Sparse Representation-Based Video Quality Assessment for Synthesized 3D Videos

Yun Zhang, *Senior Member, IEEE*, Huan Zhang, Mei Yu, Sam Kwong, *Fellow, IEEE*, and Yo-Sung Ho, *Fellow, IEEE*

***Abstract*— The temporal flicker distortion is one of the most annoying noises in synthesized virtual view videos when they are rendered by compressed multi-view video plus depth in Three Dimensional (3D) video system. To assess the synthesized view video quality and further optimize the compression techniques in 3D video system, objective video quality assessment which can accurately measure the flicker distortion is highly needed. In this paper, we propose a full reference sparse representation-based video quality assessment method toward synthesized 3D videos. First, a synthesized video, treated as a 3D volume data with spatial (X-Y) and temporal (T) domains, is reformed and decomposed as a number of spatially neighboring temporal layers, i.e., X-T or Y-T planes. Gradient features in temporal layers of the synthesized video and strong edges of depth maps are used as key features in detecting the location of flicker distortions. Second, the dictionary learning and sparse representation for the temporal layers are then derived and applied to effectively represent the temporal flicker distortion. Third, a rank pooling method is used to pool all the temporal layer scores and obtain the score for the flicker distortion. Finally, the temporal flicker distortion measurement is combined with the conventional spatial distortion measurement to assess the quality of synthesized 3D videos. Experimental results on synthesized video quality database demonstrate our proposed method is significantly superior to the other state-of-the-art methods, especially on the view synthesis distortions induced from depth videos.**

***Index Terms*— Video quality assessment, synthesized view, sparse representation, flicker distortion, temporal layer.**

I. Introduction

FREE Viewpoint Video (FVV) and Three Dimensional (3D) video systems are capable of generating arbitrary Virtual Viewpoint Images (VVIs) and providing users with more realistic visual enjoyments including interactive arbitrary viewing and 3D depth perception. They have promising prospects in many applications, such as 3D film and television, virtual reality game, distant surgery and education, and will be popular in many aspects of people's life in near future. Since the perceptual quality of VVIs is directly related to users' Quality of Experience (QoE) of 3D or FVV systems, objective VVI quality metrics are highly desired to quantify and predict the quality of videos automatically, which can be then applied to optimize many visual processing techniques, such as image/video compression, digital watermarking, and image/video reconstruction in 3D video system. Therefore, Image or Video Quality Assessment (IQA/VQA) for 3D videos has become an important issue in visual perception and visual signal processing.

The conventional 2D fidelity IQA metrics, such as the well-known Mean Squared Error (MSE) or Peak Signal-to-Noise Ratio (PSNR), are widely used in measuring the quality of image and video in various applications due to their simplicity. However, they are pixel-wise difference between distorted and source images, which can hardly reflect the real perceptual quality of visual signal and are inconsistent with human perception. To tackle this problem, many objective quality metrics have been proposed to simulate the perception mechanism of the Human Visual System (HVS) and evaluate the perceptual quality. Structural SIMilarity (SSIM) [1] and Multi-Scale SSIM (MS-SSIM) [2] were proposed to measure the inter-dependencies among spatial pixels, *e.g.*, image structural information [3], by considering the luminance masking and contrast masking effects. Videos consist of a set of successive 2D images with 25 frames per second or higher, become more complicated by adding the temporal dimension. Visual Quality Model (VQM) [4] and MOtion based Video Integrity Evaluation index (MOVIE) [5] are two typical VQA

Manuscript received December 21, 2018; revised June 17, 2019; accepted June 30, 2019. Date of publication July 29, 2019; date of current version September 23, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61871372, Grant 61671258, and Grant 61672443, in part by the Guangdong NSF for Distinguished Young Scholar under Grant 2016A030306022, in part by the Key Project for Guangdong Provincial Science and Technology Development under Grant 2017B010110014, in part by the Shenzhen International Collaborative Research Project under Grant GJHZ20170314155404913, in part by the Shenzhen Science and Technology Program under Grant JCYJ20170811160212033 and Grant JCYJ20180507183823045, in part by the RGC General Research Fund (GRF) 9042322, 9042489 (CityU 11200116, 11206317), in part by the Guangdong International Science and Technology Cooperative Research Project under Grant 2018A050506063, and in part by the Membership of Youth Innovation Promotion Association, Chinese Academy of Sciences, under Grant 2018392. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel L. Lau. *(Corresponding author: Yun Zhang.)*

Y. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: yun.zhang@siat.ac.cn).

H. Zhang is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: huan.zhang@siat.ac.cn).

M. Yu is with the Faculty of Information and Engineering, Ningbo University, Ningbo 315211, China (e-mail: yumei2@126.com).

S. Kwong is with the Department of Computer Science, City University of Hong Kong, Hong Kong, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 5180057, China (e-mail: cssamk@cityu.edu.hk).

Y.-S. Ho is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: hoyo@gist.ac.kr).

Digital Object Identifier 10.1109/TIP.2019.2929433

methods that incorporate the spatial and temporal motion information. However, these are conventional 2D image/video quality metrics, which are on the basis that the distorted 2D images or videos are spatially aligned with the reference or original images/videos. The VVIs are intermediate images rendered from the color and depth videos of neighboring views by using the Depth Image Based Rendering (DIBR) [6] technology. The quality of VVIs is actually affected by the rendering distortions and the distortions in color and depth videos of neighboring reference views, which causes geometrical misalignments. Bosc *et al* [7], [8] also pointed out that conventional metrics are not suitable in predicting the quality of the synthesized views.

To solve this problem, many efforts have been devoted to developing IQA metrics for 3D synthesized views [9], [10], [12]–[14]. Battisti *et al*. [9] proposed a 3D Synthesized view Image quality Metric (3DSwIM) in which a block-based registration was performed between synthesized view image and original image, and a discrete wavelet transform based method was then utilized to detect the disoccluded areas. To alleviate the interference of the imperceptible geometrical distortion, Li *et al*. [10] proposed a method measuring LOcal geometric and Global Sharpness (LOGS) which employed the SIFT-based warping method before the disoccluded regions detection. Yue *et al*. [11] also proposed a synthesized IQA metric combining global measures for sharpness and local measures for geometrical distortions, such as disoccluded regions and stretching. Stanković *et al*. [12]–[14] proposed a Morphological Pyramid PSNR (MP-PSNR) based on morphological pyramid decomposition and Morphological Wavelet PSNR metric (MW-PSNR) based on morphological wavelet decomposition, both of which used multi-scale image decomposition and morphological filters to detect the disoccluded areas. Gu *et al*. [15] proposed a non-reference quality metrics for synthesized images by integrating the outcomes of autoregressive predictor and visual saliency and achieved superior performance. Furthermore, multiscale natural scene statistics was also exploited for blindly assessing the quality of the synthesized images [16]. In addition, Jakhetiya *et al*. [17] modeled geometrical distortions in synthesized images as outliers and conducted median filtering based outlier detection to measure the synthesized image quality. Moreover, a new prediction model was built by considering more perceptual factors of texts in screen content images [18], which was also good in measuring geometrical distortions in synthesized images. These methods tried to reduce the impacts from the imperceptible geometrical distortion and capture the strong perceptible geometrical distortion in the disoccluded areas and other forms of distortion, such as object shifting, and stretching. As a comparison, they were superior to the conventional metrics and performed well in measuring the quality of 3D synthesized images.

Since the distortion prones to take place along the vicinity of the strong edges in the depth map, the depth map could be exploited to help establish the quality assessment model of the synthesized images/videos. There were some works that used depth maps to evaluate the quality of the VVIs [19], [20]. Ekmekcioglu *et al*. [19] proposed a perceptual quality assessment method which utilized depth map to divide the VVIs into the foreground and background region and give more weights on foreground region in evaluation. Farid *et al*. [20] proposed a 3D synthesized IQA model in which the depth image distortion metric was combined with a texture distortion metric to jointly determine the quality of the synthesized image without the reference VVI. Also, some works employed depth maps to assess the quality of 3D videos [21], [22]. Liu *et al*. [21] proposed a depth-image based objective quality assessment model to evaluate the stereo image pairs of synthesized videos, in which depth features were extracted to highlight those salient regions in calculation of the quality scores. Solh *et al*. [22] proposed a synthesized VQA method, in which a high quality depth map was first obtained and the distortion metric was calculated based on this depth map, then the combined metric was used to assess the 3D video quality. However, in the 3D synthesized videos, due to the random activity of geometrical distortion caused by depth errors along the edges among temporal successive frames, there exists annoying temporal flicker distortion, which has not been well considered in these schemes.

In Multiview Video plus Depth (MVD) video system the VVI is mainly synthesized by two or more views, in addition to the spatial distortions, such as geometrical distortion and contour artifacts, the flicker distortion has become one of the most annoying distortions in the synthesized 3D videos. Methods [23]–[25] had taken the flickering into consideration in synthesized VQA. Zhao *et al*. [23] developed a temporal distortion measurement for the 3D synthesized videos, in which the Just Noticeable Difference (JND) in temporal domain was considered. However, this method only considered the flicker distortion in the static regions and overestimated the spatial distortion. Zhou *et al*. [24] proposed a non-reference metric measuring the flicker distortion, in which the gradient differences in neighboring frames were exploited to detect the flicker distortion areas and Singular Value Decomposition (SVD) vector values were utilized to measure the flickering strengths. However, this method only assessed the synthesized distortions induced by the DIBR algorithms and the compression distortions from the color or depth videos have not been considered. In addition, Shao *et al*. [25] proposed a non-reference view synthesis prediction model to predict the perceptual quality of the stereoscopic videos, in which the influences from color and depth distortions and their interactions were analyzed. The view synthesis process and hole filling were actually not involved in this method, which may lead to inaccuracy. Recently, Liu *et al*. [26] established a subjective database comprised of multiple synthesized videos from compressed texture/depth videos called SIAT database, and proposed an effective objective metric, which could measure both the flicker distortion and spatio-temporal activity distortion existed in synthesized videos. Based on [26], Zhang *et al*. [27] proposed a low complexity Synthesized Video Quality Metric (SVQM) and developed rate-distortion optimization algorithm based on the SVQM which was then integrated into the 3D video coding framework in 3D system. The SVQM incorporated the spatial and temporal distortions in VVIs, and had a trade-off between the prediction accuracy

and computational complexity for 3D video coding. The VQA for 3D synthesized videos still needs further investigations to improve the accuracy and adaptability to 3D applications.

Nowadays, due to the outstanding capability of representing the receptive field of neurons in the V1 region of the main visual cortex, sparse representation has been successively applied into some objective visual quality assessments, which can be categorized as statistical features and local features based schemes. In terms of the first category, He *et al*. [28] proposed a sparse representation based method as a blind image quality assessment model in which the natural scene statistics features were represented by sparse coding, and the sparse coefficients of the test images were employed as the weighting terms of the DMOS of the corresponding training images, and thus the final quality was obtained. Shabeer *et al*. [29] constructed a spatial-temporal dictionary and exploited the difference of the statistical characteristics of sparse coefficients induced by the distorted videos, and proposed a non-reference VQA model to predict the video quality. With regards to the second type, Zhang *et al*. [30] first proposed a phase and amplitude distortion based image quality model through the independent subspace analysis mimicking the features of the visual neurons. Shao *et al*. [31] further extended the phase and amplitude distortion metric via sparse representation in stereoscopic IQA. Based on the free energy principle, Liu *et al*. [32] proposed a reduced-reference IQA model, in which an internal generative model was approximated by sparse representation, and the entropy of the prediction was used to predict the image quality. Jiang *et al*. [33] proposed an IQA for retargeted images in which the feature extraction was based on the sparse representation of SIFT-point features and salient features. Sparse representation has been successfully applied to conventional IQAs, however, it has scarcely been applied into the VQA measuring the synthesized videos, where the flicker distortion mainly occurs in the temporal domain.

