Deep Learning Approach for No-reference Screen Content Video Quality Assessment

Ngai-Wing Kwong, Yui-Lam Chan, *IEEE Member*, Sik-Ho Tsang, Ziyin Huang, and Kin-Man Lam, *IEEE Senior Member*

***Abstract*— Screen content video (SCV) has drawn much more attention than ever during the COVID-19 period and has evolved from a niche to a mainstream due to the recent proliferation of remote offices, online meetings, shared-screen collaboration, and gaming live streaming. Therefore, quality assessments for screen content media are highly demanded to maintain service quality recently. Although many practical natural scene video quality assessment methods have been proposed and achieved promising results, these methods cannot be applied to the screen content video quality assessment (SCVQA) task directly since the content characteristics of SCV are substantially different from natural scene video. Besides, only one no-reference SCVQA (NR-SCVQA) method, which requires handcrafted features, has been proposed in the literature. Therefore, we propose the first deep learning approach explicitly designed for NR-SCVQA. First, a multi-channel convolutional neural network (CNN) model is used to extract spatial quality features of pictorial and textual regions separately. Since there is no human annotated quality for each screen content frame (SCF), the CNN model is pre-trained in a multi-task self-supervised fashion to extract spatial quality feature representation of SCF. Second, we propose a time-distributed CNN transformer model (TCNNT) to further process all SCF spatial quality feature representations of an SCV and learn spatial and temporal features simultaneously so that high-level spatiotemporal features of SCV can be extracted and used to assess the whole SCV quality. Experimental results demonstrate the robustness and validity of our model, which is closely related to human perception.**

***Index Terms*—human visual experience, multi-channel convolutional neural network, multi-task learning, no reference video quality assessment, screen content video quality assessment, self-supervised learning, spatiotemporal features**

I. INTRODUCTION

In the era of advanced technology, massive screen content has been generated with rapid development and boosted for various screen content-related applications, such as remote office, online meetings, shared-screen collaboration, and gaming live streaming [1]. For improving compression capability and efficiently transmitting the screen content media, screen content coding has been developed as an extension of high-efficiency video coding (HEVC) and is supported in versatile video coding [2]. However, the SCV will unavoidably be processed by screen content media compression, transmission, and reproduction, thereby generating various distortions and affecting the human visual experience [1]. Consequently, quality assessment for screen content media is highly demanded to maintain the service quality and has drawn increasing attention from researchers.

Fig. 1. (a) The imagery of NSV; (b) The imagery of SCV.

Image/Video quality assessment (IQA/VQA) can generally be evaluated in subjective and objective aspects [3]. Subjective quality assessment invites humans to review distorted images/videos and score their quality based on subjective perception. Although it can estimate the most accurate video quality since it directly reflects the human visual experience, subjective methods are time-consuming and not applicable to real-time processing. In contrast, objective quality assessment is a practical alternative to automatically evaluate image/video quality without the enormous time and labor resources. An ideal objective method should be equivalent to the subjective results. Objective quality assessment methods can be further classified into three types by the degree of use of reference [4]: full-reference (FR), reduced-reference (RR), and no-reference (NR).

Many effective IQA/VQA methods have been proposed by various techniques and achieved promising results for natural scene images (NSI) [6-10] and natural scene video (NSV) [11-18]. However, these methods cannot be applied to the SCVQA task directly since the content characteristics of SCV are substantially different from NSI/NSV [19]. As illustrated in Fig. 1(a), NSI/NSV is captured by the camera from real-world scenes, which often include landscapes, people, and wildlife. This imagery typically comprises camera noise, complex textures and contents, rich colors, and smooth edge transitions. In contrast, SCV generally consists of computer-generated content, such as text, tables, animations, and computer screens. SCV may sometimes incorporate elements of traditional natural content [20]. Fig. 1(b) illustrates the distinctive features of SCV, characterized by sharp edges, prominent textual elements, repetitive patterns, and limited color variations [20]. Consequently, the content attributes of SCV differ significantly from those of NSI/NSV. Moreover, despite the content characteristics of screen content image (SCI) being similar to

Ngai-Wing Kwong, Yui-Lam Chan, Sik-Ho Tsang, Ziyin Huang, and Kin-Man Lam, are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, (e-mail: ngai-wing.kwong@connect.polyu.hk;enylchan@polyu.edu.hk; en.ho@connect.polyu.hk; ziyin.huang@connect.polyu.hk; enkmlam@polyu.edu.hk). The work presented in this article is supported by the Hong Kong Research Grants Council (RGC) under Research Grant PolyU 152069/18E, and Innovation and Technology Fund - Partnership Research Programme (ITF-PRP) under PRP/036/21FX.

SCV, the SCI quality assessment (SCIQA) method [20-26] cannot effectively evaluate the perceptual quality of the SCV since SCV has the additional temporal and spatiotemporal information compared with SCI [27]. Therefore, a precise VQA method designed explicitly for SCVs is still to be developed and highly desired.

To the best of our knowledge, only three FR methods [28-30], and one NR method [31] have been explicitly designed and proposed for SCVQA, namely FR-SCVQA and NR-SCVQA, respectively. Although these FR-SCVQA methods have demonstrated promising results, their practical application is limited as reference videos are not always accessible in real-world scenarios [5]. Besides, despite the NR-SCVQA metrics proposed by Li [31] have also shown encouraging results by exploring fifteen features from intra- and inter-frames, the generalization of perceptual SCVQA is restricted since it is a handcrafted feature-based method, which only focuses on specific distortions that limit its performance. There is plenty of room for improvement. Recently, many deep learning models that can learn the hidden features and feature representation have been proposed for NR natural scene video quality assessment (NR-NSVQA). Therefore, developing a deep learning-based NR-SCVQA method is our goal to resolve the above critical issue for the SCVQA task.

Developing a deep learning-based NR-SCVQA method, similar to the NR-NSVQA method, presents a common challenge: the substantial computational power and memory required for training when directly applying deep neural network models for VQA. Given that a raw video typically contains high resolution and frame rate, processing the entire VQA database at once for end-to-end training is impossible due to the limited computational power and memory size of the graphics processing unit. According to NR-NSVQA methods, separating the spatial and temporal learning process is realistic to alleviate the above problems. However, there is no robust label for the spatial learning process since each distorted video only provides a single perceived video quality, Mean Opinion Score (MOS), graded by reviewers as ground truth, and there is no human-annotated label (frame quality score) for each frame for the spatial learning process. Therefore, NR-NSVQA methods utilized some pre-trained models to extract spatial feature representation of frames for the VQA task. For example, the works in [14-15] and [32] used the CNN model pre-trained on ImageNet [33] to extract the frame feature representation from the image classification task to the VQA target domain and then feed all features into the regression model for temporal learning. However, given the significant difference in content characteristics between SCV and NSI/NSV, as well as images from ImageNet used for image classification tasks, the aforementioned pre-trained model may lead to a domain gap issue, resulting in an improper spatial feature representation of screen content frame (SCF) for the SCVQA task. This represents a major challenge in the development of the deep learning-based NR-SCVQA method.

To resolve the above issue, we propose a novel deep learning-based NR-SCVQA method. Our method leverages a multi-task self-supervised learning (SSL) multi-channel CNN model in combination with a time-distributed CNN transformer model (TCNNT) to learn the optimized spatial quality feature representation of SCF and the high-level spatiotemporal features of video to predict the human perceived SCV quality score. Specifically, since humans perceive textual and pictorial regions differently, resulting in diverse visual perception characteristics [30], the textual segmentation method is first employed to separate the SCF into the pictorial and textual parts. The SCF is combined with the saliency map derived from its pictorial region and the edge information map obtained from its textual region, which is then input into our proposed multi-channel CNN for handling separately and fusing. Subsequently, we employ a multi-task SSL approach by integrating the pairwise ranking task [10] alongside additional SSL tasks, such as the distortion classification task and degradation degree task, as a new multi-task SSL multi-channel CNN model to learn the optimized spatial quality feature representation of SCF, ultimately benefiting the SCVQA downstream task. Afterward, all SCF spatial quality feature representations of an SCV are extracted and fed into our proposed TCNNT model. The TCNNT uses a time-distributed CNN model to process data sequences with temporal dependencies, further enhancing spatial feature extraction for video by processing each timestamp of frames simultaneously. Then, the transformer model in TCNNT handles temporal features using a self-attention mechanism, capturing dependencies across time steps. Thus, our TCNNT model learns the spatial and temporal features concurrently, optimizing spatiotemporal features of SCV for precise SCVQA quality prediction. The contributions of this work are summarized as follows:

- To our best knowledge, this is the first deep-learning approach explicitly designed for the NR-SCVQA task. This method compensates the constraints of FR methods [28-30] for real-world application and overcomes the limitations of the handcrafted feature-based NR method [31], improving the performance and generalization of perceptual quality evaluation for SCV.
- We propose the multi-channel CNN model to learn the optimized spatial quality feature representation of SCF via multi-task SSL (pairwise ranking, distortion classification, and degradation degree tasks). This approach compensates for the shortage of human-annotated labels for SCFs during the spatial feature learning process. This strategy also enhances our model's performance and generalization capabilities.
- We are the first to combine the time-distributed CNN module and the transformer encoder module (TCNNT model) for the VQA task. Unlike the existing NR-NSVQA methods in [14-15], [32] which separate the spatial and temporal learning process, the goal for our proposed TCNNT model is to simultaneously learn spatial and temporal features in an end-to-end manner, ultimately resulting in optimized/high-level spatiotemporal features of SCV, providing a precise final SCV quality score.

