VIIDEO [20]	0.304 (±0.047)	0.312 (±0.054)	0.605 (±0.021)	0.079 (±0.126)	0.112 (±0.089)	11.879 (±1.43)
TLVQM [10]	0.776 (±0.024)	0.784 (±0.028)	0.408 (±0.026)	0.802 (±0.067)	0.788 (±0.079)	7.246 (±0.86)
VSFA [12]	0.757 (±0.023)	0.761 (±0.024)	0.435 (±0.025)	0.719 (±0.072)	0.726 (±0.084)	8.912 (±0.94)
CNN-TLVQM [13]	0.817 (±0.028)	0.819 (±0.031)	0.359 (±0.022)	0.813 (±0.067)	0.827 (±0.071)	6.937 (±0.81)
HEKE [24]	0.739 (±0.021)	0.716 (±0.027)	0.456 (±0.019)	0.702 (±0.063)	0.718 (±0.066)	9.020 (±0.78)
RAPIQUE [23]	0.796 (±0.022)	0.804 (±0.026)	0.364 (±0.023)	0.668 (±0.086)	0.691 (±0.087)	9.497 (±0.77)
VIDEVAL [71]	0.766 (±0.028)	0.781 (±0.025)	0.427 (±0.022)	0.698 (±0.105)	0.723 (±0.095)	9.229 (±0.83)
MDTVSFA [52]	0.784 (±0.025)	0.792 (±0.029)	0.378 (±0.024)	0.811 (±0.052)	0.807 (±0.052)	7.011 (±0.86)
LSCT-PHIQNet [30]	**0.856** (±0.027)	0.846 (±0.023)	0.324 (±0.021)	0.821 (±0.061)	0.811 (±0.056)	6.781 (±0.81)
Proposed	0.853 (±0.019)	**0.850** (±0.021)	**0.317** (±0.022)	**0.849** (±0.051)	**0.842** (±0.049)	**6.117** (±0.79)

Method	LIVE-VQC			Weighted Average		
	PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
BRISQUE [17]	0.614 (±0.071)	0.579 (±0.063)	13.473 (±1.52)	0.596	0.592	5.375
NIQE [15]	0.642 (±0.054)	0.617 (±0.055)	12.865 (±1.47)	0.584	0.569	5.201
V-BLIINDS [21]	0.708 (±0.052)	0.701 (±0.050)	12.176 (±1.33)	0.692	0.686	4.856
VIIDEO [20]	0.216 (±0.079)	0.167 (±0.089)	16.874 (±2.03)	0.255	0.249	6.557
TLVQM [10]	0.801 (±0.034)	0.803 (±0.041)	10.218 (±1.16)	0.786	0.790	4.001
VSFA [12]	0.736 (±0.038)	0.703 (±0.046)	11.352 (±1.24)	0.747	0.740	4.524
CNN-TLVQM [13]	0.826 (±0.046)	0.817 (±0.043)	**9.765** (±1.11)	0.819	0.819	3.806
HEKE [24]	0.751 (±0.041)	0.745 (±0.036)	10.816 (±1.09)	0.739	0.725	4.388
RAPIQUE [23]	0.743 (±0.041)	0.768 (±0.052)	10.961 (±1.13)	0.767	0.782	4.424
VIDEVAL [71]	0.735 (±0.045)	0.729 (±0.046)	11.271 (±1.24)	0.750	0.761	4.525
MDTVSFA [52]	0.786 (±0.035)	0.744 (±0.037)	10.239 (±1.16)	0.787	0.779	3.962
LSCT-PHIQNet [30]	0.822 (±0.047)	0.808 (±0.033)	9.981 (±1.11)	0.842	0.831	3.843
Proposed	**0.841** (±0.038)	**0.836** (±0.036)	9.785 (±1.08)	**0.849**	**0.845**	**3.698**

in the LIVE-VQC database in terms of PLCC and SROCC. Compared with the second-best performance method, CNN-TLVQM, the performance of our proposed method outperforms about 0.015 and 0.019 in PLCC and SROCC, respectively. In terms of the weighted average results, our proposed model obtains the best overall performance in PLCC and SROCC, with improvements of 0.007 and 0.014 compared to the second-best performance method, respectively.