In this paper, we propose a full-reference Sparse Representation based 3D Video Quality Assessment (SR-3DVQA) for synthesized videos, in which dictionary learning for temporal layers is specifically learned and sparse representation is then employed to represent the flicker distortion over the temporal layers. The main contributions of this work are summarized as three-fold: (1) Synthesized video is decomposed as a number of spatially neighboring temporal layers. Gradient features in temporal layers of the synthesized video and strong edges of depth map are used as key features in detecting the location of flicker distortions. (2) A temporal flicker distortion measurement is proposed, where novel dictionary learning and sparse representation for the temporal layers are then derived and applied to effectively represent the temporal flicker distortion. (3) Temporal flicker distortion measurement is combined with the conventional spatial distortion measurements to assess the quality of synthesized 3D videos. Pooling methods and key factors are analyzed. The remainder of this paper is organized as follows. Motivations and analyses are presented in Section II. The Sparse Representation based Flicker Distortion Measurement (SR-FDM) is proposed in Section III. The proposed SR-3DVQA model is then presented in Section IV. Afterwards, experimental results and analyses are demonstrated in Section V. Finally, the conclusions are drawn in Section VI.

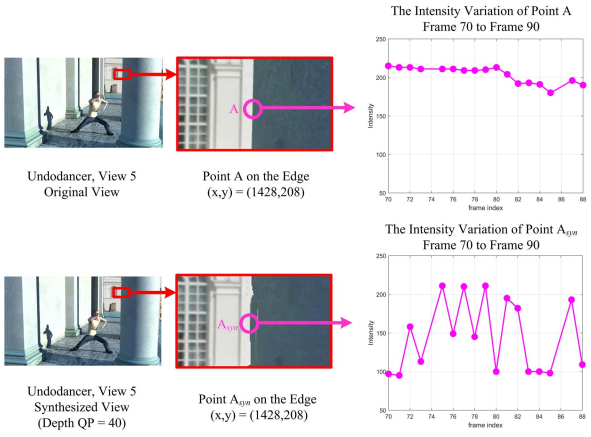


Fig. 1. The flicker distortion in synthesized video.

II. Motivations

For a synthesized view generated by DIBR technology in MVD based 3D video system, the process usually consists of two steps. The first step is warping pixels from existing views to another new virtual view via DIBR technology. Due to dis-occlusion and rounding operations in warping, small holes may exist in the rendered image. The second step is hole filling and post-processing, namely inpainting the disoccluded areas or holes from their surrounding pixels. However, due to imperfect depth images and misalignment between color and depth image, some distortions, such as contour artifacts and geometrical distortions, may occur during the DIBR based rendering process. In addition, the compression distortion in color images will be transferred to the rendered images, while the depth distortions will induce the displacement of pixels, *i.e.*, geometrical distortion [34], [35]. Furthermore, the inconsistency between temporally successive depth images caused by depth image generation and compression will induce the inconsistent geometrical distortions among frames, which is the annoying temporal flicker distortion [26], [27]. Fig. 1 illustrates an example of the flicker distortion along the rim of a pillar in the synthesized video with depth compression distortion of Undodancer. It can be observed that the intensity of point A in the original view changes little and smoothly in twenty frames. In contrast, the intensity of corresponding point point A_{syn} in synthesized view fluctuates drastically in temporal dimension, which is unnatural and annoying to the viewers. However, conventional 2D and 3D VQA metrics have not well considered the properties of the synthesized view videos and the flicker distortion. Therefore, a more effective algorithm is needed for the synthesized videos.

Traditional IQA/VQA methods extract the hand-crafted features and fuse the features to predict the quality of the distorted images/videos, *e.g.*, MSSIM [2], VQM [4], and Liu's scheme [26]. In this situation, highly professional background knowledge is required and only very limited features can be

TABLE I
DEFINITIONS OF KEY SYMBOLS OR VARIABLES

Variables	Description
$\mathbf{L}_i$	The i-th temporal layer of the video $\mathbf{V}$
$\{\mathbf{L}_i \vert 1 \leqslant i \leqslant H\}$	Temporal layers set denotation of video $\mathbf{V}$
$\{\mathbf{L}_{o,i}\}, \{\mathbf{L}_{d,i}\}$	The original and distorted video $\mathbf{V}_o$ and $\mathbf{V}_d$ denoted by temporal layers set, respectively
$V(x, y, t)$	The pixel value (x, y, t) in video $\mathbf{V}$
$V^g(x, y, t)$	The gradient value of pixel (x, y, t) in video $\mathbf{V}$
$\mathbf{G}_i$	The i-th gradient temporal layer of the video $\mathbf{V}$
$\{\mathbf{G}_{o,i}\}, \{\mathbf{G}_{d,i}\}$	The original and distorted gradient video denoted by temporal layers set, respectively
$\mathbf{S}_{V,i}$	The effective patch index set in the i-th gradient temporal layer of video $\mathbf{V}_o$ or $\mathbf{V}_d$
$D(x, y, t)$, $D^{edge}(x, y, t)$, $D^{edge'}(x, y, t)$	The pixel, edge, dilated edge pixel value of (x, y, t) of depth video $\mathbf{D}$, respectively
$\mathbf{E}_{o,k}, \mathbf{E}'_{o,k}$	The k-th edge frame, dilated edge frame of depth video $\mathbf{D}$
$\mathbf{M}_{o,i}$	The i-th edge temporal layer in depth video $\mathbf{D}$
$\mathbf{S}_{E,i}$	The effective patch index set in the i-th edge temporal layer of depth video $\mathbf{D}$
$\mathbf{S}_i$	The flicker distortion area patch index set in the i-th temporal layer of video $\mathbf{V}_o$ or $\mathbf{V}_d$
$\mathbf{y}^{o,i}_{u,v}, \mathbf{y}^{d,i}_{u,v}$	The patch at (u, v) in the i-th temporal layer of video $\mathbf{V}_o$ and $\mathbf{V}_d$, respectively
$\mathbf{U}_k$	The k-th new temporal layer of the video $\mathbf{V}$
$\mathbf{W}$, $\overline{\mathbf{W}}$	The weight map of the original, and the new temporal layer of video, respectively

developed. To further improve the representation ability and discover hidden knowledge in video data, learning effective features from data will be helpful to VQA if it could represent the visual signals more compact and discriminative. Sparse representation is a representative feature learning based method, the goal of which is to learn an over-complete dictionary with sparsity constraints, mimicking the characteristics of V1 neurons in human visual processing pathway. Therefore, we intend to use sparse representation to represent the temporal flickering and then evaluate the 3D video quality.

Conventional sparse representation and dictionary learning based on 2D spatial patches were used to represent the 2D image features. Currently, 3D dictionary learning, which further includes the temporal or depth dimension, has been employed in several applications, such as online sequence denoising [36], video super-resolution [37], and human action identification [38]. The 3D dictionary learning objective function can be formulated as [36]–[38]

$$\min_{\boldsymbol{\Psi}^{3D}, \boldsymbol{\alpha}^{3D}} \|\mathbf{X}^{3D} - \boldsymbol{\Psi}^{3D}\boldsymbol{\alpha}^{3D}\|_2^2 + \mu\|\boldsymbol{\alpha}_j^{3D}\|_1 + \lambda\boldsymbol{\rho}(\mathbf{X}^{3D}), \quad (1)$$

where $\mathbf{X}^{3D} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n]$ denotes the 3D training patches set from video data. $\mathbf{x}_j \in \mathbf{R}^3$ is a 3D volumetric data, consisting of horizontal x, vertical y, and temporal t dimensions, which we denoted as 3D-XYT. $\boldsymbol{\Psi}^{3D}$ is the learned 3D dictionary. $\boldsymbol{\alpha}^{3D}$ is a simplified form of the sparse coefficients of the overall 3D patches. $\|.\|_1$ means the $\mathbf{L}_1$ norm of the vector. $\|\boldsymbol{\alpha}_j^{3D}\|_1$ denotes that the sparse coefficient $\boldsymbol{\alpha}_j^{3D}$ of patch j should satisfy the sparsity constraint, and μ regulates the sparsity. $\boldsymbol{\rho}(\mathbf{X}^{3D})$ represents the task-driven constraint, and λ regulates the weight of this term. The advantage of 3D sparse representation is learning 3D dictionaries for better representation ability. However, the computational complexity of learning 3D dictionaries increases dramatically. One alternative solution is degrading 3D dictionary learning and approximating it with multi-layer 2D dictionaries. In fact, 2D sparse representation for 2D images (*i.e.*, 2D-XY data) can be regarded as a special case or degradation of 3D sparse representation for 3D data (*i.e.*, 3D-XYT data) by fixing the temporal dimension. Intuitively, to represent the flicker distortion in the temporal domain of synthesized video, we attempt to keep the temporal dimension and fix either X or Y in the 3D sparse representation, *i.e.*, 2D-XT or 2D-YT plane. Then, the sparse representation is customized to represent the temporal flickering features for the synthesized video, *i.e.*, SR-FDM presented in detail in the next section.

III. SPARSE REPRESENTATION BASED FLICKER DISTORTION MEASUREMENT (SR-FDM)

The SR-FDM framework consists of five main modules, *i.e.*, temporal layer conversion, gradient feature extraction, flicker distortion detection, sparse representation for flicker distortion features, and weighted layer pooling. Fig. 2 demonstrates the flowchart of the flicker distortion assessment. First, the original video and the distorted synthesized video are converted to the temporal layers via temporal layer conversion (subsection III.A), respectively. Then, each layer would be transformed to gradient feature map after gradient feature extraction in subsection III.B. And then, for each layer, the location of possible flicker distortion is identified through flicker distortion detection module with the assistance of the depth video, as illustrated in subsection III.C. Subsequently, flicker distortion strengths are measured through the sparse coefficients features at the identified flicker distortion location, and the sparse representation is based on the learned temporary dictionary through training stage in subsection III.D. Finally, the overall flicker distortion score is obtained by weighted layer pooling.