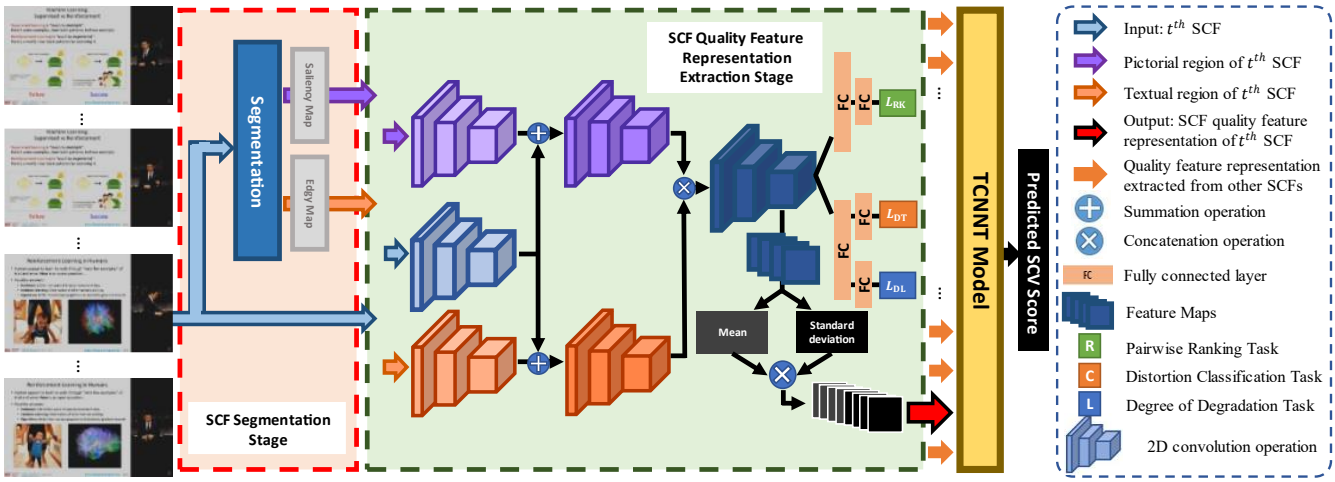

Fig. 2. The framework of our proposed NR-SCVQA model.

- By evaluating our model on two SCVQA databases, we verify that our model can predict the SCV quality via deep features representation learning, precisely close to human visual perception compared with other classic and state-of-the-art NSIQA/NSVQA and SCIQA/SCVQA methods.

The rest of this paper is organized as follows. In Section II, we present the relevant research works. 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 Works in NR-VQA

A. Natural Scene VQA Methods

FR-NSVQA methods: At the early development stage of FR-NSVQA, some successful and efficient FR-NSIQA methods, such as PSNR, SSIM [6], and GMSD [7] are used as the spatial feature extraction algorithms in some FR-NSVQA approaches and incorporate some temporal pooling methods and weight functions to obtain the NSV quality score. Since temporal information is also critical for VQA, some studies instead design models focusing on the quality of the complete video rather than the average quality of frames. MOVIE [11] developed a general, spatio-spectrally localized multiscale framework to evaluate the video quality by considering both spatial and temporal (and spatiotemporal) along motion trajectories. ST-MAD [12] extends the image-based algorithm (most apparent distortion) to quantify motion-based distortion of spatial-temporal slices created by taking time-based slices of the original and distorted videos to achieve practical quality assessment for natural video. In STRRED [13], a Gaussian scale mixture model is used to compute the amount of spatial and temporal information differences between the reference and distorted videos, which could be further combined to obtain the spatiotemporal-reduced reference entropic differences to evaluate the video quality.

NR-NSVQA methods: Similar to the FR-NSVQA methods, some NR-NSIQA methods [8-9] are used to extract the visual features of frames or the quality of frames to predict the NSV quality with some temporal pooling methods. With the rise of deep neural network models that can learn the data representation, hidden features, and abstract features automatically, most NR-NSVQA models use pre-trained deep neural network models for spatial feature extraction instead of handcrafting spatial features. For example, VSFA [14] extracts content-aware features from a CNN model pre-trained on the image classification task, which can also compensate for the lack of enormous training samples to train the robust deep CNN model, and then predicts the video quality using a gated recurrent unit (GRU) temporal-memory model. The authors in [36] also improved this method by training on mixed datasets. Also, CNN-TLVQM [15] combines the handcrafted human visual system features extracted from TLVQM [37] and the spatial features obtained from a pre-trained CNN via transfer learning. It then uses a support vector regression model to evaluate the predicted quality score. However, due to the content and characteristics of SCV being substantially different from NSV, VQA methods adopted by NSV cannot provide the optimal feature representation for SCV, which obstructs the effectiveness and accuracy of SCVQA, which is also proven by Li [31].

B. Screen Content VQA Methods

To the best of our knowledge, only three FR methods [28-30], and one NR method [31] are explicitly designed for SCVQA. The work in [28] conducted a subjective study for SCVs and established the first SCV database (SCVD). Meanwhile, it was verified that existing IQA/VQA methods cannot effectively evaluate the perceptual quality of the SCV. Therefore, the first FR-SCVQA method [28], namely the spatiotemporal Gabor feature tensor-based model (SGFTM) was designed explicitly for SCVs. The SGFTM uses the 3D-Gabor filter to extract spatiotemporal screen content visual feature representation of reference and distorted SCV and evaluate the quality of distorted SCV by measuring their similarity. Moreover, Li [29] built another SCV database for compressed SCVs (CSCVQ). The work in [29] also investigated that general IQA and VQA cannot fulfill the need for quality assessment for screen content due to content and characteristics differences. It also proposed a new FR-SCVQA method, MS-RSDS, that measures the SCF quality by

computing the similarity of multi-scale relative standard deviation difference of two continuous time references and distorted frames to obtain the final SCV quality score by average pooling. Besides, Li further proposed the first and only NR-SCVQA method [31], in which fourteen handcrafted screen content spatial features and one handcrafted temporal feature are extracted and incorporated with a support vector regression to obtain the final SCV quality score. Recently, HSFM [30], the latest FR-SCVQA method, uses 3D Laplacian of Gaussian to capture the spatiotemporal edge information and then combines it with natural spatiotemporal feature extracted by 3D mean subtracted and contrast normalized with a weighting strategy to evaluate the SCV quality score. Although the FR-SCVQA and NR-SCVQA methods described above have shown promising results, these handcrafted-based methods only focus on specific distortions and features, which restrict the generalization of perceptual quality features of SCVs and the accuracy of SCVQA.

III. Proposed Method

To better predict the SCV quality, we propose a novel deep learning-based NR-SCVQA method that adopts a multi-task SSL multi-channel CNN model with a TCNNT model. The framework of our proposed model is shown in Fig. 2. First, we divide the pictorial and textual regions of SCF via the textual segmentation method and transform them into the saliency map and edgy information map separately. Then, the multi-channel CNN model is used to explore spatial quality features of pictorial and textual regions separately via multi-task learning (including pairwise ranking task, distortion classification task, and degradation degree task). Therefore, the model can be trained to assess SCFs in an SSL manner without using any human-annotated labels, which can be the pretext task to extract the optimized spatial quality feature representation for the SCVQA downstream task. Lastly, all SCF spatial quality feature representations at each timestamp of an SCV are extracted and fed into our proposed TCNNT model. The TCNNT model then further processes the spatial and temporal features simultaneously so that high-level spatiotemporal features of SCV can be extracted to predict the final SCV quality comprehensively. We detail each part in the following sub-sections.

A. Multi-task SSL with Multi-channel CNN Model

Naive supervised learning is impossible due to the lack of a ground-truth label for each SCF. To compensate for the shortage of labels, we propose to pre-train the multi-channel CNN to learn the spatial quality feature representation of SCF with multi-task SSL strategy by identifying pairwise ranking, distortion type, and degradation degree. SSL is a form of unsupervised learning that can let the network learn critical features from unlabeled data by providing a non-human annotated supervision signal [38]. For example, the model in [39] predicts the image rotation to learn the image representation for the image classification task. [40] solved the Jigsaw puzzles via SSL and repurposed them for the object

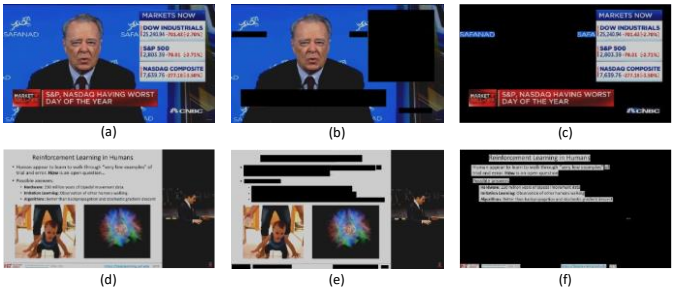

Fig. 3. Results from the SCF segmentation stage. (a) and (d) are the distorted SCF; (b) and (c) are the pictorial and textual region of (a); (e) and (f) are the pictorial and textual region of (d)

Fig. 4. (a) Saliency map of the pictorial region in Fig. 3(b); (b) Edge information map of the textual region in Fig. 3(c)

detection task. [41] designed the model to predict the relative position between the central patch and its neighboring location, which can capture visual similarity across images for visual representation learning. For the related task, RankIQA [10] pre-trained the Siamese network to rank the quality of images to learn image quality features for the IQA task.

Moreover, compared with the aforementioned task-specific SSL models, multi-task SSL aims to let the model learn multi-tasks in parallel while using shared features. By solving multiple learning tasks at the same time, the shared features and knowledge learned from each task are helpful in learning other tasks for better feature representation learning, which results in improving the learning efficiency, accuracy, and generalization learning ability of the model [42]. Motivated by this, we propose to integrate the pairwise ranking task, distortion classification task, and degradation degree task to train our multi-channel CNN model so that our new model can assess SCFs in a multi-task SSL manner with better performance.

1) Pre-processing Stage

Before the training process of our multi-channel CNN model, we first use the textual segmentation method to separate the SCF into the pictorial region (I_P) and textual region (I_T). This is because an SCF usually contains mixed screen content and natural scene information. However, the content characteristics of screen content and natural scene information are substantially different. Also, human perception varies between textual and pictorial regions, resulting in diverse visual perception characteristics for these areas [20]. Therefore, the evaluation of SCF must account for these varied visual characteristics. Consequently, it becomes essential to extract the spatial features of pictorial and textual regions of SCF independently.