From the result in Table 2, it is evident that our proposed model outperforms other NR-VQA methods and exhibits better effectiveness and generalization performance on all three video databases. It can prove that our proposed NR-VQA method is more robust and effective than other transfer learning/pre-trained model-based methods.

D. PERFORMANCE EVALUATION ON TRADITIONAL VQA DATABASES

Unlike the UGC databases, which focus on in-capture distortion and videos in the wild, the traditional VQA databases focus on the distortion produced in the compression and transmission process, called post-capture distortions. Most existing NR-VQA methods are designed for either UGC or traditional databases. Therefore, we also tested our proposed model and compared the performance with other state-of-the-art NR-VQA approaches, including V-BLIINDS [21], VIIDEO [20], SACONVA [46], TLVQM [10], VSFA [12], CNN-TLVQM [13], Wang's [72], RIR-Net [25], and HEKE [24], on the two traditional VQA databases, the LIVE and CSIQ video databases, to further demonstrate the effectiveness and generalization. It is noted

TABLE 3. Performance comparison of NR-VQA models on two traditional databases. Note that * are performances taken from the method's original papers.

Method	LIVE		CSIQ	
	PLCC↑	SROCC↑	PLCC↑	SROCC↑
V-BLIINDS [21]	0.846(0.03)	0.831(0.04)	0.846(0.05)	0.861(0.05)
VIIDEO [20]	0.668(0.04)	0.675(0.04)	0.673(0.05)	0.652(0.06)
SACONVA [46] *	0.871	0.857	0.867	0.864
TLVQM [10]	0.636(0.04)	0.642(0.03)	0.718(0.05)	0.722(0.04)
VSFA [12]	0.697(0.04)	0.712(0.04)	0.723(0.04)	0.719(0.05)
CNN-TLVQM [13]	0.812(0.03)	0.835(0.02)	0.827(0.03)	0.841(0.04)
Wang's [72] *	0.853	0.857	0.852	0.860
RIRNet [25] *	0.809	0.783	0.843	0.857
HEKE [24]	0.861(0.03)	0.876(0.02)	0.919(0.03)	0.928(0.04)
Proposed	**0.914**(0.02)	**0.927**(0.02)	**0.939**(0.03)	**0.942**(0.03)

that the results of SACONVA [46], Wang's [72], and RIRNet [25] are duplicated from their papers. Also, the best and the second- best performances of PLCC and SROCC are highlighted in bold and underlined, respectively.

The PLCC and SROCC results of the LIVE and CSIQ video databases are shown in Table 3, where the standard deviation is also provided in the bracket. As we can see, the results of ours outperforms all other NR-VQA methods. It is worth noting that although Wang's [72] adopts the saliency map and frame difference information for learning the CNN model to achieve video quality assessment, it ignores the motion information that we believe is crucial to the frame-level quality prediction, which affect the accuracy of perceptual quality prediction. In contrast, we conjecture that our proposed method focuses on the HVP and motion information, taking account of the motion-aware region map into the CNN model to learn the motion-aware and spatial-aware fusion features at the same time via the semi-supervised learning, and fine-tuning strategies. As a result, Table 3 shows that our proposed model yields significantly higher PLCC and SROCC than Wang's [72]. Besides, the proposed model achieves significant improvements on both PLCC and SROCC compared to the second-best performance HEKE. In the LIVE video database, our proposed method outperforms HEKE [24] by about 0.053 and 0.051 in PLCC and SROCC. Additionally, PLCC and SROCC improvements of 0.02 and 0.014 are achieved in the CSIQ video database. Therefore, it is demonstrated that our proposed model is universal and can achieve a strong correlation with human perception in both UGC and traditional databases.

E. PERFORMANCE EVALUATION ON CROSS DATABASES

To further verify the generalization capability of our proposed model with diverse contents and distortions, this section shows the performances of LIVE, KoNViD-1k, and LIVE-Qualcomm in a cross-database scenario, where LIVE is the traditional database focusing on post-capture distortions, KoNViD-1k is the UGC database focusing on video in the wild and various video contents, and LIVE-Qualcomm is the UGC database focusing on in-capture distortions. We trained our proposed model on one database and tested it with another two databases. Then, under the same experiment scheme, the SROCC performance was compared with TLVQM [10], VSFA [12], CNN-TLVQM [13], and LSCT-PHIQNet [30]. Again, the best and the second-best performances of SROCC are highlighted in bold and underlined, respectively. Table 4 clearly shows that the generalization ability of our proposed model trained on the three databases outperforms others in SROCC.