A. Temporal Layer Conversion

Generally, a video can be regarded as 3D volumetric data $\mathbf{V} = \{V(x, y, t) \vert 1 \leqslant x \leqslant W, 1 \leqslant y \leqslant H, 1 \leqslant t \leqslant T\}$ where H, W, and T represent video height (Y), width (X), and frame length (T), respectively. If we divide the 3D-XYT video into multiple 2D-XT layers, the video could be redefined as $\mathbf{V} = \{\mathbf{L}_i \vert 1 \leqslant i \leqslant H\}$, where $\mathbf{L}_i = \{V(x, y, t) \vert 1 \leqslant x \leqslant W, y = i, 1 \leqslant t \leqslant T\}$ is the i-th temporal layer, *i.e.*, 2D-XT planes, and the height H is also the number of temporal layers. For VQA of synthesized video, there are three main advantages of segmenting a video into temporal layers: (1) visuality: it helps visualize the temporal features and can provide explicit and intuitional cues; (2) capability: the temporal layer picture, a surface formed by space lines varying with time, can be used to present the long-term temporal features of the video;

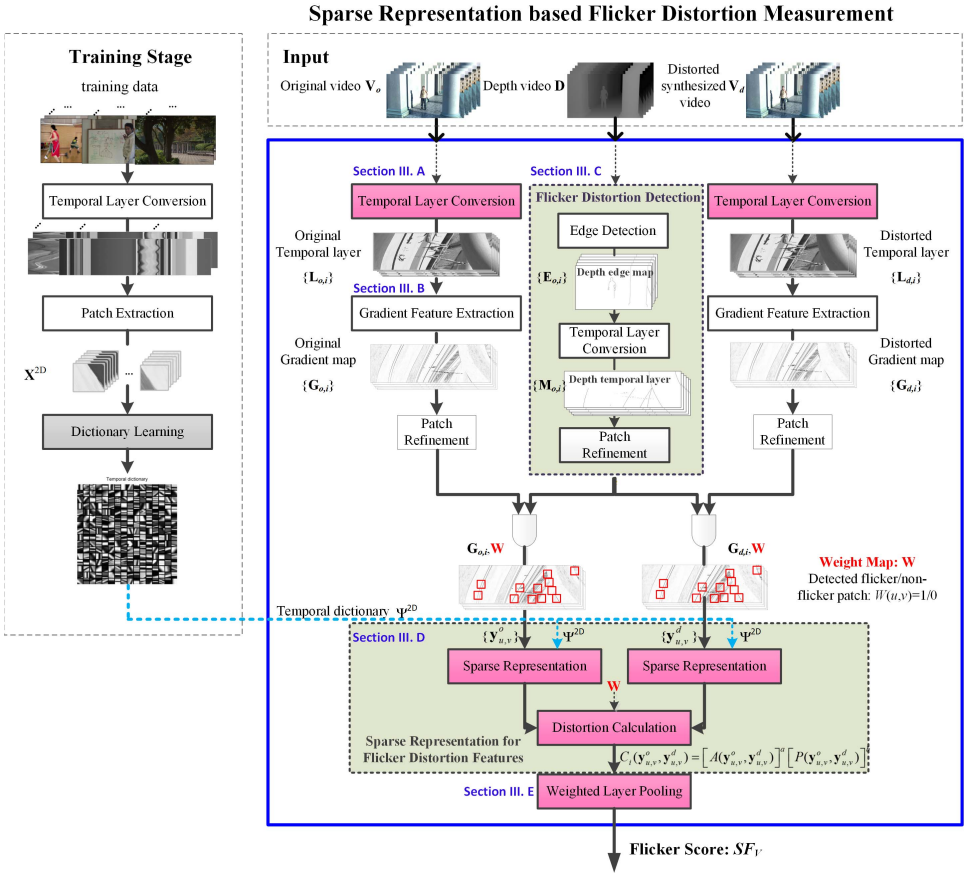


Fig. 2. The flowchart of the SR-FDM.

(2) simplicity: it avoids employing motion estimation method to match the patches between t-1-th frame and t-th frame to capture the motion features.

Therefore, to assess the flicker distortion more effectively, we convert the distorted 3D synthesized video $\mathbf{V}$ into temporal layers $\{\mathbf{L}_i\}$, *i.e.*, 2D-XT planes in this paper, as shown in the left part of the Fig. 3. We can observe that the intense movement of the human and the variable motion of one point along the rim of the pillar represent as a drastically twisted stripe and a smooth curve line, respectively, which denotes the temporal layer can capture the temporal features. In addition, the distortion in the patches with flicker distortion is obvious, *e.g.*, the crumbling and disorderly edges, while the non-flicker patch has clear edges. This phenomenon implies that the flicker distortion could be captured in the temporal layer. Thus, the original view $\mathbf{V}_o$ and the distorted synthesized view $\mathbf{V}_d$ are converted to sets of temporal layers $\{\mathbf{L}_{o,i}\}$ and $\{\mathbf{L}_{d,i}\}$, respectively.

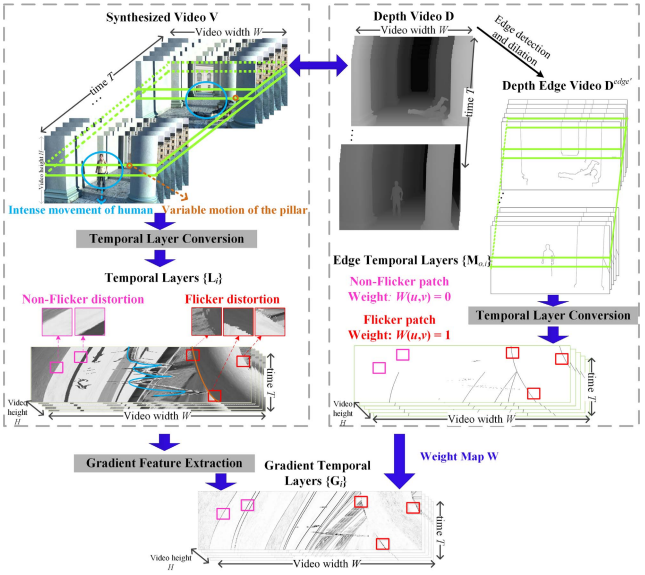


Fig. 3. The visual effects of flicker distortion in temporal layer.

B. Gradient Feature Extraction

The gradient features are more suitable to extract the flicker distortion as compared with the pixel intensity itself. The reasons are two folds: one is that human eyes are sensitive to the change rate of the intensity which leads to the significant change in the gradient; the other is that the flicker distortion caused by temporal inconsistency of depth map usually locates at edges or regions with gradient. Similarly, there are also some VQA methods utilizing the gradient features of the synthesized view [24], [26]. Note that we use the vertical gradient features of the temporal layers to capture the flicker distortion to avoid the interference of the static situation. For a

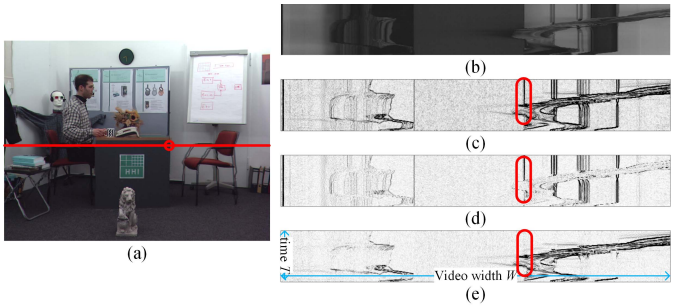


Fig. 4. The comparison of gradient maps with and without horizontal gradients in temporal layer. (a) is the first frame of Bookarrival synthesized with original texture videos and depth videos of QP 44; (b) is the temporal layer generated by the red line in (a); (c), (d), and (e) are the corresponding gradient feature maps of (a), where (c) is the gradient map with both horizontal and vertical gradients, (d) has only the horizontal gradients, and (e) has only the vertical gradients. The red rectangles in (c), (d), and (e) correspond to the feature areas generated by the red circle in (a).

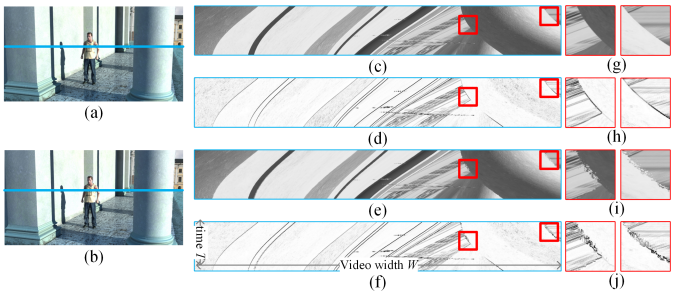


Fig. 5. The gradient map comparison in the gradient temporal layers of sequence Undodancer (a) is the first frame of original video; (b) is the first frame of the synthesized video from distorted MVD data; (c) and (e) are the temporal layers generated by the blue line in (a) and (b), respectively; (d) and (f) are the corresponding gradient feature maps of (c) and (e), respectively. (g)–(j) are zoomed subfigures of the marked red rectangles in (c)–(e).

static object, an arbitrary point on its boundaries as time varies would be a vertical line, which would result in large horizontal gradient in temporal layer. Fig. 4 is an example of gradient maps with and without horizontal gradients in temporal layer of static objects. It can be observed that the gradient map with horizontal and vertical gradients in Fig. 4(c) and the gradient map with horizontal gradients in Fig. 4(d) have strong strengths in rectangle areas while the gradient map with vertical gradients in Fig. 4(e) has very weak features. As shown in Fig. 4, the static objects can generate strong horizontal gradients, which reflects no meaningful motion information, thus affecting the true flicker distortion detection. In addition, if there exists flicker distortion along the boundaries of the static object, the vertical gradient can capture it.

For the pixel (x, i, t) in $\mathbf{L}_i$, the vertical gradient is computed as

$$V^g(x, i, t) = V(x, i, t+1) - V(x, i, t). \tag{2}$$

The temporal gradient $\mathbf{G}_i = \{V^g(x, i, t) | 1 \leqslant x \leqslant W, 1 \leqslant t \leqslant T\}$ is thus acquired. Therefore, the temporal gradient set $\{\mathbf{G}_{o,i}\}$ and $\{\mathbf{G}_{d,i}\}$ for the original and distorted synthesized videos could be obtained accordingly, and one example of them is shown in Fig. 5. In the red rectangles in Fig. 5, taking Undodancer as an example, the distorted synthesized view has more obviously black flocculus along the motion trajectory in the gradient temporal layer, compared with the original view, which implies the flicker distortion corresponds to the pattern changes in gradients. In practice, the gradient map needs patch refinement to exclude noises. The random noises in the original video captured by the cameras can also cause some small changes in gradients in temporal layer, which are actually not the flicker distortion. To reduce this influence of the noise from the capturing [26], we exclude the patches in the temporal layer if their gradient variance is small. The effective patches set in $\mathbf{G}_i$ layer can be defined as

$$\mathbf{S}_{V,i} = \{(u, v) | \frac{\sum_{x=u}^{u+w-1} \sum_{t=v}^{v+w-1} (V^g(x, i, t) - \overline{V_p^g(x, i, t)})^2}{w^2} > g\}, \tag{3}$$

$$\overline{V_p^g(x, i, t)} = \frac{\sum_{x=u}^{u+w-1} \sum_{t=v}^{v+w-1} V^g(x, i, t)}{w^2}, \tag{4}$$

where (u, v) denotes the patch index, indicating the location of one patch in the i-th temporal layer of the distorted synthesized video. w is the patch size. In our method, the variance threshold g is set as 5.

C. Depth Image Based Flicker Distortion Area Detection

Since not all the areas in the temporal layer include flicker distortion, we propose a flicker distortion area detection algorithm to locate the flicker distortion more precisely. In fact, the flicker distortion of synthesized videos usually locates at the object edges, which is mainly caused by the depth temporal inconsistency among frames and misalignment between depth and color videos at the depth edges or discontinuous regions. As shown in the right part of Fig. 3, the flicker distortion mainly exists in the areas of synthesized view corresponding to depth discontinuities, *e.g.*, strong edges or borders marked as red rectangles. Therefore, depth map and its discontinuities can be utilized to detect the flicker distortion. We use the edge detection operator, Canny, to detect the depth edges of the synthesized depth image and a large threshold is used to get the strong depth edges. The depth edges in the depth map is presented as $\mathbf{E}_{o,k} = \{D^{edge}(x, y, t) | 1 \leqslant x \leqslant W, 1 \leqslant y \leqslant H, t = k\}$. In section V. B, the threshold determination and analyses on the impacts of canny edge thresholds will be presented.

In addition, for the distorted synthesized video with depth compression, the location of flicker distortion in synthesized video usually deviates from the texture edges for a few pixels. The main reason lies in that the misalignment between color texture and depth videos and the depth errors induced by compression along the depth edges would easily generate the contour artifacts and neighborhood misplacement in synthesized views [39]. Therefore, to avoid the missed detection of the flicker area, image dilation is employed to expand the detected edges width for the depth edge map $\mathbf{E}_{o,k}$, and the dilated depth edge map $\mathbf{E}'_{o,k} = \{D^{edge'}(x, y, t) | 1 \leqslant x \leqslant W, 1 \leqslant y \leqslant H, t = k\}$ is obtained by using a squared

dilation mask. Since temporal layer images are divided into patches as processing units in our method, dilation radius of 2 is enough to capture the patches with the flicker distortion.