Segmentation: In this paper, we implement the fast CNN-based document layout analysis algorithm [43] to divide the

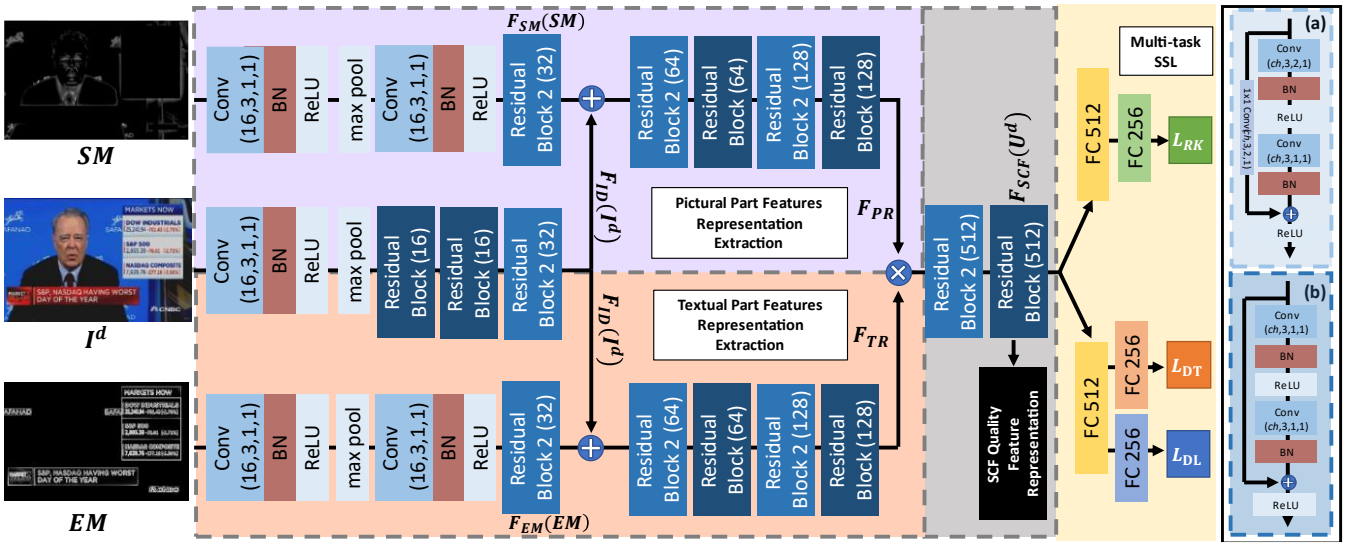

Fig. 5. The network architecture of our proposed CNN model. (a) The structure of Residual Block 2 (*ch*); (b) The structure of Residual Block (*ch*). Conv(*ch*, *kn*, *st*, *pd*) represents the 2D convolution operation where *ch* is the output channel, *kn* × *kn* is the kernel size, *st* represent the size of stride and *pd* is the padding size. BN, FC and GP represent the batch normalization operation, fully connected layer and global pooling.

SCF into pictorial and textual regions. First, the SCF is segmented into blocks of content by the running length algorithm described in [44] and a 3×3 dilation operation. Then, the horizontal and vertical projections of these SCF blocks are determined and fed into a one-dimensional CNN model to classify SCF blocks into text, table, or graph. Finally, the classification results are used to group all text boxes and table boxes into textual regions. Graph boxes, along with the remaining background, are categorized as part of pictorial regions of the frame. As a result, the SCF is divided into pictorial I_P and textual I_T regions, as shown in Fig. 3.

Pictorial Region: After separating the pictorial I_P and textual I_T regions of SCF, and acknowledging the complex nature of human visual perception as well as the diverse ways in which viewers process and perceive different content types and regions, we calculate the saliency map of I_P as the visual-aware map for the pictorial region of SCF as shown in Fig. 4(a). This approach is based on the understanding that pictorial regions often contain intricate and diverse visual information. The utilization of saliency detection is crucial for identifying regions within the pictorial regions that are most likely to capture the viewer's attention. This valuable insight enables our model to assign variable weights to different regions, proportionate to their relevance. This ensures that the most visually compelling areas exert a more prominent impart on the quality assessment process.

To obtain the saliency map of I_P, we first implement the method in [45] to determine the saliency residuals using the log-spectrum algorithm. Then, we invert the saliency residuals $\mathcal{R}(f)$ from the spectral domain back to the spatial domain to compute the preliminary saliency map as follows:

$$\mathcal{R}(f) = \mathcal{L}(f) - \mathcal{A}(f) \tag{1}$$

$$PSM(f) = \mathcal{F}^{-1}(\mathcal{R}(f)) \tag{2}$$

where f is the form of the pictorial region of the input distorted SCF I_P after Fourier Transform ($f = FT(I_P)$), $\mathcal{A}(f)$ is the real part of f, and $\mathcal{L}(f)$ is the log spectrum of $\mathcal{A}(f)$. $\mathcal{F}^{-1}(\cdot)$ indicates the inversion process of $\mathcal{R}(f)$ from the spectral back to the spatial domain. Moreover, to further extract the fine-grained features of the preliminary saliency map, the visual saliency feature (VSF) method in [46] is applied to calculate the center-surround differences on the salient region in the preliminary saliency map that helps to define borders and compute the final saliency map SM of I_P as follows:

$$SM = VSF(PSM(f)) \tag{3}$$

This saliency map SM represents important pictorial regions and it is used as the input of our multi-channel CNN.

Textual Region: As we all know, the textual region of SCF I_T contains various textual elements with plenty of sharp edges. These sharp edges are vital for ensuring text clarity and legibility. However, video compression may induce artifacts, particular along edges, which negatively impact the readability of characters and thus affect the perceived overall quality. Moreover, it is a well-established fact that human visual attention is naturally drawn to areas with high contrast and clear boundaries, such as the edges of text. Consequently, edge detection plays a pivotal role in quality assessment by highlighting potential focal points within the textual region that are likely to capture viewer attention.

Therefore, to capture the textual visual characteristics of SCF, we compute the Gabor feature map as the edge information map for the textual region of distorted SCF since the receptive field of edge information can be well reflected by the Gabor response. As shown in Fig. 4(b), the Gabor feature map shows the rich edge information of the textual region of Fig. 3(c). To obtain the Gabor feature map of I_T, we convolve I_T with the horizontal- and vertical-oriented odd Gabor filters, $G^h(x,y)$ and $G^v(x,y)$, individually to compute the horizontal and vertical Gabor responses, $H(x,y)$ and $V(x,y)$, as follows:

$$G^h(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left\{\frac{-1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2\right]\right\} \sin(2\pi\omega x)$$

$$G^v(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left\{\frac{-1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2\right]\right\} \sin(2\pi\omega y)$$
$$H(x,y) = G^h(x,y) * I_T(x,y)$$
$$V(x,y) = G^v(x,y) * I_T(x,y) \quad (4)$$

where (x, y) denotes the pixel coordinate in the SCF, ω is the frequency of the sinusoidal function, σ_x and σ_y are the standard deviations of the Gaussian function in the x-direction and y-direction, respectively. After that, the edge information map of I_T, EM, is then constructed by adding the horizontal and vertical Gabor responses together as follows:

$$EM = H(x,y) + V(x,y) \quad (5)$$

This edge information map EM represents important screen content regions and rich edge information used as the input of our multi-channel CNN.

2) Multi-task SSL Multi-channel CNN Model Pretext Task

To train our multi-channel CNN model with no robust label of SCF, we use quality ranking, distortion type, and degradation degree of SCF as non-human annotated supervision signals, or so-called pseudo labels $PL = [RK, DT, DL]$ where RK is the quality ranking indicator of the pairwise SCF, DT indicates the distortion type and DL represents the degradation degree of SCF. Then, PL can be used for the unlabeled data $U^d = [I^d, SM, EM]$, which includes the distorted SCF (I^d) and its corresponding saliency map of I_P (SM) and edge information map of I_T (EM), for our multi-channel CNN model to conduct the multi-task SSL to learn the spatial quality feature representation of SCF for the SCVQA downstream task, which also compensates for the shortage of human-annotated labels for SCFs.

Multi-channel Mechanism: As shown in Fig. 5, in our model, the multi-channel mechanism is used to explore the spatial quality feature of pictorial and textual regions separately. The spatial quality features of the pictorial region, F_{PR}, are first extracted from the channel of I^d and the channel of SM (the violet region in Fig. 5), which is given by:

$$F_{PR} = PRNet\big(F_{ID}(I^d) \oplus F_{SM}(SM)\big) \quad (6)$$

where the symbol $\oplus$ is the element-wise summation operation, $F_{ID}(\cdot)$ and $F_{SM}(\cdot)$ represent the shallow feature extraction processes on I^d and SM, and $PRNet(\cdot)$ is used to extract the spatial quality features focusing on the pictorial region. Thus, we can extract spatial quality features of pictorial regions by concentrating on the critical and visually significant regions within the pictorial area, utilizing the saliency map. This approach enables the model to identify and emphasize the most relevant visual information, leading to a more accurate representation of the pictorial regions of SCF.

In the meantime, the channel of I^d and the channel of EM (the orange region in Fig. 5) focus on learning spatial quality features of the textual region, F_{TR}, which is given by:

$$F_{TR} = TRNet\big(F_{ID}(I^d) \oplus F_{EM}(EM)\big) \quad (7)$$

where $F_{EM}(\cdot)$ represents the shallow feature extraction process on EM, and $TRNet(\cdot)$ is the operation of extracting spatial quality features of the textual region. This approach can effectively integrate the edge information from the textual region into the spatial quality features, emphasizing the significance of edge features within the SCF. By incorporating these edge features, the model can better learn spatial quality features of the textual region, thereby improving the representation of the textual regions of SCF.