TABLE 4. SROCC results for cross-databases and combined database training and testing.

Training Database	Method	Testing Database		
		LIVE	KoNViD-1k	LIVQ-Qualcomm
LIVE	TLVQM [10]	-	0.517	0.559
	VSFA [12]	-	0.624	0.614
	CNN-TLVQM [13]	-	0.667	0.673
	LSCT-PHIQNet [30]	-	0.641	0.663
	Proposed	-	**0.713**	**0.721**
KoNViD-1k	TLVQM [10]	0.497	-	0.472
	VSFA [12]	0.616	-	0.627
	CNN-TLVQM [13]	0.651	-	0.598
	LSCT-PHIQNet [30]	0.642	-	0.623
	Proposed	**0.691**	-	**0.682**
LIVQ-Qualcomm	TLVQM [10]	0.574	0.533	-
	VSFA [12]	0.645	0.624	-
	CNN-TLVQM [13]	0.621	0.569	-
	LSCT-PHIQNet [30]	0.652	0.637	-
	Proposed	**0.704**	**0.687**	-
All Combined	TLVQM [10]	0.554	0.514	0.531
	VSFA [12]	0.623	0.635	0.659
	CNN-TLVQM [13]	0.667	0.656	0.661
	LSCT-PHIQNet [30]	0.656	0.668	0.656
	Proposed	**0.708**	**0.703**	**0.711**

Besides, we also randomly picked up samples from all three video databases as training, and then used the rest of all samples for evaluation. When the model is trained on this combined database, as shown in Table 4, the performance of our model significantly outperforms the second-best method, CNN-TLVQM and LSCT-PHIQNet, improving SROCC by 0.041, 0.035 and 0.04 on the LIVE, KoNViD-1k and LIVE-Qualcomm video databases, respectively. Since our proposed model can achieve satisfactory results in the cross-database and combined- database scenario, in which three databases contain different and diverse video contents and distortions, the generalization capability of our proposed model is demonstrated, and it can be concluded that our model is universal for different video contents and all types of distortions.

F. ABLATION STUDY OF MOTION-AWARE REGION MAP, SEMI-SUPERVISED LEARNING, AND FINE-TUNING STRATEGIES

As mentioned in Sections III, based on the concept of ROI, SM in (5) can guide image quality feature learning

TABLE 5. Ablation study of our proposed model with various training settings on multi-channel CNN.

Method	LIVE		KoNViD-1k	
	PLCC↑	SROCC↑	PLCC↑	SROCC↑
Multi-CNN$_{SM+w/oSF}$	0.873	0.884	0.824	0.813
Multi-CNN$_{OFM+w/oSF}$	0.883	0.891	0.828	0.829
Proposed	0.939	0.947	0.886	0.891

by focusing on salient regions, as OFM in (6) can guide frame-level quality feature representation learning based on motion-aware regions, and the domain gap can then be reduced by the semi-supervised learning and fine-tuning strategies using (13). Therefore, to demonstrate the effects of SM, OFM, and these techniques for frame-level quality feature extraction, the ablation study was performed on the multi-channel CNN model using various training settings. It includes three combi- nations: VQA by using SM on the pre-trained multi-channel CNN model without performing the semi-supervised learning and fine-tuning strategies (Multi-CNN$_{SM+w/oSF}$), VQA by using OFM on the pre-trained multi-channel CNN model without performing semi-supervised learning and fine-tuning strategies (Multi-CNN$_{OFM+w/oSF}$), and VQA by using OFM to perform semi-supervised learning and fine-tuning strategies on the pre-trained multi-channel CNN model (Proposed). It is noted that we only used the training set and validation set data to perform the ablation study. The experimental results on the LIVE and KoNViD-1k video databases are shown in Table 5.