After the edge detection and dilation, the areas where flicker distortion takes place could be detected roughly. Since we measure the flicker distortion on the temporal layer, we convert the edge maps set $\{\mathbf{E}'_{o,k}\}$ into temporal edge layers $\{\mathbf{M}_{o,i}\}$, where $\mathbf{M}_{o,i} = \{D^{edge'}(x, y, t) | 1 \leqslant x \leqslant W, y = i, 1 \leqslant t \leqslant T\}$. If the distortion along the edges flicks or changes in a very short time, *i.e.*, the number of edge pixels in temporal layer is very small, human eyes can hardly perceive this flickering. In this case, we assume that only if the number of edge pixels in patches of the temporal layer $\mathbf{M}_{o,i}$ is more than a threshold B in a period, it may cause possible flicker perception. The edge patches set in the i-th temporal layer are refined as

$$\mathbf{S}_{E,i} = \{(u, v) | \sum_{x=u}^{u+w-1} \sum_{t=v}^{v+w-1} D^{edge'}(x, i, t) > B\}, \quad (5)$$

where (u, v) is the indices of edge patch in the i-th temporal edge layer $\mathbf{M}_{o,i}$ of the original depth video. B is set as 1.

The final flicker distortion area $\mathbf{S}_i$ in the i-th temporal layer could be obtained based on the textural gradient and depth edges, which is $\mathbf{S}_i = \mathbf{S}_{V,i} \cap \mathbf{S}_{E,i}$. The flicker distortion area of whole video consists of flicker area from all the temporal layers, *i.e.*, $\{\mathbf{S}_i\}$. In addition, the binarization weight map $\mathbf{W}_i$ in each temporal layer $\mathbf{G}_i$ can be obtained accordingly as

$$W_i(u, v) = \begin{cases} 1, & \text{if } (u, v) \in \mathbf{S}_i \\ 0, & \text{otherwise} \end{cases}. \quad (6)$$

where $W_i(u, v)$ is the element of $\mathbf{W}_i$. With the assistance of the depth map, the flicker distortion area is located more accurately.

D. Sparse Representation for Flicker Distortion Features

In this section, the sparse representation is used to measure the distortion between the original and distorted videos in the detected flicker distortion areas, which includes temporal dictionary learning phase and sparse representation phase.

1) Temporal Dictionary Learning: To represent the flicker distortion in the synthesized video, the dictionary that aims to learn the temporal flicker features of 2D-XT or 2D-YT plane could be learned. Since the 2D-XT and 2D-YT plane have the similar effects in capturing the motion features, the dictionary learning function for 2D-XT plane is used and can be derived from (1) as

$$\min_{\mathbf{\Psi}^{2D}, \boldsymbol{\alpha}^{2D}} \|\mathbf{X}^{2D} - \mathbf{\Psi}^{2D}\boldsymbol{\alpha}^{2D}\|_2^2, s.t. \|\boldsymbol{\alpha}_j^{2D}\|_0 \leq \varepsilon, \quad (7)$$

where $\mathbf{X}^{2D}$ denotes the training patches set of 2D-XT from video data. $\mathbf{\Psi}^{2D}$ is the learned 2D dictionary for temporal layers. $\|.\|_0$ means the number of nonzero entries in the vector. During dictionary learning, the number of nonzero entries of $\boldsymbol{\alpha}_j^{2D}$ should not be greater than a given ε. In our case, ε is set as 6. The dictionary is learned by using K-SVD [40]. During learning, the sparse coefficients are solved by OMP algorithm [41]. The learned temporal dictionary is 64×256.

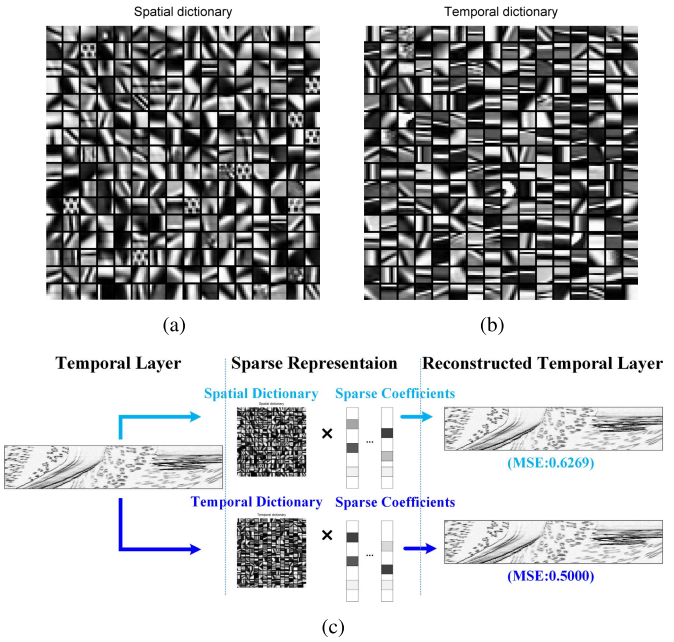


Fig. 6. The comparison between the learned spatial and temporal dictionaries. (a) spatial dictionary; (b) temporal dictionary; (c) Reconstructed temporal layer by using temporal or spatial dictionary.

To demonstrate the capability of representing flicker distortion by using temporary dictionary, a temporal layer with flicker distortion was represented by temporary dictionary learned from temporal layers or spatial dictionary learned from conventional images, as shown in Fig. 6(a) and (b). Both the two dictionaries were learned with the same training sequences and learning parameters. From Fig. 6(c), it is observed that the reconstructed error, measured by MSE between the original layer and reconstructed temporal layer using temporal dictionary is 0.5000, which is less than that of reconstructed temporal layer by using spatial dictionary, *i.e.*, 0.6269. The temporal dictionary is more effective since it is learned from temporary layers. Therefore it is suitable to be used to capture the temporal activity and flickering of the synthesized video. The goal of temporal dictionary learning is to learn the normal temporal activity of non-flicker distortion. Thus, when the learned dictionary is employed upon the original and distorted synthesized videos via sparse representation, the flicker distortion existing in the distorted videos could be distinguished.

2) Sparse Representation for Flicker Distortion: To represent the flicker distortion, two types of features based on sparse representation are used as the distortion features. One is the phase distortion which is employed to measure the flocculus shape features of the flicker distortion; the other is the amplitude distortion which can capture the strength of the flicker distortion. For patches $\mathbf{y}_{u,v}^{o,i}$ in the original video and its corresponding patch $\mathbf{y}_{u,v}^{d,i}$ in the distorted synthesized video in the i-th temporal layer, the two features can be written as

$$P(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i}) = \begin{cases} \frac{\left|\left\langle \boldsymbol{\alpha}_{u,v}^{o,i}, \boldsymbol{\alpha}_{u,v}^{d,i} \right\rangle\right| + c}{\left\|\boldsymbol{\alpha}_{u,v}^{o,i}\right\|_2 \cdot \left\|\boldsymbol{\alpha}_{u,v}^{d,i}\right\|_2 + c}, & \text{if } W_i(u, v) = 1 \\ 1, & \text{otherwise,} \end{cases} \quad (8)$$

$$A(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i}) = \begin{cases} 1 - \frac{\left|\left\|\boldsymbol{\alpha}_{u,v}^{o,i}\right\|_2 - \left\|\boldsymbol{\alpha}_{u,v}^{d,i}\right\|_2\right| + c}{\left\|\boldsymbol{\alpha}_{u,v}^{o,i}\right\|_2 + \left\|\boldsymbol{\alpha}_{u,v}^{d,i}\right\|_2 + c}, & \text{if } W_i(u,v) = 1 \\ 1, & \text{otherwise,} \end{cases} \tag{9}$$

where $\boldsymbol{\alpha}_{u,v}^{o,i}$ and $\boldsymbol{\alpha}_{u,v}^{d,i}$ are the sparse coefficients of the original patch $\mathbf{y}_{u,v}^{o,i}$ and the distorted patch $\mathbf{y}_{u,v}^{d,i}$ with respect to the learned dictionary $\boldsymbol{\Psi}^{2D}$ by (7), respectively. $\langle\cdot\rangle$ denotes the inner product. c is a constant with a small value added to prevent the denominator to be zero and is set as 0.02. $P(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})$ computes the phase similarity between sparse coefficients $\boldsymbol{\alpha}_{u,v}^{o,i}$ and $\boldsymbol{\alpha}_{u,v}^{d,i}$ and can be used to measure the structural similarity between $\mathbf{y}_{u,v}^{o,i}$ and $\mathbf{y}_{u,v}^{d,i}$. $A(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})$ measures the amplitude similarity between $\mathbf{y}_{u,v}^{o,i}$ and $\mathbf{y}_{u,v}^{d,i}$ through sparse coefficients. $P(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})$ and $A(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})$ are both among the range [0, 1]. The combination of (8) and (9) can be used to measure the integral similarity between patch $\mathbf{y}_{u,v}^{o,i}$ and $\mathbf{y}_{u,v}^{d,i}$ [31]. Therefore, $P(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})$ and $A(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})$ will be small in representing large flicker distortions, and vice versa.

Since human eyes tend to perceive the flicker distortion in the form of regions instead of lines, the flicker distortion can be computed over multiple temporal layers instead of a single layer. For simplicity, the sparse coefficients $P(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})$ and $A(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})$ of a group of temporal layers, *i.e.*, $\mathbf{U}_k = \{\mathbf{L}_i | h_s(k-1)+1 \leqslant i \leqslant h_s k\}$, $k \in [1, \frac{H}{h_s}]$, are averagely merged, and the integral similarity for patches locating at (u, v) is

$$\overline{C_k(u,v)} = \frac{1}{h_s} \sum_{i=h_s(k-1)+1}^{h_s k} [A(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})]^a [P(\mathbf{y}_{u,v}^{o,i}, \mathbf{y}_{u,v}^{d,i})]^b, \tag{10}$$

where h_s is the number of temporal layers for averaging, k is an index of $\mathbf{U}_k$, and a and b are parameters denoting weights of amplitude and phase distortion. We set a and b as 1, respectively for simplicity. Finally, the score of the flicker distortion of $\mathbf{U}_k$ can be obtained as

$$SF_k = \frac{\sum_u \sum_v (1 - \overline{C_k(u,v)})}{\sum_u \sum_v \overline{W(u,v)}}, \tag{11}$$

where the weight map $\overline{W(u,v)}$ is obtained by

$$\overline{W(u,v)} = \begin{cases} 1, & \text{if } (u,v) \in \bigcup_{i=h_s(k-1)+1}^{h_s k} \mathbf{S}_i \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

E. Weighted Pooling for Temporal Layers

Since the group of temporal layers $\mathbf{U}_k$ in fact contributes unevenly to the visual perception, a weighted pooling scheme is proposed for the temporal layers. It is observed that the temporal layer with more edge patches probably makes more contribution to the final perception of the flicker distortion. Therefore, we employ a rank-based method [10], [42] to pool the scores among temporal layers. The flicker score of the whole distorted video SF_V is

$$SF_V = \frac{\sum_{k=1}^{H_s} w_k SF_k}{\sum_{k=1}^{H_s} w_k}, \tag{13}$$

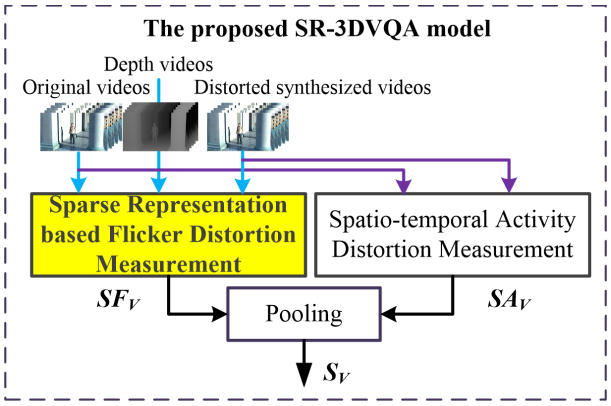


Fig. 7. The framework of the proposed SR-3DVQA model.

where w_k represents the weight of each layer, $H_s = \frac{H}{h_s}$, and SF_k represent the index and the flicker score of the k-th group temporal layer $\mathbf{U}_k$, respectively. This SF_V score is normalized to range [0, 1] through the normalization of the summation of the weight of w_k, which is calculated as

$$w_k = log_2\left(1 + \frac{Rank_k}{H_s}\right), \tag{14}$$

where $Rank_k$ represents the rank of the $\mathbf{U}_k$ among all layers in terms of the importance, *i.e.*, the number of edge patches. In this way, the flicker distortion score SF_V of the distorted synthesized video is obtained.