Following the extraction of spatial quality features from the pictorial and textual regions by $PRNet(\cdot)$ and $TRNet(\cdot)$ individually, it is essential to fuse these features (F_{PR} and F_{TR}) to evaluate the overall visual perception of the entire SCF, which comprises both pictorial and textual content. Therefore, F_{PR} and F_{TR} are then concatenated and processed to represent the whole spatial quality features of distorted SCF I^d (F_{SCF}) as follows:

$$F_{SCF}(U^d) = SCFNet([F_{PR}, F_{TR}]) \quad (8)$$

where$[\cdot]$ is the concatenation operation, and $SCFNet(\cdot)$ is the feature fusion operation of spatial quality features of both pictorial and textual regions. This fusion process allows for a more holistic assessment of the SCF by considering the interplay between the natural content and screen content, ultimately providing an optimized spatial feature representation of the perceived visual quality of SCF. These SCF spatial quality features, F_{SCF}, are then further employed for various SSL pretext tasks.

Pairwise Ranking Task: The main pretext task of the multi-task SSL is the pairwise ranking task [10]. For two SCFs with the same content and distortion type but with different degradation degrees, I_i^d and I_j^d where i and j are the distortion intensity index and $i \neq j$, the relative SCF quality ranking is known since the SCF quality would be lower with a higher degradation degree generally, e.g., increasing the distortion level of noise or blur results in a less clear SCF and a lower perceptual quality compared to the same SCF with a weaker increase in distortion. Therefore, it is conjectured that by encouraging the model to distinguish the pairwise quality ranking of I_i^d and I_j^d, our new model can learn the crucial spatial quality feature representations of SCF. To conduct the SSL pairwise ranking task, first, spatial quality features of I_i^d, $F_{SCF}(U_i^d)$, are processed by two dense layers, FC_R, to output one scalar, denoted by $F_{PRK}^i = FC_R(F_{SCF}(U_i^d))$ where $U_i^d = [I_i^d, SM_i, EM_i]$. After that, we randomly select a pairwise SCF I_j^d for I_i^d to evaluate the pairwise ranking level. We use the same model and network parameter treated as the Siamese network to compute the output, $F_{PRK}^j = FC_R(F_{SCF}(U_j^d))$, for I_j^d. Therefore, the loss of pairwise ranking task is computed as:

$$L_{PRK}^i = \begin{cases} \max\left(0, F_{PRK}^i - F_{PRK}^j + \varepsilon\right) \ if\ RK = 0 \\ \max\left(0, F_{PRK}^j - F_{PRK}^i + \varepsilon\right) \ if\ RK = 1 \end{cases}$$
$$L_{RK} = \frac{1}{M}\sum^M L_{PRK} \quad (9)$$

where M is the batch size, and ε is the margin. It is noticed that $RK = 1/0$ when the quality ranking of I_i^d is higher/lower than the pairwise SCF I_j^d. Therefore, our model can learn the crucial spatial quality features of SCF by distinguishing the pairwise

quality ranking of I_i^d and I_j^d to improve the SCF quality feature representation learning process.

Distortion Classification Task: Moreover, SCFs with different distortion types, such as transmission, display, and compression distortions, result in explicitly different visual distortion and perceived visual quality [47]. Therefore, in addition to the pairwise ranking task, the distortion classification task is also trained in the multi-task SSL manner for our new model to characterize the distortion class of SCF to enhance the generalization learning ability and efficiency of our model by solving multiple learning tasks at the same time. To conduct the distortion classification pretext task, we also use two dense layers, different from the dense layers used for the pairwise ranking task, to compute the probability of distortion type, denoted by $F_{DT} = FC_{D2}(FC_{D1}(F_{SCF}(U^d)))$. Then, the loss can be computed by F_{DT} and its pseudo labels DT.

$$L_{DT} = \frac{1}{M}\sum^{M} L_{CE}\,(F_{DT}, DT) \tag{10}$$

where $L_{CE}(\cdot)$ is the categorical cross-entropy loss function. By doing so, the SCF quality feature representation learning process of our model can be further improved by exploring explicitly different visual distortions and perceived visual quality from different distortion classes.

Degradation Degree Task: The degree of degradation is I^d is also the crucial clue for the SCF quality feature representation learning. By focusing on the degradation degree of SCF, the model can more effectively comprehend and learn quality variations, thereby refining the spatial quality feature representation and improving the overall prediction accuracy. Therefore, apart from the pairwise ranking task, the degradation degree task is also included. Since each distortion type of SCVs results in a different perceived visual quality range and the distortion level prediction is correlated to distortion type, we use the first dense layer from the distortion classification task, FC_{D1}, as shared features and connect it with a new dense layer, FC_{DD}, for the degradation degree task to improve the learning ability. Therefore, the loss function for the degradation degree task is computed as follows:

$$L_{DL} = \frac{1}{M}\sum^{M} L_{CE}\,(F_{DL}, DL) \tag{11}$$

where $F_{DL} = FC_{DD}(FC_{D1}(F_{SCF}(U^d)))$ and DL is the pseudo labels for distortion level.

Multi-task Learning: Finally, we combine several SSL tasks by weighting and adding the losses together as the overall loss function for our multi-channel CNN model, in which the multi-task learning can improve the learning efficiency, prediction accuracy, and generalization ability of our new model for better SCF quality feature representation learning for the following SCVQA downstream task. The resulting overall loss is

$$L_{overall} = L_{RK} + \beta(L_{DT} + L_{DL}) \tag{12}$$

where β is the weighting parameter.

3) Screen Content Quality Feature Representation Extraction

After completing the training process of the multi-task SSL multi-channel CNN model, we can use it to extract the optimized spatial quality feature representation of SCFs for the SCVQA downstream task. Specifically, the feature maps are extracted at the last convolutional layer in the well-trained multi-channel CNN model as SCF spatial quality feature representation F_{SCF}, as shown in Fig. 5. In practice, we divide the SCF into B non-overlapping frame patches, and each frame patch obtains its spatial quality feature representation F_{SCFP}^b through the well-trained multi-channel CNN model. Ultimately, the mean and standard deviation of all quality feature representations of frame patches within the SCF are taken as its SCF quality feature representation F_{SCF}^t for the SCVQA task as follows:

$$F_{SCF}^t = \left\{\mu\{F_{SCFP}^b\}, \sigma\{F_{SCFP}^b\}\right\}_{b=1}^{b=B} \tag{13}$$

where B is the total number of frame patches in t^{th} SCF, and $\mu(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation operation, respectively. This aggregated feature representation effectively captures the variability and central tendency of the quality features F_{SCFP}^b, providing a more comprehensive quality feature representation of SCF F_{SCF}^t.

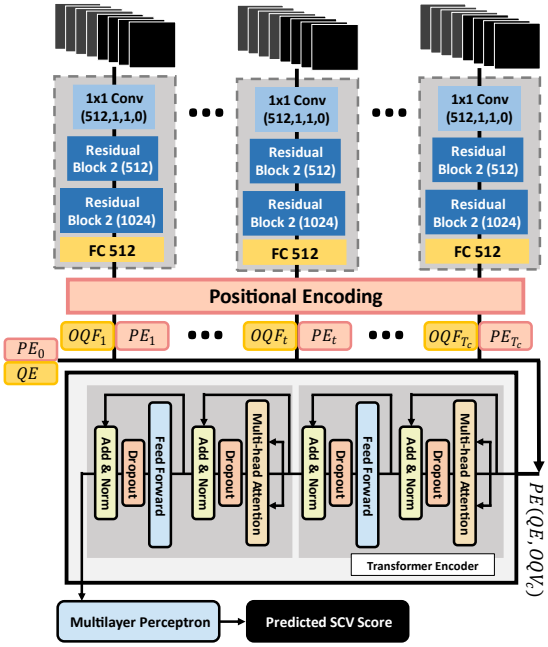

Fig. 6. The network architecture of our proposed TCNNT model.

B. Screen Content Video Quality Prediction via TCNNT Model

Most existing deep learning-based VQA methods [14-15], [32] separate the spatial and temporal learning process for the VQA task, which can only lead to a low-level or sub-optimal spatiotemporal feature learning for the VQA task since the model cannot learn the spatial and temporal features simultaneously (in an end-to-end manner) to evaluate the quality of the whole video. Hence, we propose the TCNNT model, designed to concurrently learn spatial and temporal features in an end-to-end manner, which can distill the high-level and optimal spatiotemporal features to obtain the final precise predicted SCV quality for the SCVQA task.

First, we concatenate all SCF quality feature representations of the c^{th} distorted SCV as a feature vector $SCVQ_c = [F_{SCF}^1, F_{SCF}^2, \ldots, F_{SCF}^{T_c-1}, F_{SCF}^{T_c}]$, where $c = 1,2,3,\ldots,C$, C is the total

number of videos in the SCVQA database, F_{SCF}^{t} is the SCF quality feature representation of t^{th} SCF, and T_c is the total number of frames of the c^{th} distorted SCV. After that, the time-distributed CNN module, $TCNN$, is used to process the $SCVQ_c$ at different timestamps and output the optimized spatial quality feature vector OQV_c as follows:

$$OQV_c = TCNN(SCVQ_c) \tag{14}$$

where $OQV_c = [OQF_1, OQF_2, \dots, OQF_{T_c-1}, OQF_{T_c}]$, OQF_t is the optimized spatial quality feature representation of t^{th} SCF with a shape of 1×512 in each timestamp, as shown in Fig. 6. Unlike the conventional CNN module used in [14-15], [24], [32] for image/spatial feature extraction, time-distributed CNN modules are often employed in video action recognition tasks, as they possess the ability to extract spatiotemporal features and identify patterns in the data, enabling to recognize and classify actions happening over time. Leveraging these capabilities, we propose to adopt the time-distributed CNN module for the VQA task. Through the above process, the time-distributed CNN module is applied to every temporal slice individually, treating samples as different timestamps to extract an optimized spatial quality feature representation for each timestamp and capturing temporal information in the series data, which is different from the standard CNN module that can only process spatial features and cannot handle the time series data. This salient feature of the time-distributed CNN module can be integrated with the temporal pooling method as an end-to-end model, which can simultaneously process the spatial and temporal features to explore the spatiotemporal feature to boost the performance of our model for more precise SCV quality prediction. The effectiveness of the time-distributed CNN model will be further analyzed in Section IV.H.