As we can see in Table 5, PLCC and SROCC of Multi-CNN$_{SM+w/oSF}$ and Multi-CNN$_{OFM+w/oSF}$ are similar. This proves that both of SM and OPM can be used to highlight important regions on a frame. However, due to the domain gap between stillness salient regions on images and motion-aware regions on video frames, the improvement is limited when directly applying OFM on the CNN model pre-trained on SM. After performing semi-supervised learning and fine-tuning strategies, the proposed model can outperform Multi-CNN$_{SM+w/oSF}$ and Multi-CNN$_{OFM+w/oSF}$ on LIVE and KoNViD-1k in terms of PLCC and SROCC. Thus, it shows that the semi-supervised learning and fine-tuning strategies can reduce the domain gap, resulting in more accurate frame-level quality feature representation while considering motion information.

G. ABLATION STUDY OF CAHNNEL ATTENTION MECHANISM

Table 6 shows an ablation study using regular residual blocks instead of SENet blocks (channel attention mechanism) on the multi-channel CNN model. The results show an improvement when using the SENet blocks to the multi-channel CNN model. It demonstrates that channel attention mechanism can reinforce crucial features and weaken inconsequence features, resulting in better quality feature representation learning.

TABLE 6. Ablation study on multi-channel CNN model with and without channel attention mechanism.

Method	LIVE		KoNViD-1k	
	PLCC↑	SROCC↑	PLCC↑	SROCC↑
Without the channel attention mechanism	0.917	0.925	0.871	0.879
Proposed	0.939	0.947	0.886	0.891

TABLE 7. Performance comparison of NR-VQA models. Note that FQFR, TF, and CF represents the frame quality feature representation, temporal features and color-aware features, respectively.

Features			LIVE		KoNViD-1k	
FQFR	TF	CF	PLCC↑	SROCC↑	PLCC↑	SROCC↑
✓			0.898	0.912	0.861	0.867
✓	✓		0.917	0.919	0.868	0.872
✓		✓	0.915	0.921	0.871	0.869
✓	✓	✓	0.939	0.947	0.886	0.891

TABLE 8. Computational complexity (seconds per video) comparison of NR-VQA models.

Method	1280 × 720	1920 × 1080
TLVQM [10]	68.2 sec	167.4 sec
CNN-TLVQM [13]	56.8 sec	114.1 sec
HEKE [24] (≈9.6 frame/sec)	16.4 sec	26.7 sec
RAPIQUE [23] (1 frame/sec)	7.3 sec	11.5 sec
VIDEVAL [71]	107.3 sec	284.9 sec
LSCT-PHIQNet [30]	45.8 sec	80.4 sec
Proposed	37.8 sec	66.9 sec
Proposed_Light	23.2 sec	38.7 sec

H. ABLATION STUDY OF FEATURES LEARNING

This ablation study performs feature selection to analyze the performance gain from features, which also demonstrates the effectiveness of the feature learning ability of our proposed model. There are four groups in the experiment: prediction by our proposed model with the frame-level quality feature representation only (FQFR), prediction by our proposed model with the frame-level quality feature representation and temporal features (FQFR+TF), prediction by our proposed model with the frame-level quality feature representation and color-aware features (FQFR+CF), and prediction by our proposed model with all features (FQFR+TF+CF). The experimental results are shown in Table 7.

Although the frame-level quality feature representation with the GRU model can achieve a satisfactory result, it can be seen that incorporating the temporal or color-aware features

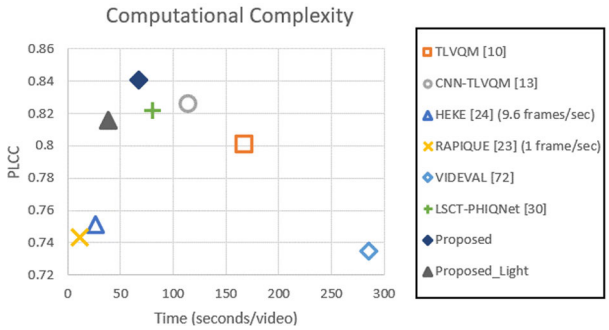


FIGURE 4. The PLCC results in LIVE-VQC video database (collected from Table 2) against the computational complexity runtime with 1080p resolution.

can also be of great help in predicting precise video quality scores with HVP characteristics. When only frame-level quality feature representation is used, the PLCC results are around 0.898 and 0.861 in the LIVE and KoNViD-1k video databases. When incorporating the frame-level quality feature representation with temporal features or color-aware features, the PLCC results in LIVE improved from 0.898 to 0.917 and 0.915. Furthermore, after combining the frame-level quality feature representation, temporal features, and color-aware features into our model, it can achieve the best results by learning the gradient of temporal features and temporal variation of spatial features along with the time series as spatiotemporal features, thereby enhancing video quality prediction with the help of HVP characteristics.