IV. The Proposed SR-3DVQA Model

The distortions of synthesized video mainly have two categories [26]. One is the flicker distortion, the other is about the conventional spatio-temporal activity distortions in synthesized video, such as compression artifacts, rendering distortion, contour and hole artifacts. The proposed SR-3DVQA model, as shown in Fig. 7, is mainly composed of two modules, including SR-FDM and the spatio-temporal activity distortion measurement. Both the original video and the distorted synthesized video are input into the two modules, and additionally, the synthesized depth video is input into the SR-FDM module. The overall quality score of a compressed synthesized video is predicted by pooling the flicker distortion score and the spatio-temporal activity distortion score.

A. Spatial-Temporal Activity Distortion Measurement

In our method, we employ the same method as that in Liu [26] to assess spatial activity distortion, which mainly measures the variance difference of the pixel gradients in a spatio-temporal tube. The concepts of Quality Assessment Group of Pictures (QA-GoP) and Spatio-Temporal (S-T) tube are introduced in the spatio-temporal activity distortion measurement method, which is illustrated in Fig. 8. A video is divided into several QA-GoPs, which is made up of a number of frames, *e.g.*, $2N + 1$ frames as shown in Fig. 8. A QA-GoP consists of multiple S-T tubes, which are concatenated by matched blocks via motion estimation algorithms in adjacent frames, denoting the motion trajectory. Given the original

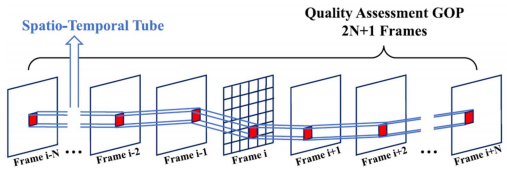


Fig. 8. The spatio-temporal Tube and GoP for quality assessment [26].

video $\mathbf{V}_o$ and distorted synthesized video $\mathbf{V}_d$, first of all, the pixel gradients $G_o^S(x, y, t)$, and $G_d^S(x, y, t)$, are computed by calculating the norm of the pixel gradient vector composed of horizontal and vertical spatial gradients at frame t, respectively.

$$G_\varphi^S(x, y, t) = \sqrt{\left|\nabla G_{\varphi,x}^S(x, y, t)\right|^2 + \left|\nabla G_{\varphi,y}^S(x, y, t)\right|^2}, \quad (15)$$

where $\varphi \in \{o, d\}$, $\nabla G_{\varphi,x}^S(x, y, t)$, $\nabla G_{\varphi,y}^S(x, y, t)$ are gradients in the horizontal and vertical directions, respectively. Then, the gradients are organized by S-T tube, and the standard deviation of the gradients in the i-th S-T tube in a QA-GoP is computed as [26]

$$\sigma_\varphi^{tube}(x_n, y_n, t_n) = \sqrt{\frac{\sum_{t=t_n-N}^{t_n+N} \sum_{x=x_n}^{x_n+w-1} \sum_{y=y_n}^{y_n+h-1} (G_\varphi^S(x, y, t) - \overline{G_{\varphi,tube}^S(x_n, y_n, t_n)})^2}{w \times h \times (2N+1)}}, \quad (16)$$

$$\overline{G_{\varphi,tube}^S(x_n, y_n, t_n)} = \frac{\sum_{t=t_n-N}^{t_n+N} \sum_{x=x_n}^{x_n+w-1} \sum_{y=y_n}^{y_n+h-1} G_\varphi^S(x, y, t)}{w \times h \times (2N+1)}, \quad (17)$$

where w and h are width and height of the tube in spatial domain. N is the number of forward or backward frames involved in a S-T tube. The spatio-temporal activity $R_{\varphi,i}^{tube}$ can be then obtained by clipping $\sigma_\varphi^{tube}(x_n, y_n, t_n)$, where i is the index of tube $\{x_n, y_n, t_n\}$. They are calculated as

$$R_{\varphi,i}^{tube} = \begin{cases} \sigma_\varphi^{tube}(x_n, y_n, t_n), & \text{if } \sigma_\varphi^{tube}(x_n, y_n, t_n) > \tau \\ \tau, & \text{otherwise}, \end{cases} \quad (18)$$

where τ is the perceptible threshold for spatio-temporal gradient standard deviation, and is set as 180 in [26]. Afterwards, the distortion score of a QA-GoP is calculated through worst-case pooling strategy, and the overall spatio-temporal distortion score of the whole video is obtained as.

$$SA_V = \frac{1}{N_{all}} \sum \frac{1}{N_\Phi} \sum_{i \in \Phi} \left| log_{10}\left(\frac{R_{d,i}^{tube}}{R_{o,i}^{tube}}\right)\right|, \quad (19)$$

where Φ denotes the set of the worst 5% S-T tubes in a QA-GoP [26], N_Φ denotes the number of tubes in set Φ, N_{all} represents the number of QA-GoP in a test video.

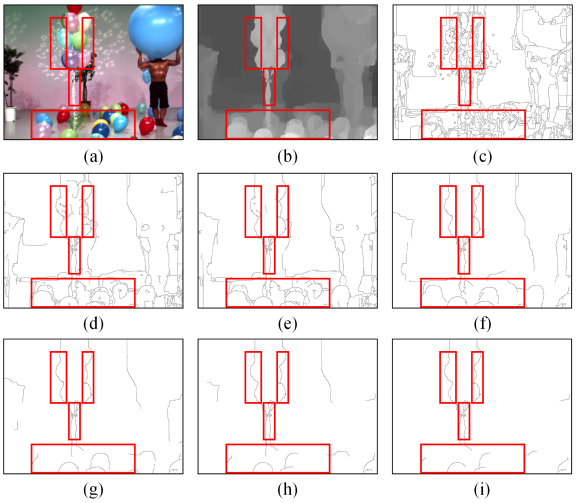


Fig. 9. The edge detection effects among different canny thresholds in edge detection of depth map for Balloons. (a) the distorted synthesized view by depth map QP46; (b) the original depth map; (c)–(i) are edge maps by canny with threshold 0.03, 0.07, 0.1, 0.2, 0.3, 0.4, and 0.5, respectively.

B. Pooling

A general pooling method combining the summation and the multiplication is explored to integrate the flickering and spatio-temporal distortions in assessing the synthesized video quality, which can be written as

$$S_V = c \times (w_1 SF_V + w_2 SA_V) + d \times f(SF_V) \times SA_V, \quad (20)$$

where c, d are weighted parameters to balance the relative importance of the summation and multiplication pooling items. $f(.)$ denotes the map function of the flicker distortion score in multiplication pooling. w_1, w_2 are used to weigh the flicker distortion score and the spatio-temporal activity distortion score in summation pooling. When d is set to zero, (20) is degenerated to the summation. Similarly, when c is set to zero, (20) is degenerated to the multiplication. In this paper, c, d, w_1, w_2 are set as 1, 0, 0.5, 0.5, respectively, which denotes the flicker distortion and spatio-temporal activity distortion are summated in the pooling stage. The impacts of the pooling method, weight parameters and mapping function $f(.)$ are discussed in Section V.D.

V. Experimental Results and Analyses

In this section, we firstly present the canny threshold determination for edge detection. Then, the quality assessment performance is compared among the proposed SR-3DVQA model and state-of-the-art metrics. The statistical significance test is conducted subsequently. Finally, the impacts of some factors and key parameters in our proposed method are analyzed and discussed.

A. Canny Threshold Determination

It's important to choose a suitable canny threshold in depth edge detection for flicker area detection. We compare the edge detection effects among different canny thresholds. Figs. 9, 10 demonstrate the relationship between the flicker areas in the

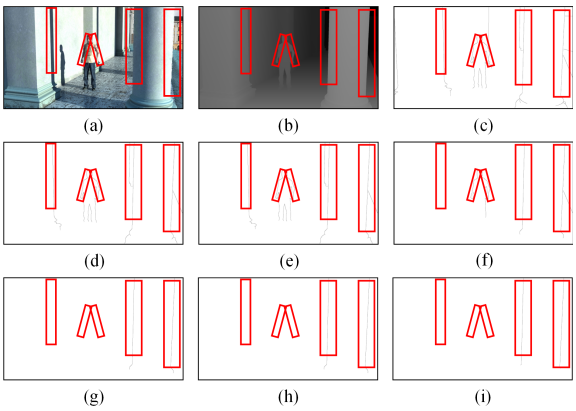


Fig. 10. The edge detection effects among different canny thresholds in edge detection of depth map for Undodancer. (a) the distorted synthesized view by depth map QP40; (b) the original depth map; (c)–(i) are edge maps by canny with threshold 0.03, 0.07, 0.1, 0.2, 0.3, 0.4, and 0.5, respectively.

synthesized view and the canny thresholds in depth edge detection on sequence Balloons and Undodancer, where the marked red rectangles in (a) to (i) are the areas with the flicker distortion. While selecting an optimal threshold, the depth edges in the red rectangles, which is corresponding to flicker areas, shall be completely detected, and the edges outside the rectangles shall be excluded or included as few as possible. It's observed that edge maps generated by thresholds from 0.1 to 0.5 are more consistent to this flickering area for sequence Balloons, where strong edges of balloons and ribbons inside the red rectangles are detected and less edges of plants and the man outside the rectangles are detected. In this case, thresholds among range 0.1 to 0.5 seem suitable for sequence Balloons. For Undodancer, the flicker mainly concentrates on the upper body of the dancer and the rightmost and foremost pillar, so thresholds among range 0.07 to 0.2 are feasible. Large threshold will only include the strong edges and exclude some flickering regions by mistake, while small threshold will falsely include more non-flickering regions, which both degrades the accuracy of the proposed algorithm. According to results among the test sequences, threshold ranges from 0.1 to 0.4 are more reasonable and 0.2 is selected for the edge detection in this paper.

B. Quality Prediction Performance Comparisons

The training dataset, testing dataset and settings of the SR-3DVQA model are first introduced. Then, the quality prediction performance among different methods are compared.