Furthermore, for the temporal processing, instead of using the long short-term memory (LSTM) or GRU used in [14-15], we adopt the transformer encoder model since the self-attention mechanism of the transformer model allows the model to capture the attention allocation along the whole sequence. It is used to capture the influence of attention distribution crossing temporal and is suitable combined with a time-distributed CNN model for the VQA task to evaluate the whole video quality comprehensively [34-35]. More experimental results of the model combination will be presented in Section IV.H.

Since the training process of our TCNNT model is in an end-to-end manner, as shown in Fig. 6, in which the time-distributed CNN model is responsible for extracting the optimized spatial feature as each timestamp, and the transformer model is capable of exploring temporal features and capturing dependencies across time steps simultaneously so that high-level spatiotemporal features of SCV can be extracted to predict the SCV quality score as follows:

$$\hat{v}_c^L = TCNNT(SCVQ_c) = MLP(TE(PE(QE, OQV_c)))$$
$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{15}$$

where QE is the quality learnable parameter, $PE(\cdot)$ represents the positional encoding process, $TE(\cdot)$ is the operation of the transformer encoder model implemented by following [34] including the multi-head self-attention layers $MultiHead(\cdot)$ (where $Q = X \cdot W^Q, K = X \cdot W^K, V = X \cdot W^V$ are Query, Key, Value, X is the input features ($PE(QE, OQV_c)$) and W represents the parameter matrix) and position-wise fully connected feed-forward layers. Finally, $MLP(\cdot)$ contains two dense layers to predict the final SCV quality score $\hat{v}_c^L$. The loss function of the supervised learning TCNNT model is defined as:

$$\mathcal{L}_{TCNNT} = \frac{1}{N}\sum_{i=0}^{N-1}(\hat{v}_i^L - v_i^L)^2 \tag{16}$$

where N is the batch size, $\hat{v}^L$ represents the final predicted video quality score by our TCNNT model and v^L is the ground truth label (MOS) of the corresponding SCV collected from the subjective study, provided by published databases [28-29]. Therefore, in this end-to-end training process, our proposed TCNNT model can boost performance and enhance SCV quality prediction by learning the high-level spatiotemporal features and temporal attention information simultaneously.

TABLE I
THE SUMMARY OF THE SCV DATABASES

Database	Resolution	# of Videos		Frame Rate (fps)	# of Distortion		Duration (seconds)
		Ref.	Dis.		Type	Level	
SCVD	1920×1080	16	800	30	10	5	10
CSCVQ	1280×720	11	165	30	3	5	10

IV. EXPERIMENTAL RESULTS

A. Screen Content Video Quality Databases and Evaluation

To demonstrate the validity of our proposed model, the performance of our proposed model and various existing classical and latest IQA/VQA methods are evaluated on the two existing SCVQA databases, SCVD [28] and CSCVQ [29]. The summary of the above SCVQA databases is shown in Table I.

1) *SCVD* [28] is the first subjective video database explicitly designed for the SCVQA, containing 800 distorted SCVs. These videos are generated from 16 reference SCVs with 5 degrees of quality degradation on ten different distortion types, including the acquisition and transmission distortions (Gaussian noise, Gaussian blur, motion blur, and packet loss), display distortions (contrast change, color saturation change, and color quantization with dithering), and compression distortions (H.264, HEVC and SCC). Each distorted SCV is of resolution 1920×1080 and a 10-second video with a frame rate of 30 frame per second (fps). The MOS is provided in the range of 20.12 to 74.08.

2) *CSCVQ* [29] is another subjective video database for SCVQA, specifically focusing on compression distortion. It contains 165 distorted SCVs with a resolution of 1280×720 compressed from 11 screen application scenario reference videos using the H.264, HEVC, and HEVC-SCC with five degrees of quality degradation. All SCVs have a duration of 10 seconds with a frame rate of 30 fps, and the MOS ranges from 20.53 to 72.76.

To evaluate the performance of our proposed model, three commonly used metrics: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), and Root Mean Square Error (RMSE), are utilized to measure the accuracy and monotonic consistency between

TABLE II
PERFORMANCE COMPARISON OF SCVQA MODELS ON TWO SCVQA DATABASES. THE TOP THREE BEST RESULTS OF EACH PERFORMANCE INDEX, PLCC, SROCC, AND RMSE, ARE SHOWN IN BOLD WITH COLORS IN RED, BLUE, AND BLACK, RESPECTIVELY. A NOTATION OF * IS USED TO INDICATE DEEP LEARNING-BASED METHODS.

IQA/VQA	FR/NF	Method	SCVD			CSCVQ		
			PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
NSIQA	FR	PSNR	0.6267	0.6213	10.5931	0.8139	0.7862	8.1364
		SSIM [6]	0.7063	0.6896	9.7032	0.8263	0.7949	7.9237
		GMSD [7]	0.7135	0.7768	9.3967	0.8916	0.8763	5.4866
	NR	BRISQUE [8]	0.5867	0.6012	11.1861	0.7113	0.6886	9.6452
		NIQE [9]	0.5541	0.5679	11.5851	0.6984	0.6975	9.9134
SCIQA	FR	ESIM [21]	0.7798	0.7736	8.5236	0.8936	0.8872	5.6687
		SQMS [25]	0.7525	0.7662	8.8216	0.8864	0.8617	5.9376
		SVQI [22]	0.7213	0.7322	9.3024	0.8571	0.8414	6.3653
	NR	Yang's Work[20]*	0.7794	0.7715	8.5578	0.8772	0.8568	6.0684
NSVQA	FR	MOVIE [11]	0.6132	0.5569	10.7596	0.8916	0.8797	5.6716
		STMAD [12]	0.7368	0.7307	9.1362	0.8633	0.8513	6.109
		STRRED [13]	0.7531	0.7448	8.9673	0.8911	0.8962	5.4964
	NR	VSFA [14]*	0.7514	0.7678	9.0108	0.7968	0.8124	7.6718
		VIDEVAL [16]	0.7462	0.7466	9.0964	0.8775	0.8569	5.9511
SCVQA	FR	SGFTM [28]	**0.8315**	**0.8232**	**7.4371**	0.8941	0.8831	5.5108
		MS-RSDS [29]	0.8196	0.8062	7.7934	**0.9308**	**0.9225**	**4.9163**
		HSFM [30]	**0.8486**	**0.8308**	**7.1956**	**0.9463**	**0.9240**	**4.4720**
	NR	Li's Work [31]	0.8063	0.8136	8.1762	0.9232	0.9013	5.0618
		Proposed*	**0.8945**	**0.9073**	**6.1979**	**0.9387**	**0.9312**	**4.7517**

IQA/VQA	FR/NF	Method	Direct Average			Weighted Average		
			PLCC↑	SROCC↑	RMSE↓	PLCC↑	SROCC↑	RMSE↓
NSIQA	FR	PSNR	0.7203	0.7038	9.3648	0.6587	0.6495	10.173
		SSIM [6]	0.7663	0.7423	8.8135	0.7268	0.7076	9.3989
		GMSD [7]	0.8026	0.8266	7.4417	0.744	0.7938	8.7281
	NR	BRISQUE [8]	0.649	0.6449	10.4157	0.608	0.6161	10.9226
		NIQE [9]	0.6263	0.6327	10.7493	0.5788	0.5901	11.2993
SCIQA	FR	ESIM [21]	0.8367	0.8304	7.0962	0.7993	0.793	8.0355
		SQMS [25]	0.8195	0.8140	7.3796	0.7754	0.7825	8.3285
		SVQI [22]	0.7892	0.7868	7.8339	0.7445	0.7509	8.8002
	NR	Yang's Work[20]*	0.8283	0.8142	7.3131	0.7961	0.7861	8.1322
NSVQA	FR	MOVIE [11]	0.7524	0.7183	8.2156	0.6608	0.6121	9.8896
		STMAD [12]	0.8001	0.7910	7.6226	0.7584	0.7513	8.6186
		STRRED [13]	0.8221	0.8205	7.2319	0.7767	0.7707	8.3738
	NR	VSFA [14]*	0.7741	0.7901	8.3413	0.7592	0.7754	8.7819
		VIDEVAL [16]	0.8119	0.8018	7.5238	0.7687	0.7655	8.5586
SCVQA	FR	SGFTM [28]	0.8628	0.8532	6.4740	**0.8422**	**0.8334**	**7.1077**
		MS-RSDS [29]	**0.8752**	**0.8644**	**6.3549**	0.8386	0.8261	7.3015
		HSFM [30]	**0.8975**	**0.8774**	**5.8338**	**0.8653**	**0.8467**	**6.7299**
	NR	Li's Work [31]	0.8648	0.8575	6.6190	0.8263	0.8286	7.6437
		Proposed*	**0.9166**	**0.9193**	**5.4748**	**0.9021**	**0.9114**	**5.9506**

the objective prediction and subjective assessment, higher PLCC/SROCC, or lower RMSE means better performance. Before estimating the above metrics, a nonlinear logistic regression process is performed to map prediction results as the same scale space as the subjective scores with different value domains to ensure a fair performance comparison according to the video quality experts group (VQEG) [48] as follows:

$$\hat{Q} = \beta_1 \left[\frac{1}{2} - \frac{1}{1+\exp\left[\beta_2 (Q-\beta_3)\right]}\right] + \beta_4 Q + \beta_5 \qquad (17)$$

where β_1, β_2, β_3, β_4, β_5 are the parameters to be determined.