I. COMPUTATIONAL COMPLEXITY

Computational complexity is another major concern in order to apply VQA methods in practical applications. Therefore, we evaluate the computational complexity of our proposed model and six competing NR-VQA models for benchmarking. For a fair comparison, all methods were tested on the same device operating on Windows 10 platform with Intel i9-10900K CPU, 64G RAM, and NVIDIA GeForce RTX 3090 24G GPU.

The comparison results of computational complexity are shown in Table 8. Two videos with different resolutions (720p and 1080p) from the LIVE-VQC were tested to compute the required runtime of a video for each NR-VQA method. Although our proposed model requires higher computational complexity than HEKE [24] and RAPIQUE [23], which perform temporal downsampling, our proposed model delivers better prediction accuracy results, as shown in Fig. 4. Meanwhile, for our proposed model, we can also perform temporal downsampling by a factor of two to form a lighter mode (Proposed_Light), i.e., feature extraction every 2 frames instead of every frame, which can further reduce the computational complexity by sacrificing a slight accuracy. The Proposed_light now has a complexity closer to HEKE. Notably, the Proposed_Light can achieve high accurate performance as shown in Fig. 4. Also, our proposed model is faster than other NR-VQA methods and has better PLCC performance. Specifically, compared with LSCT-PHIQNet [30] and CNN-TLVQM [13], our proposed model achieves about 20% and 50% reduction in computational complexity. Overall, Fig. 4 demonstrates that our proposed model has the best performance in the trade-off between accuracy and computational complexity.

V. CONCLUSION

In this paper, by using SSL, we developed a quality feature learning through multi-channel CNN using non-human annotated labels, and GRU while considering HVP characteristics of NR-VQA. First, we solve the limitations of the lack of available human-annotated label data for the VQA task by a SSL-based multi-channel CNN based on the image quality feature learning method in IQA domain. Second, we bridge the domain gap between the IQA and VQA tasks by adopting semi-supervised learning and fine-tuning strategies in the pre-trained CNN model. The model takes motion-aware information into consideration to optimize frame-level quality feature representation learning. Finally, a GRU model is established to explore spatiotemporal features and gradient of temporal features of video to estimate the video quality by incorporating the frame-level quality feature representation, and HVP-related temporal and color-aware features. Experimental results demonstrate the robustness and generalization of our proposed model, which is practical in real applications and is strongly related to human perception. In the future, one of the ways is to replace GRU with transformer to exploit the long-range dependencies for further improvement. Also, a lightweight NR-VQA model is essential for the further development of real-time applications.

REFERENCES

[1] Cisco. (2017). *Cisco Visual Networking Index: Forecast and Trends, 2017 to 2022*. Whitepaper. Accessed: Apr. 2021. [Online]. Available: https://cyrekdigital.com/uploads/content/files/white-paper-c11-741490.pdf

[2] 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, Jan. 2010.

[3] H. C. Soong and P. Y. Lau, "Video quality assessment: A review of full-referenced, reduced referenced and no-referenced methods," in *Proc. IEEE 13th Int. Colloq. Signal Process. Appl. (CSPA)*, Penang, Malaysia, Mar. 2017, pp. 232–237.

[4] P. V. Vu and D. M. Chandler, "ViS$_3$: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imag.*, vol. 23, no. 1, Feb. 2014, Art. no. 013016.

[5] 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.

[6] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2505–2508.

[7] 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.

[8] 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, Apr. 2013.

[9] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

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

[11] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 546–554.

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

[13] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3311–3319.

[14] T. J. Liu, W. Lin, and C. C. J. Kuo, "Recent developments and future trends in visual quality assessment," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Submit Conf.*, Oct. 2011, pp. 1–10.

[15] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Apr. 2012.

[16] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

[17] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[18] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, p. 32, Sep. 2016.