1) Settings for Temporal Dictionary Learning: Due to the temporal inconsistency in the generated depth video, the synthesized video rendered from the original texture videos and depth videos may also have noticeable flicker distortion. Therefore, the original view instead of the original synthesized video is preferred for temporal dictionary learning. For the original texture video in MVD system has similar temporal properties with the conventional single-view video, we selected the conventional videos from HEVC test sequences as training sequences so as to separate the training sequences from the test sequences. To cover different spatial resolution and content, eight 2D video sequences were selected in the temporal dictionary learning, *i.e.*, BasketballDrive, FourPeople, Flowervase, Johnny, KristenAndSara, ParkScene, RaceHorses, and Vidyo3. The properties of the training sequences are shown in Table II. The first 300 or 240 frames of each sequence were kept in training. For each sequence, four temporal layers were extracted at a uniform sampling way along the frame height. Then 32 different temporal layers in total were extracted. Temporal layer images with pixel intensity are directly employed as the feature maps in dictionary learning. These images were then divided into patches with size 8×8 and one-pixel overlap, which were collected as the training dataset for temporal dictionary learning. Note that the dictionary could be learned from either the intensity or the gradient map of training videos. The reason is that both the two temporary dictionaries are capable of learning atoms with sharp edges and boundaries for capturing conventional motion. In addition, similar performances can be found when using the dictionaries that learned from different training groups and layer images, which validates the robustness and reliability of the learned dictionaries.

TABLE II
THE PROPERTIES OF THE TRAINING SEQUENCES

Training Sequences	Resolution	Frame Length	Extracted Layer Index
BasketballDrive	1920×1080	300	216,432,648,864
FourPeople	1280×720	300	144,288,432,576
Flowervase	832×480	300	96,192,288,384
Johnny	1280×720	300	144,288,432,576
KristenAndSara	1280×720	300	144,288,432,576
ParkScene	1280×720	240	144,288,432,576
RaceHorses	416×240	300	48,96,144,192
Vidyo3	1280×720	300	144,288,432,576

2) Dataset and Settings for SR-3DVQA Prediction: The SIAT synthesized video database [26] was adopted as the testing dataset, which is totally different from the learning dataset. It consists of 10 MVD sequences and 140 synthesized videos in 1024×768 and 1920×1088 resolution which were obtained by 14 different combinations of compressed texture/depth videos, namely generated with different quantization parameters. Each video was synthesized by two views composed of two texture videos and their corresponding depth videos. The number of frames in the test sequences is 200 except sequence Bookarrival, whose number is 100. According to whether the texture/depth video is "compressed" or "uncompressed", the generated distorted synthesized videos are categorized into four subsets: $U_T U_D$, $U_T C_D$, $C_T U_D$, and $C_T C_D$. 'C' and 'U' mean the videos are "compressed" and "uncompressed" respectively while the subscripts 'T' and 'D' denote texture videos and depth videos respectively. Taking $C_T C_D$ for example, it represents the synthesized video subset were synthesized from the texture and depth videos with compression distortions. The subjective experiment was conducted by single stimulus paradigm with continuous score. Difference Mean Opinion Scores (DMOS) were provided.

TABLE III
THE PLCC, SROCC, AND RMSE COMPARISON BETWEEN DIFFERENT METHODS. THE BEST RESULTS ARE MARKED IN BOLD

Methods	$U_T C_D$			$C_T U_D$			$C_T C_D$			ALL Data		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
PSNR	0.545	0.481	0.093	0.569	0.566	0.109	0.658	0.666	0.085	0.645	0.627	0.098
SSIM [1]	0.576	0.465	0.102	0.533	0.534	0.112	0.708	0.704	0.079	0.629	0.598	0.100
WSNR [42]	0.316	0.295	0.105	0.770	0.778	0.085	0.610	0.645	0.089	0.605	0.589	0.102
MSSSIM [2]	0.709	0.626	0.078	0.714	0.718	0.093	0.843	0.849	0.060	0.743	0.731	0.086
IW-PSNR [43]	0.514	0.458	0.095	0.711	0.723	0.093	0.686	0.671	0.082	0.686	0.663	0.093
IW-SSIM [43]	0.751	0.741	0.073	0.811	0.811	0.078	0.868	0.863	0.056	0.795	0.792	0.078
VQM [4]	0.608	0.527	0.088	0.753	0.755	0.087	0.632	0.643	0.087	0.671	0.655	0.095
MOVIE [5]	0.603	0.573	0.088	0.655	0.649	0.100	0.703	0.713	0.080	0.710	0.693	0.090
Bosc [7]	0.317	0.293	0.105	0.385	0.376	0.122	0.535	0.489	0.095	0.461	0.431	0.114
MP-PSNR [12]	0.576	0.490	0.090	0.508	0.496	0.1142	0.595	0.598	0.090	0.601	0.587	0.103
MP-PSNRr [14]	0.560	0.523	0.092	0.489	0.485	0.116	0.543	0.547	0.094	0.589	0.581	0.104
MW-PSNR [13]	0.570	0.501	0.091	0.466	0.480	0.117	0.555	0.568	0.094	0.580	0.569	0.105
MW-PSNRr [14]	0.569	0.520	0.091	0.458	0.451	0.118	0.548	0.565	0.094	0.575	0.569	0.105
3DSwIM [9]	0.322	0.266	0.105	0.105	0.193	0.132	0.382	0.243	0.104	0.386	0.268	0.118
LOGS [10]	0.656	0.625	0.083	0.511	0.523	0.114	0.571	0.550	0.092	0.641	0.570	0.099
PSPTNR [23]	0.487	0.527	0.096	0.339	0.391	0.125	0.353	0.338	0.105	0.433	0.453	0.116
Liu [26]	0.824	0.824	0.063	0.843	0.838	0.071	0.868	0.863	0.056	0.868	0.869	0.064
FDI [24]	–	–	–	–	–	–	–	–	–	0.595	0.570	0.103
SR-3DVQA	**0.916**	**0.886**	**0.044**	**0.909**	**0.920**	**0.055**	**0.894**	**0.888**	**0.051**	**0.910**	**0.914**	**0.053**

As for the proposed SR-3DVQA, the parameters are set as follows: (1) parameters in training: patch size is 8×8, sparsity ε is 6, dictionary size is 64×256 [45]; (2) parameters in flicker detection module: patch variance g is 5, canny threshold is set as 0.2, and the dilation mask is 2×2, and threshold B for classifying edge patches is 1; (3) parameters in sparse representation: the layer size h_s is 8. Detailed analyses on their impacts will be presented in Section V.F. The comparison methods include three categories: eight conventional 2D IQA/VQA metrics, *i.e.*, PSNR, SSIM [1], WSNR [43], MSSSIM [2], IW-SSIM [44], IW-PSNR [44], VQM [4] and MOVIE [5], seven 3D synthesized IQA metrics, *i.e.*, Bosc [7], MP-PSNR [12], MW-PSNR [13], MP-PSNRr [14], MW-PSNRr [14], 3DSwIM [9] and LOGS [10], three synthesized VQA metrics, *i.e.*, PSPTNR [23], Liu [26], and Zhou *et al*. [24] (denoted as FDI). Note for those IQA metrics, the score of each video was obtained by averaging the scores of all the frames in the video. In addition, the results of FDI, a non-reference method, are referred from [24] with the whole SIAT synthesized video database.

To measure the performances of the VQA methods, the quality scores of all the proposed and benchmark objective VQA methods were obtained first when they were tested on the database. Then, these quality scores were fitted into the predicted DMOS with the same range as that of the groundtruth DMOS via a nonlinear regression function. In this paper, a five-parameter nonlinear regression function was used, which is

$$f(x) = \eta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\eta_2 (x - \eta_3)}}\right) + \eta_4 x + \eta_5, \qquad (21)$$

where η_1 to η_5 are fitting parameters, x denotes the objective score of the quality metrics, and $f(x)$ is the predicted subjective score obtained by nonlinearly fitting x to range [0, 1] [46]. Finally, Spearman Rank Order Correlation Coefficient (SROCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Squared Error (RMSE) between the predicted DMOS and the true DMOS are computed to measure the performance of VQA metrics [47]. The prediction performance of the VQA scheme is better if PLCC and SROCC approach to 1.0 and RMSE approaches to 0.

Table III shows the performance comparison between the state-of-the-art benchmark methods and the proposed SR-3DVQA method on the SIAT database, which includes three subsets $U_T C_D$, $C_T U_D$, $C_T C_D$, and the ALL dataset consisting of the three subsets. Video samples in $U_T U_D$ are included into $C_T C_D$, which is the same as [26]. In terms of all performance indices SROCC, PLCC, and RMSE on different subsets and ALL dataset, the best one is marked in bold. As shown in the Table III, for $U_T C_D$ dataset where the depth video was distorted, conventional 2D metrics IW-SSIM and MSSSIM perform better than other benchmark schemes. It is because the depth distortion causes the geometrical distortion in the rendered view, and conventional 2D metric, such as PSNR, may overestimate the geometrical distortion. The performance of LOGS, a metric proposed for 3D synthesized image, follows the IW-SSIM and MSSSIM. The PLCC and SROCC values of Liu and the proposed SR-3DVQA method are much higher than all the rest methods. Our method has the highest performance with dominant superiority, which indicates our method can predict the flicker distortion very well and have better consistency with human perception. For $C_T U_D$, three 2D quality metrics IW-SSIM, WSNR, and VQM are good, since they are designed for compression and structural distortion for 2D images/videos which are probably the main distortions

in $C_T U_D$. All the 3D synthesized image/video metrics don't perform well except Liu and our method since their methods haven't considered the distortion induced by the compressed texture videos. Similar to $U_T C_D$, the proposed SR-3DVQA method performs the best in terms of the PLCC, SROCC and RMSE. Similarly, on $C_T C_D$, our proposed method performs the best among them while Liu and IW-SSIM have very similar performance, and MSSSIM and SSIM perform fairly good and are better than other methods. In the ALL dataset, IW-SSIM places the third after Liu and our method. Liu's method ranks the second while the proposed SR-3DVQA method is the best among the benchmark schemes.

C. Statistical Significance Test for SR-3DVQA

Moreover, to further verify the effectiveness of the proposed method, statistical significance test is performed. F-test based on the result of the variance ratio of the predicted residuals between two methods was used to indicate the significance. The predicted residual is obtained from DMOS_P predicted by test model and the ground truth DMOS, which can be described as

$$res(k) = DMOS_P(k) - DMOS(k), \tag{22}$$

where $res(k)$ represents the predicted residual of the test model on video k. The variance of the residuals, termed as Var_i, of the test model i on all the videos could be calculated. Then, the variance ratio $R_{i,j}$ between test models i and j could be computed, which could be written as

$$R_{i,j} = \frac{Var_i}{Var_j}. \tag{23}$$

If $R_{i,j}$ is greater than the F-ratio threshold which is determined by the sample size and the significance level, it means the performance of test model j is significantly superior to that of test model i; otherwise, the difference is insignificant. Based on the variance ratios between the benchmark schemes and the proposed method on the four datasets, the significance test results can be obtained. The variance ratios and significance results are listed in column as '$R_{i,j}$/sig.' in Table IV, where $R_{i,j}$ is the variance ratio and sig. is the significance result. The symbol '1', '-1', or '0' denote the proposed method is 'significantly superior', 'significantly inferior', or 'insignificant' to the benchmark method in the row, respectively. It can be observed that on $C_T U_D$, our method is significantly superior to all benchmark schemes except for Liu's scheme. On $C_T C_D$ subset, except for Liu, MSSSIM, and IW-SSIM, our method is significantly superior to all other methods. Liu, MSSSIM, and IW-SSIM may be comparable in evaluating the synthesized videos whose distortions are coming from color videos. In addition, the proposed method have significantly superior performance than all the other methods on the ALL dataset and subset $U_T C_D$. It is because the proposed SR-3DVQA is mainly proposed to evaluate the flicker distortion in synthesized video caused by the depth map and it works very well. Overall, the significance test has further validated the superiority of the proposed method in predicting the quality of the synthesized videos.