B. Implementation Details

For the multi-task SSL multi-channel CNN model training, all input SCFs and the corresponding saliency map and edge information map were split into 512×512 image patches. The parameter β in (12) was set as 0.6 through the experiments. We trained the model for 1000 epochs with an initial learning rate of 0.0001 using an Adam optimizer. Also, (12) is used as the loss function as a multi-task learning for pre-training our CNN model.

The mean and standard deviation of all quality feature representations of frame-patches within the SCF as its SCF spatial quality feature representation was then extracted from our multi-channel CNN model to form the feature vector, $SCVQ_c$, as the input of the TCNNT model. Specifically, two transformer encoder layers were used in our TCNNT model.

We used (16) as a loss function to train the whole model with an initial learning rate of 0.0001 using the Adam optimizer.

C. Performance Evaluation on SCVQA Databases

We trained and tested our model on the two SCVQA databases individually and compared the performance with other classical and state-of-the-art SCIQA/SCVQA and NSIQA/NSVQA approaches. Eighteen IQA/VQA approaches (including FR and NR approaches): Five NSIQA methods (PSNR, SSIM [6], GMSD [7], BRISQUE [8], and NIQE [9]), four SCIQA methods (ESIM [21], SQMS [25], SVQI [22] and Yang's work [20]), five NSVQA methods (MOVIE [11], STMAD [12], STRRED [13], VSFA [14] and VIDEVAL [16]) and four SCVQA methods (SGFTM [28], MS-RSDS [29], HSFM [30] and Li's work [31]), were included. It is important to note that SGFTM, MS-RSDS, HSFM, and Li's work [31] represent the most recent methods proposed for the SCVQA task. However, SGFTM, MS-RSDS, and HSFM are FR-SCVQA methods, and their practical application is constrained, considering reference videos are not always available in real-world scenarios. Also, Li's work [31] relies on handcrafted features, concentrating solely on specific distortions and selected features, which inevitably limits the method's generalization capabilities and the accuracy of SCVQA.

The mean performance of PLCC, SROCC, and RMSE results of the above competitors and the proposed model on the SCVD and CSCVQ SCV databases are given in Table II. The direct average (the mean performance of two SCVQA databases) and weighted average (weighted mean performance based on the size of two SCVQA databases) performances are also presented as the overall performance. As we can see, our proposed model achieves almost all the best (10 out of 12) performances in terms of PLCC, SROCC, and RMSE. On the SCVD database, the correlation and accuracy of our proposed model are superior to other methods. Our proposed model achieves a remarkable performance of 0.8945 and 0.9073 in terms of PLCC and SROCC. It demonstrates superior performance, improving PLCC and SROCC by 5.4% and 9.2%, respectively, compared to the second-best model, HSFM [30], which is the best non-deep learning method. Furthermore, when compared to other deep learning-based methods on the SCVD database, our model demonstrates a clear distinction. The deep learning-based SCIQA method, Yang's work [20], achieves PLCC and SROCC scores of only 0.7794 and 0.7715. Similarly, the deep learning-based NSVQA method, VSFA [14], registers score of 0.7514 and 0.7678 in PLCC and SROCC. These results highlight that both deep learning-based SCIQA and NSVQA methods are not fully equipped to accurately measure the perceptual quality of SCVs. This shortfall is primarily due to the distinct content characteristics of SCVs, which differ significantly from NSVs and the SCVs contain additional temporal and spatiotemporal information compared to SCIs.

On the CSCVQ database, our performance is the second-best in terms of PLCC and RMSE and is on par with HSFM [30] with a slight difference of 0.0076 and 0.2797, but our proposed model obtains a higher SROCC. Also, our proposed model outperforms HSFM [30] on the SCVD database with about 0.0459, 0.0765, and 0.9977 improvements in terms of PLCC, SROCC, and RMSE. These results show that our proposed model has better generalizability. The success of our model can be attributed to its adaptability and robustness in handling diverse distortions. The CSCVQ database, which is tailored to compression distortion, yields promising results when evaluated with HSFM [30]. On the other hand, the SCVD database introduces a more complex and diverse set of distortions, offering ten different types of transmission, display, and compression challenges. Existing FR-SCVQA methods [28-30], relying on filters like 3D-Gabor and 3D Laplacian of Gaussian for edge information extraction, as well as NR-SCVQA [31] with its dependence on handcrafted features, show effectiveness within certain types of distortions. However, their ability to address a wide range of distortions may be limited. In contrast, our proposed deep learning-based NR-SCVQA model is crafted to learn from and generalize across a wide range of data and distortions, thereby exhibiting superior generalization capabilities and outperforming others in the varied distortion landscape of the SCVD. It is also worth noting that HSFM [30] is the FR method with the help of reference information when doing the quality assessment. However, our proposed model is an NR-SCVQA metric, which is more practical for assessing the perceived video quality in real applications without reference to SCV. Compared with another NR-SCVQA method [31], our proposed model improves with large margins, of approximately 10.9% in PLCC and 11.5% in SROCC on the SCVD database as well as 1.7% in PLCC and 3.3% in SROCC on the CSCVQ database, which again confirms the effectiveness of our proposed model.

TABLE III
CROSS-DATABASES AND JOINT-DATABASES EXPERIMENT.

Training on	Methods	Testing on	
		SCVD	CSCVQ
SCVD	Li's Work [31]	/	0.764
	Proposed	/	**0.837**
CSCVQ	Li's Work [31]	0.659	/
	Proposed	**0.787**	/
Combined	Li's Work [31]	0.731	0.807
	Proposed	**0.828**	**0.884**

Furthermore, in terms of overall performance, our proposed model outperforms other competitors with respect to PLCC, SROCC, and RMSE. It shows improvements of 2.1%, 4.8%, and 6.2% using a direct average, and 4.3%, 7.6%, and 11.6% using a weighted average, respectively, over the second-best model, HSFM [30], the leading non-deep learning method. Our proposed model evidently outperforms other IQA/VQA methods and exhibits the best effectiveness and generalization performance on two SCV databases. Also, it proves that our proposed deep learning-based method is more robust and effective than other non-deep learning-based SCVQA [28-31] methods.

D. Performance Evaluation on Cross-databases

To further validate the generalization capability of our proposed model, we examine its performance on the SCVD and CSCVQ databases in a cross-database validation. This section describes the outcomes when our model is trained on one

TABLE IV
PERFORMANCE VARIATIONS ON ADOPTED SUB-ALGORITHMS.

Items	*Saliency Map*		*Edge Map*		**CSCVQ**
	[45-46]	**[49]**	**Gabor**	**Sobel**	PLCC↑
1 proposed	✓		✓		0.9462
2		✓	✓		0.9397
3	✓			✓	0.9153

(a) (b)

Fig. 7. (a) Saliency map of pictorial region in Fig. 3(b) using [45-46]; (b) Saliency map of pictorial region in Fig. 3(b) using [49].

database and evaluated on another. Using the same experimental framework, we then compare its PLCC performance with the NR-SCVQA method proposed in [31]. Table III clearly shows that the generalization ability of our proposed model, when assessed in cross-database scenarios, outperforms the method in [31] in terms of PLCC. Besides, we conducted a supplementary experiment in which we randomly selected samples from both SCVQA databases for training and utilized the remaining samples for evaluation. When our model is trained on this joint database scenario, the results in Table III show a significant improvement over the competing NR-SCVQA method. Specifically, the PLCC performance of our model has 13.2% and 9.5% improvement on the SCVD and CSCVQ SCVQA databases, respectively. These results clearly demonstrate the robust generalization capability of our proposed model. Its success in achieving promising results in both cross-database and joint-database situations underscores its potential for effective real-world implementation.

E. Ablation Study of Sub-algorithms at Pre-processing State

To thoroughly assess the effectiveness of our proposed method, the performance variations of our model were investigated when employing different sub-algorithms at the pre-processing stage, including saliency map (*SM*) extraction and edge map (*EM*) extraction methods. Instead of the methods employed in this paper ([45-46] for *SM* and the Gabor filter for *EM*), we also investigated the method proposed in [49] for *SM* extraction and the Sobel filter for *EM* extraction. Table IV showcases the outcomes achieved by two *SM* extraction methods (Items 1 and 2) on the CSCVQ database. The results indicate that both methods achieve similar and satisfactory results. It is evident that both methods, effectively focusing on the relevant regions, as depicted in Fig. 7(a) and Fig. 7(b), which highlight the face and tie area. However, when it comes to *EM* extraction methods, the Gabor filter outperforms the Sobel filter (Items 1 and 3). This can be attributed to the capability of the Gabor filter to capture texture information and more complex edge features in an image compared to the Sobel filter, making it better suited to represent features in edge regions. Overall, all combinations and variations of different sub-algorithms achieve satisfactory results on the CSCVQ database, with PLCC values exceeding 0.91. This demonstrates the effectiveness of our proposed method.

TABLE V
ABLATION STUDY OF OUR PROPOSED MODEL WITH VARIOUS TASK LEARNING ON MULTI-CHANNEL CNN.

Multi-task SSL			**SCVD**		**CSCVQ**	
L_{RK}	L_{DT}	L_{DL}	PLCC↑	SROCC↑	PLCC↑	SROCC↑
✓			0.8471	0.8602	0.8878	0.8843
✓	✓		0.8957	0.9104	0.9236	0.9183
✓		✓	0.8629	0.8766	0.8925	0.9019
✓	✓	✓	**0.9136**	**0.9241**	**0.9462**	**0.9417**

F. Ablation Study of Multi-Task Learning

As mentioned in Section III, multi-task SSL allows the model to learn multi-tasks in parallel while using shared features. In other words, the features learned from each task are also used for other tasks for better feature representation learning, which can improve the learning efficiency, prediction accuracy, and generalization learning ability of the model by solving multiple learning tasks at the same time [42]. Therefore, to demonstrate the effects of multi-task learning (pairwise ranking task in (9), distortion classification task in (10), and degradation degree task in (11)) for the SCF spatial quality feature representation learning, the ablation study was performed on the multi-channel CNN model using various training settings. It contains four combinations: performing the pairwise ranking task only (L_{RK}), performing the multi-task learning by the pairwise ranking task and the distortion classification task ($L_{RK}+L_{DT}$), performing the multi-task learning by the pairwise ranking task and the degradation degree task (L_{RK}+L_{DL}), and performing the multi-task learning with all three tasks (L_{RK}+L_{DT}+L_{DL}). It is noted that we only used the training set and validation set data to perform the ablation study. Experimental results are shown in Table V.