[19] K. Manasa and S. S. Channappayya, "An optical flow-based no-reference video quality assessment algorithm," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2400–2404.

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

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

[22] F. Yi, M. Chen, W. Sun, X. Min, Y. Tian, and G. Zhai, "Attention based network for no-reference UGC video quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1414–1418.

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

[24] Y. Liu, J. Wu, L. Li, W. Dong, J. Zhang, and G. Shi, "Spatiotemporal representation learning for blind video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3500–3513, Jun. 2022.

[25] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "RIRNet: Recurrent-in-recurrent network for video quality assessment," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 834–842.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[28] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2656–2666.

[29] M. Ebrahim, M. Al-Ayyoub, and M. A. Alsmirat, "Will transfer learning enhance ImageNet classification accuracy using ImageNet-pretrained models?" in *Proc. 10th Int. Conf. Inf. Commun. Syst. (ICICS)*, Jun. 2019, pp. 211–216.

[30] J. You, "Long short-term convolutional transformer for no-reference video quality assessment," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2112–2120.

[31] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th ACM Int. Conf. Multimedia*, Oct. 2006, pp. 815–824.

[32] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, *arXiv:1803.07728*.

[33] G. Larsson, M. Maire, and G. Shakhnarovich, "Learning representations for automatic colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 577–593.

[34] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6874–6883.

[35] C. Vondrick, A. Shrivastava, A. Fathi, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," in *Proc. ECCV*, 2018, pp. 391–408.

[36] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 69–84.

[37] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1422–1430.

[38] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Nov. 2020, pp. 1691–1703.

[39] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[40] C. J. Howard and A. O. Holcombe, "Unexpected changes in direction of motion attract attention," *Attention, Perception, Psychophys.*, vol. 72, no. 8, pp. 2087–2095, Nov. 2010.

[41] A. von Mühlenen and M. Conci, "The role of unique color changes and singletons in attention capture," *Attention, Perception, Psychophys.*, vol. 78, no. 7, pp. 1926–1934, Oct. 2016.

[42] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot, "No-reference image blur assessment based on discrete orthogonal moments," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 39–50, Jan. 2016.

[43] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang, and B. Chen, "No-reference quality assessment of deblocked images," *Neurocomputing*, vol. 177, pp. 572–584, Feb. 2016.

[44] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, "No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1030–1040, May 2017.

[45] Y. Zhang, X. Gao, L. He, W. Lu, and R. He, "Blind video quality assessment with weakly supervised learning and resampling strategy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2244–2255, 2019.

[46] Y. Li, L. M. Po, C. H. Cheung, X. Xu, L. Feng, F. Yuan, and K. W. Cheung, "No-reference video quality assessment with 3D shearlet transform and convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1044–1057, Jun. 2016.

[47] S. Ahn and S. Lee, "Deep blind video quality assessment based on temporal human perception," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 619–623.

[48] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

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

[50] K. Greff, R. K. Srivastava, J. Koutnìk, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space Odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017.

[51] W. Wu, Z. Liu, Z. Chen, and S. Liu, "No-reference video quality assessment based on similarity map estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 181–185.

[52] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-Wild videos with mixed datasets training," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1238–1257, Apr. 2021.

[53] Z.-L. Chu, T.-J. Liu, and K.-H. Liu, "No-reference video quality assessment by a cascade combination of neural networks and regression model," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 4116–4121.

[54] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ: 'Patching up' the video quality problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14019–14029.