TABLE IV

THE SIGNIFICANCE RESULTS OF THE PROPOSED METHOD AGAINST COMPARISON METHODS BASED ON THE RESIDUAL VARIANCE RATIO RESULTS BETWEEN COMPARISON METHODS AND THE PROPOSED METHOD. THE SYMBOL '1' INDICATES THAT THE PROPOSED METHOD IS SIGNIFICANTLY SUPERIOR TO THE COMPARISON METHOD, AND THE SYMBOL '−1' INDICATES THE OPPOSITE SITUATION, WHILE THE SYMBOL '0' INDICATES THAT THERE IS NO SIGNIFICANT DIFFERENCE BETWEEN THE PROPOSED METHOD AND THE COMPARISON METHOD

Methods	$U_T C_D$	$C_T U_D$	$C_T C_D$	ALL
	$R_{i,j}$/sig.	$R_{i,j}$/sig.	$R_{i,j}$/sig.	$R_{i,j}$/sig.
PSNR	4.4000/1	3.9355/1	2.8077/1	3.3448/1
SSIM [1]	5.3000/1	4.1613/1	2.4615/1	3.4483/1
WSNR [42]	5.6500/1	2.3548/1	3.1154/1	3.6207/1
MSSSIM [2]	3.1000/1	2.8387/1	1.4231/0	2.5517/1
IW-PSNR [43]	4.6000/1	2.8710/1	2.6154/1	3.0345/1
IW-SSIM [43]	2.7000/1	2.0000/1	1.2308/0	2.1034/1
VQM [4]	3.9500/1	2.5161/1	2.9615/1	3.1379/1
MOVIE [5]	4.0000/1	3.3226/1	2.5000/1	2.8276/1
Bosc [7]	5.6000/1	4.9677/1	3.5385/1	4.5172/1
MP-PSNR [12]	4.2000/1	4.3226/1	3.1923/1	3.6552/1
MP-PSNRr [14]	4.3000/1	4.4194/1	3.5000/1	3.7241/1
MW-PSNR [13]	4.2000/1	4.5484/1	3.4231/1	3.7931/1
MW-PSNRr [14]	4.2500/1	4.5806/1	3.4615/1	3.8276/1
3DSwIM [9]	5.6000/1	5.7419/1	4.2308/1	4.8621/1
LOGS [10]	3.5500/1	4.2903/1	3.3462/1	3.3793/1
PSPTNR [23]	4.7500/1	5.1613/1	4.3462/1	4.6552/1
Liu [26]	2.0000/1	1.6774/0	1.2308/0	1.4138/1
F-ratio threshold	1.6927	1.6927	1.5994	1.3217

D. Impacts of Pooling Methods

The pooling methods of flicker distortion measurement and spatio-temporal activity distortion measurement are analyzed in this subsection. For (20), when (c, d) was set as (1, 0), *i.e.*, the summation pooling, the effects of weight parameter pair (w_1, w_2) on the performance in quality assessment were explored. (w_1, w_2) were set as (0.5, 0.5), (0.6, 0.4), (0.8, 0.2), (0.4, 0.6), and (0.2, 0.8). The performance measured by PLCC, SROCC, and RMSE of the five combinations of (w_1, w_2) is listed in the second to sixth row in Table V. It can be observed that (0.5, 0.5) has better performance than other (w_1, w_2) pairs in terms of PLCC, SROCC and RMSE. For the multiplication pooling, *i.e.*, (c, d) was set as (0, 1), the role of the map function $f(.)$ was analyzed by comparing the performance of six types of $f(.)$, *i.e.*, $f(x) = x$, 'log10', 'log2', 'cubic', 'square', 'square root'. The corresponding results are demonstrated in the eighth to fifteenth row in Table V. It is noted that function 'square' excels all the other five functions. But the multiplication pooling is a little inferior to the summation pooling even with different mapping function $f(x)$. Based on the best performance of summation and multiplication pooling, the weight parameters (c, d) in the combination of the two methods had also been investigated. The values of (c, d) were set the same range as (w_1, w_2). The last five rows in Table V show the performance. It can be found that when the value of c is equal or greater than d, it achieves the best performance among the five (c, d) combinations. In fact,

TABLE V

THE IMPACTS OF POOLING METHODS EMPLOYED IN THE PROPOSED SR-3DVQA MODEL

Pooling methods	(w_1, w_2)	PLCC	SROCC	RMSE
Summation (c, d)=(1, 0)	(0.5,0.5)	0.910	0.914	0.053
	(0.6,0.4)	0.900	0.905	0.056
	(0.8,0.2)	0.851	0.852	0.067
	(0.4,0.6)	0.907	0.908	0.054
	(0.2,0.8)	0.862	0.862	0.065
Pooling methods	$f(.)$	PLCC	SROCC	RMSE
Multiplication (c, d)=(0, 1)	$f(x) = x$	0.889	0.889	0.059
	'log10'	0.880	0.879	0.061
	'log2'	0.880	0.879	0.061
	'cubic'	0.896	0.899	0.057
	'square'	0.904	0.906	0.055
	'square root'	0.853	0.846	0.067
Pooling methods	(c, d)	PLCC	SROCC	RMSE
Combination (w_1, w_2)=(0.5, 0.5) $f(.)$='square'	(0.5,0.5)	0.909	0.913	0.053
	(0.6,0.4)	0.910	0.913	0.053
	(0.8,0.2)	0.910	0.914	0.053
	(0.4,0.6)	0.909	0.912	0.053
	(0.2,0.8)	0.908	0.911	0.054

the best performance is obtained via the summation pooling with (w_1, w_2) as (0.5, 0.5) or the combination pooling when c is larger than 0.6. Overall, the pooling methods have noticeable impacts on the final performance; the best performance is obtained via the summation pooling with (w_1, w_2) as (0.5, 0.5) and it is a simpler form as compared with the combination pooling. Therefore, the summation pooling is employed and (w_1, w_2) is set as (0.5, 0.5) in the proposed SR-3DVQA model.

E. Impacts of the Reference Depth Video

We also analyzed the impacts from the reference depth video employed in the proposed SR-3DVQA model. The depth videos employed can be the original depth video or the synthesized depth video at the virtual viewpoint. The advantage of using the original depth video is that it has better picture quality as compared with using the synthesized depth video. However, the disadvantage is the original depth video at the virtual viewpoint may not be available. Using the synthesized depth video is more practical. A comparative experiment was conducted to analyze the influence from different reference depth videos used in the proposed model. Since sequence Lovebird1 and Newspaper don't have the corresponding original depth video at the virtual viewpoint, all the rest eight sequences in the database were used for comparison. Similarly, the testing database is also categorized as four datasets, $U_T C_D$, $C_T U_D$,$C_T C_D$, and ALL dataset. The PLCC, SROCC, RMSE results are demonstrated in Table VI. To distinguish these reduced datasets from those in Section V-B, we mark them with '*'. The left three columns are the results from the synthesized depth video. The values of most PLCC, SROCC and RMSE using the synthesized depth video on the test datasets are a little better than the results of using the original depth video. Basically, they are comparable. It indicates that

TABLE VI

THE IMPACTS OF DEPTH VIDEOS EMPLOYED IN THE PROPOSED SR-3DVQA MODEL

Datasets	Synthesized Depth Video			Original Depth Video		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
$U_T C_D{}^*$	0.925	0.893	0.043	0.916	0.887	0.045
$C_T U_D{}^*$	0.908	0.914	0.060	0.909	0.920	0.059
$C_T C_D{}^*$	0.923	0.921	0.052	0.921	0.915	0.053
ALL Data*	0.913	0.914	0.054	0.908	0.908	0.056

TABLE VII

IMPACTS OF DICTIONARY SIZES IN DICTIONARY LEARNING

Indices	128	256	512	1024	2048
PLCC	0.913	0.910	0.903	0.896	0.885
SROCC	0.918	0.914	0.909	0.896	0.885
RMSE	0.052	0.053	0.055	0.057	0.060

although the original depth video has more precise depth values, the synthesized depth video, which are generated through DIBR from two original depth videos, is comparable or a little better in the synthesized video quality prediction. Moreover, using the synthesized depth video is more practical. The main reason is the depth video is used to help locate the flicker area by using edge detection and dilation. The original depth video has more precise depth values but may have geometrical misalignment with the synthesized texture video. Therefore, the synthesized depth video is better to be used for both practical usage and better performance. MVD data have different types of depth maps, such as captured by depth cameras, generated by stereo-matching algorithms and computer graphics. However, extensive analyses show that the prediction performances of the SR-3DVQA are similar and robust with different types of depth maps.

F. Impacts of Key Parameters

To analyze the impacts of some key parameters in the proposed SR-3DVQA method, some univariate experiments were conducted, in which only one parameter was changed while the rest were fixed.

1) Dictionary Size and Sparsity in Dictionary Learning: In this paper, the dictionary size is set as 64×256 and sparsity is 6 [45]. To further analyze their impacts in SR-3DVQA, different dictionary sizes and sparsity were tested for learning dictionary. Table VII lists the results of using different dictionary sizes, *i.e.*, the number of atoms in the dictionary. We can observe that the PLCC, SROCC and RMSE of SR-3DVQA are similar when the dictionary size range from 128 to 512. As the dictionary size becomes larger than 512, the performance decreases significantly. The reason is that in quality assessment, too large dictionary size may include atoms with unimportant details and lead to a slight performance degradation. Proper dictionary size is 128 to 512, which can represent the main features.

Table VIII shows the performance of SR-3DVQA when it has different sparsity in learning dictionary. We can find

TABLE VIII
IMPACTS OF THE SPARSITY IN DICTIONARY LEARNING

Indices	3	6	9	12	15	18
PLCC	0.911	0.910	0.907	0.905	0.898	0.895
SROCC	0.913	0.914	0.912	0.911	0.900	0.898
RMSE	0.053	0.053	0.054	0.055	0.056	0.057

TABLE IX
IMPACTS OF DIFFERENT PATCH VARIANCE THRESHOLD g

Indices	0	1	3	5	7	9	20	50	100
PLCC	0.757	0.903	0.908	0.910	0.911	0.910	0.893	0.890	0.885
SROCC	0.706	0.895	0.910	0.914	0.915	0.915	0.897	0.891	0.887
RMSE	0.084	0.055	0.054	0.053	0.053	0.053	0.058	0.059	0.060

TABLE X
IMPACTS OF DIFFERENT B IN EDGE PATCH DETERMINATION

Indices	1	2	4	8	16
PLCC	0.910	0.909	0.900	0.878	0.830
SROCC	0.914	0.912	0.903	0.885	0.832
RMSE	0.053	0.054	0.056	0.061	0.072

TABLE XI
COMPARISON ON WITH AND WITHOUT GRADIENT FEATURES AND SPARSE REPRESENTATION IN THE PROPOSED SR-3DVQA

Indices	w/o gradient features	w/o sparse representation	SR-3DVQA
PLCC	0.822	0.877	0.910
SROCC	0.801	0.884	0.912
RMSE	0.073	0.062	0.053

that the PLCC and SROCC decrease slightly as the sparsity increases. They achieve their peaks when the sparsity is 3 or 6, which means 3 or 6 atoms out of the dictionary are enough to capture the key features of the flicker distortion in feature representation. More atoms bring no benefits in characterizing the features.