Although the SCF spatial quality feature representation extracted from the multi-channel CNN model, which is pre-trained by the pairwise ranking task only (L_{RK}), with the TCNNT model achieves a satisfactory result, it can be seen that multi-task learning (L_{RK}+L_{DT}, L_{RK}+L_{DL}, and L_{RK}+L_{DT}+L_{DL}) can also be of great help in obtaining the optimized SCF spatial quality feature representation to predict SCV quality scores more precisely. When only L_{RK} is performed on our multi-channel CNN model, the PLCC results are 0.8471 and 0.8878 in the SCVD and CSCVQ databases, respectively. The PLCC results, when incorporating the distortion classification or degradation degree task with the pairwise ranking task ($L_{RK}+L_{DT}$ and $L_{RK}+L_{DL}$) performed on the SCVD, have improved 5.7% and 1.9%. Furthermore, performing the multi-task learning with all three tasks (L_{RK}+L_{DT}+L_{DL}) on the multi-channel CNN model achieves the best results with a 7.9% improvement. It proves that compared with task-specific or single-task learning, the multi-task SSL lets our model learn multi-tasks in parallel while using shared features to improve the feature representation learning. Therefore, by solving

TABLE VI
ABLATION STUDY OF OUR PROPOSED MODEL WITH VARIOUS FEATURE LEARNING ON MULTI-CHANNEL CNN.

	F_{PR}	F_{TR}	SCVD PLCC↑	SCVD SROCC↑	CSCVQ PLCC↑	CSCVQ SROCC↑
Group 1			0.8409	0.8437	0.8786	0.8812
Group 2	✓		0.8665	0.8742	0.9072	0.9167
Group 3		✓	0.8917	0.8893	0.9274	0.9210
Group 4 (Proposed)	✓	✓	**0.9136**	**0.9241**	**0.9462**	**0.9417**

TABLE VII
ABLATION STUDY OF SPATIOTEMPORAL FEATURES LEARNING WITH VARIOUS COMBINATION OF MODEL

	Spatial FCL	*Spatial* TCNN	*Temporal* LSTM	*Temporal* TEM	SCVD PLCC↑	SCVD SROCC↑	CSCVQ PLCC↑	CSCVQ SROCC↑
FCL+LSTM	✓		✓		0.8589	0.8731	0.8964	0.8853
FCL+TEM	✓			✓	0.8715	0.8864	0.9135	0.9207
TCNN+LSTM		✓	✓		0.9067	0.9089	0.9271	0.9186
TCNNT (proposed)		✓		✓	0.9136	0.9241	0.9462	0.9417

multiple learning tasks (distinguishing pairwise ranking L_{RK}, characterizing the distortion type L_{DT}, and identifying the degradation degree of SCF L_{DL}) simultaneously, the learning efficiency, prediction accuracy, and generalization learning ability of our new model are improved.

G. Ablation Study of Multi-Channel Learning

To assess the effectiveness of our multi-channel CNN model in learning spatial quality features of pictorial and textual regions, we conducted a series of experiments under different configurations by removing either the channel of the pictorial region (F_{PR} in (6)) or the channel of the textual region (F_{TR} in (7)) or both from our proposed model to evaluate their contributions and performance individually. In **Group 1**, we extracted spatial quality feature representations solely from the channel of I^d without implementing edge and saliency extraction, increasing its channel depth to align with the feature extraction capabilities of our proposed method. In **Group 2**, we removed the channel of the textual region, F_{TR} in (7). Conversely, in **Group 3**, we excluded the channel of the pictorial region, F_{PR} in (6), from our proposed model. Finally, **Group 4** represents the full implementation of our proposed multi-channel CNN model.

The results in Table VI reveal that solitary implementation of F_{TR} or F_{PR} (Group 2 and Group 3) can surpass the performance of the fundamental spatial quality feature learning strategy of Group 1 which does not include edge and saliency extraction. This highlights the significant advantages that extra either edge or and saliency information provides to the model's capability to identify spatial quality features with greater precision. Compared to the results of Group 2, the PLCC value of Group 3 increased from 0.8665 to 0.8917 on SCVD and from 0.9072 to 0.9274 on CSCVQ. It proves that the channel of the textual region, F_{TR}, is more effective in learning the spatial quality feature of SCF since the SCF often contains relatively sharp edges and texts, and human visual system is more sensitive to the edges. Moreover, due to the fact that SCF usually contains mixed screen content and natural content, the improvement is limited when considering either pictorial or textual regions only. After performing the multi-channel mechanism to incorporate F_{PR} and F_{TR} as in (8), our multi-channel CNN model (Group 4) can achieve the best results on SCVD and CSCVQ regarding PLCC and SROCC. It shows that our multi-channel CNN can obtain the best performance by considering both spatial quality features of the pictorial and textural parts.

H. Ablation Study of Spatiotemporal Features Learning

Moreover, we analyze the effectiveness of the high-level spatiotemporal features learning by our TCNNT model. There are four groups in the experiment: prediction of the SCV quality score using SCV quality feature representation vectors, $SCVQ_c$, by fully connected layer (FCL) integrated with LSTM without exploring the high-level spatiotemporal features and temporal attention information (**FCL+LSTM**), prediction by FCL combined with Transformer encoder model (TEM) to investigate the temporal attention information only (**FCL+TEM**), prediction by time-distributed CNN model (TCNN) incorporated with LSTM which explores the spatiotemporal features only (**TCNN+LSTM**), and prediction by TCNN incorporated with TEM as our TCNNT model to explore both high-level spatiotemporal features and temporal attention information simultaneously (**TCNNT**). The experimental results are shown in Table VII.

Although the prediction by FCL integrated with the LSTM model (FCL+LSTM) can achieve a satisfactory result, it can be seen that replacing the FCL with TCNN (TCNN+LSTM) or replacing the LSTM with TEM (FCL+TEM) can also be of great help in predicting precise video quality scores via learning either spatiotemporal features or temporal attention information. When the FCL+LSTM model is used, the PLCC results are 0.8589 and 0.8964 in the SCVD and CSCVQ SCVQA databases. The PLCC results of TCNN+LSTM and FCL+TEM on SCVD improved from 0.8589 to 0.9067 and 0.8715, respectively. However, our proposed TCNNT further boosts the performance and achieves the best results on SCVD and CSCVQ by learning the spatiotemporal features and temporal attention information together, thereby enhancing the SCV quality prediction with the help of high-level spatiotemporal features learning.

I. Runtime

We assess the runtime of our proposed model and benchmark it against four FR/NR-SCVQA models. For a fair comparison, all methods were tested on the same device running a Windows 10 platform, equipped with an Intel i9-10900K CPU, 64GB RAM, and NVIDIA GeForce RTX 3090 24GB GPU.

Table VIII shows the results of the runtime comparison. We evaluated two videos of varying resolutions (720p and 1080p)

TABLE VIII
RUNTIME OF SCVQA METHODS.

SCVQA	Methods	720P	1080P
FR	SGFTM [28]	59.7 sec	108.6 sec
	MS-RSDS [29]	41.8 sec	85.1 sec
	HSFM [30]	77.1 sec	136.5 sec
NR	Li's Work [31]	282.7 sec	409.4 sec
	Proposed	154.6 sec	213.8 sec

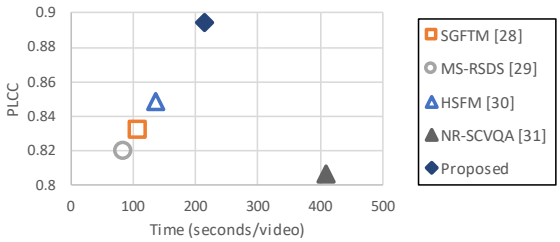

Fig. 8. The PLCC results on SCVD database (collected from Table II) against the runtime with 1080p.

from the CSCVQ and SCVD databases to determine the necessary runtime for each SCVQA method. To ensure a fair and accurate comparison, the analysis measured the time taken from the input of a raw video to the generation of the final SCV quality score. While our proposed model has longer runtime compared to FR-SCVQA methods, it provides superior prediction accuracy results, as depicted in Fig. 8. Notably, our model employs an NR approach, which can evaluate video quality without the need for an original reference video. This is particularly useful in real-world scenarios where reference videos are not always available. Broadcasters can leverage our NR-SCVQA approach to monitor the quality of transmitted content, ensuring a consistent visual experience for viewers. Similarly, streaming services can employ our NR-SCVQA method to dynamically adjust video quality in response to the intricacies of the content and fluctuating network conditions. Such an application not only optimizes bandwidth utilization but also upholds the integrity of video quality. This is critical for educational content video, news, and other SCV.

Compared to the NR-SCVQA method presented in [31], our proposed model not only surpasses it in terms of PLCC but also offers a significant computational advantage. For both 720p and 1080p resolution videos, our model achieves an approximate 45% reduction in runtime. Overall, Fig. 8 clearly demonstrates that our proposed NR-SCVQA model can better balance prediction accuracy and requested processing time, making it a superior choice for applications where both accuracy and speed are critical.