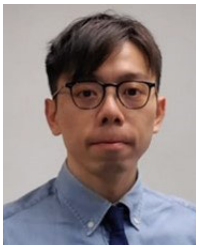
[55] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[56] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[57] S. Montabone and A. Soto, "Human detection using a mobile platform and novel features derived from a visual saliency mechanism," *Image Vis. Comput.*, vol. 28, no. 3, pp. 391–402, Mar. 2010.
[58] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scandin. Conf. Image Anal.* Berlin, Germany: Springer, 2003, pp. 363–370.
[59] K. Sheng, W. Dong, M. Chai, G. Wang, P. Zhou, F. Huang, and C. Ma, "Revisiting image aesthetic assessment via self-supervised feature learning," in *Proc. AAAI Conf. AI*, 2020, pp. 5709–5716.
[60] J. Pfister, K. Kobs, and A. Hotho, "Self-supervised multi-task pretraining improves image aesthetic assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 816–825.
[61] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Jun. 2020.
[62] D. H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn.*, 2013, vol. 3, no. 2, pp. 1–12.
[63] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[64] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Sziranyi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1 k)," in *Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2017, pp. 1–6.
[65] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K. C. Yang. (2017). *LIVE-Qualcomm Mobile in-Capture Video Quality Database*. Accessed: Apr. 2021. [Online]. Available: http://live.ece.utexas.edu/research/incaptureDatabase/index.html
[66] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, Feb. 2019.
[67] *CSIQ Video Database*, Lab. Comput. Perception Image Quality, Oklahoma State Univ., Stillwater, OK, USA, 2013.
[68] F. Bellard. *FFMPEG Tool*. Accessed: Apr. 2021. [Online]. Available: http://www.ffmpeg.org
[69] Video Quality Experts Group. *Final Report From the Video Quality Experts Group on the Validation of Objective Quailty Metrics for Video Quality Assessment*. Accessed: Apr. 2021. [Online] Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx
[70] E. C. Larson and D. M. Chandler. *Categorical Image Quality (CSIQ) Database*. Accessed: Apr. 2021. [Online]. Available: http://vision.okstate.ed
[71] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.
[72] X. Wang, P. Shi, and D. Pan, "Spatial-temporal network for no reference video quality assessment based on saliency," in *Proc. Int. Conf. Culture-Oriented Sci. Technol. (ICCST)*, Oct. 2020, pp. 107–111.

NGAI-WING KWONG received the B.Eng. degree (Hons.) from The Hong Kong Polytechnic University, Hong Kong, in 2018, where he is currently pursuing the Ph.D. degree with the Digital Signal Processing Laboratory. His current research interests include deep learning and image and video processing.

YUI-LAM CHAN (Member, IEEE) received the B.Eng. (Hons.) and Ph.D. degrees from The Hong Kong Polytechnic University, Hong Kong, in 1993 and 1997, respectively. In 1997, he joined The Hong Kong Polytechnic University, where he is currently an Associate Professor with the Department of Electronic and Information Engineering. He is actively involved in professional activities. He has authored over 150 research papers in various international journals and conferences. His research interests include multimedia technologies, signal processing, image and video compression, video streaming, video transcoding, video conferencing, digital TV/HDTV, 3DTV/3DV, multiview video coding, machine learning for video coding, and future video coding standards, including screen content coding, light-field video coding, and 360-degree omnidirectional video coding. He was the Secretary of the 2010 IEEE International Conference on Image Processing. He was also the Special and Demo Session Co-Chair of the IEEE International Conference on Visual Communications and Image Processing and the Publication Chair of the IEEE International Conference on Multimedia and Expo. He was an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING.



SIK-HO TSANG received the Ph.D. degree from The Hong Kong Polytechnic University (PolyU), Hong Kong, in 2013. He was a Research Fellow with PolyU. He is currently a Postdoctoral Fellow with the Centre for Advances in Reliability and Safety (CAiRS). His research interests include video compressions, such as HEVC and VVC, as well as image and video processing and imaging sensor techniques, such as video quality assessment and blur detection, using deep learning. He is a Reviewer of international journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON BROADCASTING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and IEEE ACCESS.

DANIEL PAK-KONG LUN (Senior Member, IEEE) received the B.Sc. degree (Hons.) from the University of Essex, U.K., in 1988, and the Ph.D. degree from The Hong Kong Polytechnic University, in 1991. He is currently an Associate Professor and the Associate Head of the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. He is active in research. He has published over 160 international journal articles and conference papers. His research interests include signal and image enhancement, sparse representation and applications, 3-D data acquisition, and computational imaging. He is a member of the Digital Signal Processing and Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He is a Chartered Engineer, a fellow of IET, and a Corporate Member of HKIE. He was a General Co-Chair, a Technical Co-Chair, and an Organizing Committee Member of a number of international conferences, including APSIPA 2015, ICME 2017, and VCIP 2020. He was the Chairperson of the IEEE Hong Kong Chapter of Signal Processing, from 1999 to 2000. He was an associate editor or editorial board member of a number of international journals.