V. CONCLUSIONS

In this paper, we developed an SCF spatial quality feature representation learning through multi-channel CNN using multi-task SSL and proposed a TCNNT model further to extract the high-level spatiotemporal features of the entire sequence to assess the perceptual quality of SCV. First, we solve the limitations of the lack of available human-annotated label data for the SCVQA via the multi-channel CNN using multi-task SSL by solving the pairwise ranking, distortion classification, and degradation degree tasks. This is one of the major bottlenecks that we break through in VQA. Furthermore, the multi-channel CNN can learn the spatial quality feature of the pictorial and textural parts separately and then concatenate those features to learn the optimized SCF spatial quality feature representation. Finally, the TCNNT model is proposed to further process all SCF spatial quality feature representation in an SCV to explore high-level spatiotemporal features by jointly learning spatial and temporal features, thereby providing the optimized spatiotemporal features to obtain the final precise predicted SCV quality for SCVQA. Experimental results demonstrate the robustness and effectiveness of our proposed model, which outperforms other handcrafted features-based FR/NR-SCVQA methods. It proves that our NR-SCVQA model, the first deep learning-based SCVQA method, compensates for the shortcomings of handcrafted feature-based methods that improve the performance and generalization of perceptual quality evaluation for SCV. However, our proposed model currently may not operate in real-time due to its runtime requirements. Looking ahead, a lightweight NR-SCVQA model should be developed by exploring the adoption of temporal downsampling or spatial-temporal sampling techniques. These strategies have the potential to significantly reduce computational complexity and requested processing time. By integrating such methods into our model, we can increase its feasibility for real-time applications, broadening its practical utility for broadcasting.

REFERENCES

[1] Min, X., Gu, K., Zhai, G., Yang, X., Zhang, W., Le Callet, P., and Chen, C. W., "Screen content quality assessment: Overview, benchmark, and beyond," in *Proc. of the ACM Comput. Surv*. 2021, pp. 1-36.

[2] X. Xu and S. Liu, "Overview of screen content coding in recently developed video coding standards," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 839-852, Feb. 2022.

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

[4] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, pp. 1–52, Apr. 2020.

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

[6] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Process.*, vol. 13, no. 4, pp. 600-612, April 2004.

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

[8] 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, 2012.

[9] 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, 2013.

[10] X. Liu, J. Van De Weijer and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. of IEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 1040-1049.

[11] 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, Oct. 2010.

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

[13] 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, Aug. 2013.

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

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

[16] 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. on Image Process.*, vol. 30, pp. 4449-4464, 2021.

[17] S. Jiang, Q. Sang, Z. Hu and L. Liu, "Self-supervised representation learning for video quality assessment," *IEEE Trans. on Broadcasting*, vol. 69, no. 1, pp. 118-129, March 2023.

[18] W. Shen, M. Zhou, X. Liao, W. Jia, T. Xiang, B. Fang, and Z. Shang, "An end-to-end no-reference video quality assessment method with hierarchical spatiotemporal feature representation," *IEEE Trans. on Broadcasting*, vol. 68, no. 3, pp. 651-660, Sept. 2022.

[19] Sik-Ho Tsang, Yui-Lam Chan, and Wei Kuang, "Mode Skipping for HEVC Screen Content Coding via Random Forest," *IEEE Trans. on Multimedia*, vol. 21, no. 10, pp. 2433-2446, Oct. 2019.

[20] J. Yang, Y. Zhao, J. Liu, B. Jiang, Q. Meng, W. Lu, and X. Gao, "No reference quality assessment for screen content images using stacked autoencoders in pictorial and textual regions," *IEEE Trans Cybernetics*, vol. 52, no. 5, pp. 2798-2810, May 2022.

[21] Z. Ni, L. Ma, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "ESIM: Edge similarity for screen content image quality assessment," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4818–4831, Oct. 2017.

[22] K. Gu, J. Qiao, X. Min, G. Yue, W. Lin, and D. Thalmann, "Evaluating quality of screen content images via structural variation analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 10, pp. 2689–2701, Oct. 2018.

[23] Y. Fu, H. Zeng, L. Ma, Z. Ni, J. Zhu and K. -K. Ma, "Screen content image quality assessment using multi-scale difference of gaussian," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2428-2432, Sept. 2018.

[24] C. Zhang, Z. Huang, S. Liu and J. Xiao, "Dual-channel multi-task CNN for no-reference screen content image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5011-5025, Aug. 2022.

[25] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

[26] Y. Fang, R. Du, Y. Zuo, W. Wen and L. Li, "Perceptual quality assessment for screen content images by spatial continuity," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4050-4063, Nov. 2020.

[27] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging HEVC screen content coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 50–62, Jan. 2016.

[28] S. Cheng, H. Zeng, J. Chen, J. Hou, J. Zhu and K. -K. Ma, "Screen content video quality assessment: Subjective and objective study," *IEEE Trans. Image Process.*, vol. 29, pp. 8636-8651, 2020.

[29] T. Li, X. Min, H. Zhao, G. Zhai, Y. Xu and W. Zhang, "Subjective and objective quality assessment of compressed screen content videos," *IEEE Trans. on Broadcasting*, vol. 67, no. 2, pp. 438-449, June 2021.

[30] H. Zeng, H. Huang, J. Hou, J. Cao, Y. Wang and K. -K. Ma, "Screen content video quality assessment model using hybrid spatiotemporal features," *IEEE Trans. Image Process.*, vol. 31, pp. 6175-6187, 2022.

[31] T. Li, X. Min, W. Zhu, Y. Xu, and W. Zhang, "No-reference screen content video quality assessment," *Displays*, vol. 69, 102030, 2021.

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

[33] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of the IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[34] A. Vaswani, N. Shazeer, N. Parmar, and J. Uszkoreit, "Attention is all you need," in *Proc. Advances in Neural Inf. Process. Syst.*, 2017.

[35] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

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

[37] 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.

[38] Y. Liu, M. Jin, S. Pan, C. Zhou, Y. Zheng, F. Xia, and P. S. Yu, "Graph self-supervised learning: A survey," *IEEE Trans on Knowledge and Data Engineering*, 2022.

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

[40] M. Noroozi, and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles", in *Proc. European conf. on computer vision (ECCV)*, 2016, pp. 69-84.

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

[42] Crawshaw and Michael. "Multi-task learning with deep neural networks: A survey." arXiv preprint arXiv:2009.09796, 2020.

[43] M. P. Viana and D. A. B. Oliveira, "Fast CNN-based document layout analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*,Venice, Italy, 2017, pp. 1173–1180.

[44] K. Y. Wong, R. G. Casey and F. M. Wahl, "Document analysis system," *IBM Journal of Research and Development*, vol. 26, no. 6, pp. 647-656, Nov. 1982.

[45] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*. (CVPR), 2007, pp. 1-8.

[46] S. Montabone and A. Soto." Human detection using a mobile platform and novel features derived from a visual saliency mechanism," *Image and Vision Computing*, vol. 28(3), pp. 391-402, 2010.

[47] W. Jing, Y. Bai, Z. Zhu, R. Zhang, and Y. Jin, "Dual-anchor metric learning for blind image quality assessment of screen content images," *Electronics*, vol. 11(16), 2510, 2022.

[48] Video Quality Experts Group, "Final report from the video quality experts group on the validation of objective quailty metrics for video quality assessmen," Accessed on: April, 2021, [Online]. Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx.

[49] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 FPS," in *Proc. 2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1404-1412.

Ngai-Wing Kwong received the B.Eng. (Hons.) from The Hong Kong Polytechnic University, Hong Kong, in 2018. He is currently pursuing the Ph.D. degree in the Digital Signal Processing Laboratory, The Hong Kong Polytechnic University, Hong Kong. His current research interests include deep learning, and image and video processing.

Yui-Lam Chan (S'94–A'97–M'00) received the B.Eng. (Hons.) and Ph.D. degrees from The Hong Kong Polytechnic University, Hong Kong, in 1993 and 1997, respectively. He joined The Hong Kong Polytechnic University in 1997, where he is currently an Associate Professor with the Department of Electrical and Electronic Engineering. He is actively involved in professional activities. He has authored over 140 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. Dr. Chan served as an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the Secretary of the 2010 IEEE International Conference on Image Processing. He was also the Special and Demo Sessions Co-Chairs, IEEE International Conference on Visual Communications and Image Processing, the Publications Chairs of the IEEE International Conference on Multimedia and Expo.

Sik-Ho Tsang received the Ph.D. degree from The Hong Kong Polytechnic University (PolyU), Hong Kong, in 2013. His research interests involve computer vision (CV), natural language processing (NLP), and acoustic signal processing using artificial intelligence (AI) and 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.

Ziyin Huang received M.Sc. from The Guangdong University of Technology, China, in 2020. She is currently pursuing the Ph.D. degree in the Digital Signal Processing Laboratory, The Hong Kong Polytechnic University, Hong Kong. Her current research interests include deep learning, and video enhancement.

Kin-Man Lam received his Associateship in Electronic Engineering with distinction from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic) in 1986, his M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College, U.K., in 1987, and his Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Australia, in 1996. From 1990 to 1993, he was a lecturer at the Department of Electronic Engineering of The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University again as an Assistant Professor in October 1996. He became an Associate Professor in 1999, and has been a Professor since 2010. He was actively involved in professional activities. He has been a member of the organizing committee or program committee of many international conferences. Prof. Lam was the Chairman of the IEEE Hong Kong Chapter of Signal Processing between 2006 and 2008, and was the Director-Student Services and the Director-Membership Services of the IEEE SPS between 2012 and 2014, and between 2015 and 2017, respectively. He was also the VP-Member Relations and Development and VP-Publications of the Asia-Pacific Signal and Information Processing Association (APSIPA) between 2014 and 2017, and between 2017 and 2021, respectively. He was an Associate Editor of IEEE Trans. on Image Processing between 2009 and 2014, and Digital Signal Processing between 2014 and 2018. He was also an Editor of HKIE Transactions between 2013 and 2018, and an Area Editor of the IEEE Signal Processing Magazine between 2015 and 2017. Currently, he is the Member-at-Large of APSIPA and the IEEE SPS VP-Membership. Prof. Lam also serves as a Senior Editorial Board member of APSIPA Trans. on Signal and Information Processing, and an Associate editor of EURASIP International Journal on Image and Video Processing. His current research interests include image and video processing, computer vision, and human face analysis and recognition.