2) Thresholds g and B in Flicker Detection: Since patch variances usually indicate textures importance in quality assessment, the patch variance threshold g is used to exclude the less important patches. Threshold B is to determine whether the current patch is edge or not in (5). To further analyze their impacts, different g and B were tested in the proposed SR-3DVQA. Table IX and X show the SROCC, PLCC and RMSE of the SR-3DVQA when it has different g and B, respectively. According to Table IX, the PLCC and SROCC indices increase when g varies from 0 to 7 and decrease when g varies from 9 to 100. Small g includes more patches as flicker and even some non-flicker patches may be falsely included, while large g leads to missing out some important flicker patches. A reasonable range for g is from 3 to 9 according to the experimental results. In addition, the PLCC and SROCC decrease as B increases according to Table X. More patches will be determined as edge patch when B becomes smaller, which makes the SR-3DVQA more sensitive in capturing the flicker distortion.

G. Impacts of Gradient Feature and Sparse Representation

To testify the contributions of using gradient features and sparse representation in representing the temporal layers, two extensive experiments were performed to test the SR-3DVQA if without using gradient features or without using the sparse representation, respectively. In the first case, the gradient feature extraction was removed in Fig.2 directly. In the second case, the sparse representation was not used, where the phase and amplitude distortions were computed directly on the gradient features. Table XI shows the performances of SR-3DVQA when it is without gradient feature and sparse representation, respectively. We can observe that the PLCC, SROCC and RMSE are 0.822, 0.801, and 0.073, respectively, when without using the gradient feature. In addition, they are 0.877, 0.884, and 0.062, respectively, when without using the sparse representation. These values are significantly inferior to those of the proposed SR-3DVQA, which indicates the gradient feature and sparse representation are effective and significantly contribute to the synthesized video quality prediction.

VI. CONCLUSION

In this paper, we propose a full-reference Sparse Representation based 3D View Quality Assessment (SR-3DVQA) for 3D synthesized view. Since the flicker distortion in synthesized video could be embodied in temporal and textural regions, it is measured by structuring the synthesized video into temporal layers. Moreover, we utilize the textural gradient and corresponding depth map to locate the possible flicker distortion. Sparse representation is used to represent the features difference between the flicker patches and non-flicker patches. The flicker distortion is then obtained by a layer pooling method. Finally, the flicker distortion measurement is combined with spatio-temporal activity measurement to build the SR-3DVQA. Extensive experiments on the SIAT database demonstrate our proposed SR-3DVQA is significantly superior to the state-of-the-art benchmark methods.

REFERENCES

[1] 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.
[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Conf. Rec. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2., Nov. 2003, pp. 1398–1402.
[3] G. Zhai, W. Zhang, X. Yang, W. Lin, and Y. Xu, "Efficient image deblocking based on postfiltering in shifted windows," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 122–126, Jan. 2008.
[4] 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.
[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] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5291, pp. 93–104, May 2004.

[7] E. Bosc *et al.*, "Towards a new quality metric for 3D synthesized view assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1332–1343, Nov. 2011.

[8] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, "A quality assessment protocol for free-viewpoint video sequences synthesized from decompressed depth data," in *Proc. 5th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 100–105.

[9] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Process., Image Commun.*, vol. 30, pp. 78–88, Jan. 2015.

[10] L. D. Li, Y. Zhou, K. Gu, W. S. Lin, and S. Q. Wang, "Quality assessment of DIBR-synthesized images by measuring local geometric distortions and global sharpness," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 914–926, Apr. 2018.

[11] G. Yue, C. Hou, K. Gu, T. Zhou, and G. Zhai, "Combining local and global measures for DIBR-synthesized image quality evaluation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2075–2088, Apr. 2019.

[12] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Multi-scale synthesized view assessment based on morphological pyramids," *J. Elect. Eng.*, vol. 67, no. 1, pp. 3–11, 2016.

[13] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR synthesized image quality assessment based on morphological wavelets," in *Proc. IEEE 7th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, May 2015, pp. 1–6.

[14] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "DIBR-synthesized image quality assessment based on morphological multi-scale approach," *EURASIP J. Image Video Process.*, vol. 1, pp. 1–23, Mar. 2017.

[15] K. Gu, V. Jakhetiya, J.-F. Qiao, X. Li, W. Lin, and D. Thalmann, "Model-based referenceless quality metric of 3D synthesized images using local image description," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 394–405, Jan. 2018.

[16] K. Gu, J. Qiao, S. Lee, H. Liu, W. Lin, and P. Le Callet, "Multiscale natural scene statistical analysis for no-reference quality evaluation of dibr-synthesized views," *IEEE Trans. Broadcast.*, to be published. doi: 10.1109/TBC.2019.2906768.

[17] V. Jakhetiya, K. Gu, W. Lin, Q. Li, and S. P. Jaiswal, "A prediction backed model for quality assessment of screen content and 3-D synthesized images," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 652–660, Feb. 2018.

[18] V. Jakhetiya, K. Gu, T. Singhal, S. C. Guntuku, Z. Xia, and W. Lin, "A highly efficient blind image quality assessment metric of 3-D synthesized images using outlier detection," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4120–4128, Jul. 2019. doi: 10.1109/TII.2018.2888861.

[19] E. Ekmekcioglu, S. Worrall, D. De Silva, A. Fernando, and A. M. Kondoz, "Depth based perceptual quality assessment for synthesised camera viewpoints," in *User Centric Media* (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), vol. 60. Berlin, Germany: Springer-Verlag, 2012, pp. 76–83.

[20] M. S. Farid, M. Lucenteforte, and M. Grangetto, "Evaluating virtual image quality using the side-views information fusion and depth maps," *Inf. Fusion*, vol. 43, pp. 47–56, Sep. 2018.

[21] H.-C. Liu and H.-M. Hang, "Quality assessment of synthesized 3D video with distorted depth map," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2014, pp. 1054–1057.

[22] M. Solh, G. AlRegib, and J. M. Bauza, "3VQM: A vision-based quality measure for DIBR-based 3D videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2011, pp. 1–6.

[23] Y. Zhao and L. Yu, "A perceptual metric for evaluating quality of synthesized sequences in 3DV system," *Proc. SPIE*, vol. 7744, Aug. 2010, Art. no. 77440X.

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

[25] F. Shao, Q. Yuan, W. Lin, and G. Jiang, "No-reference view synthesis quality prediction for 3-D videos based on color–depth interactions," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 659–674, Mar. 2018.

[26] X. Liu, Y. Zhang, S. Hu, S. Kwong, C.-C. J. Kuo, and Q. Peng, "Subjective and objective video quality assessment of 3D synthesized views with texture/depth compression distortion," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4847–4861, Dec. 2015.

[27] Y. Zhang, X. Yang, X. Liu, G. Jiang, and S. Kwong, "High-efficiency 3D depth coding based on perceptual quality of synthesized video," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5877–5891, Dec. 2016.

[28] L. He, D. Tao, X. Li, and X. Gao, "Sparse representation for blind image quality assessment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1146–1153.

[29] P. M. Shabeer, S. Bhati, and S. S. Channappayya, "Modeling sparse spatio-temporal representations for no-reference video quality assessment," in *Proc. IEEE Global Conf. Signal Inform. Process. (GlobalSIP)*, Nov. 2017, pp. 1220–1224.

[30] F. Zhang, W. Jiang, F. Autrusseau, and W. Lin, "Exploring V1 by modeling the perceptual quality of images," *J. Vis.*, vol. 14, no. 1, Jan. 2014.

[31] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Full-reference quality assessment of stereoscopic images by learning binocular receptive field properties," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 2971–2983, Oct. 2015.

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

[33] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "Learning sparse representation for objective image retargeting quality assessment," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1276–1289, Apr. 2018.

[34] Y. Zhang, S. Kwong, L. Xu, S. Hu, G. Jiang, and C.-C. J. Kuo, "Regional bit allocation and rate distortion optimization for multiview depth video coding with view synthesis distortion model," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3497–3512, Sep. 2013.

[35] Y. Zhang, S. Kwong, S. Hu, and C.-C. J. Kuo, "Efficient multiview depth coding optimization based on allowable depth distortion in view synthesis," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4879–4892, Nov. 2014.

[36] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 27–35, Jan. 2009.

[37] H. Xiong, Z. Pan, X. Ye, and C. W. Chen, "Sparse spatio-temporal representation with adaptive regularized dictionary learning for low bit-rate video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 710–728, Apr. 2013.

[38] S. C. W. Tim, M. Rombaut, and D. Pellerin, "Dictionary of gray-level 3D patches for action recognition," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2014, pp. 1–6.

[39] Y. Zhao, C. Zhu, Z. Chen, D. Tian, and L. Yu, "Boundary artifact reduction in view synthesis of 3D video: From perspective of texture-depth alignment," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 510–522, Jun. 2011.

[40] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[41] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Conf. Rec. 27th Asilomar Conf. Signals, Syst. Comput.*, vol. 1, Nov. 1993, pp. 40–44.

[42] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *IEEE Trans. Broadcast.*, vol. 62, no. 2, pp. 446–456, Jun. 2016.

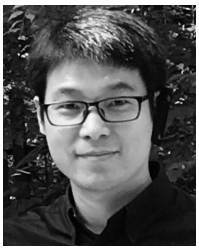
[43] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[44] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[45] L. Li, D. Wu, J. Wu, H. Li, W. Lin, and A. C. Kot, "Image sharpness assessment by sparse representation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1085–1097, Jun. 2016.

[46] *Research on Image Quality Assessment*. Accessed: Jun. 2019. [Online]. Available: http://sse.tongji.edu.cn/linzhang/IQA/IQA.htm

[47] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.



Yun Zhang (M'12–SM'16) received the B.S. and M.S. degrees in electrical engineering from Ningbo University, Ningbo, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2010. From 2009 to 2014, he was a Postdoctoral Researcher with the Department of Computer Science, City University of Hong Kong, Hong Kong. From 2010 to 2017, he was an Assistant Professor and an Associate Professor at the Shenzhen Institutes of Advanced Technology (SIAT), CAS, Shenzhen, China, where he is currently a Professor. His research interests are video compression, 3D video processing, and visual perception.

Huan Zhang received the B.S. degree from the Civil Aviation University of China, Tianjin, China, in 2010, and the M.S. degree from Tsinghua University, Beijing, China, in 2013. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences, China. Her research interests include image restoration and image/video quality assessment.

Mei Yu received the M.S. degree from the Hangzhou Institute of Electronics Engineering, Hangzhou, China, in 1993, and the Ph.D. degree from Ajou University, South Korea, in 2000. Then, she joined Ningbo University, Ningbo, China, where she has been a Professor with the Faculty of Information Science and Engineering, since 2005. Her research interests include image/video coding and video perception.

Sam Kwong (F'13) received the B.S. degree from the State University of New York at Buffalo in 1983, the M.S. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1985, and the Ph.D. degree from the University of Hagen, Germany, in 1996. From 1985 to 1987, he was a Diagnostic Engineer with Control Data Canada. He joined Bell Northern Research Canada as a member of the Scientific Staff. In 1990, he became a Lecturer at the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, where he is currently a Professor at the Department of Computer Science. His research interests are video and image coding and evolutionary algorithms.

Yo-Sung Ho (SM'06–F'16) received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, South Korea, in 1981 and 1983, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California at Santa Barbara, in 1990. He joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, in 1983. From 1990 to 1993, he was with North America Philips Laboratories, Briarcliff Manor, NY, USA, where he was involved in the development of the Advanced Digital High-Definition Television (AD-HDTV) system. In 1993, he rejoined the technical staff of ETRI and was involved in development of the Korean DBS Digital Television and High-Definition Television Systems. Since 1995, he has been with the Gwangju Institute of Science and Technology (GIST), where he is currently a Professor of the School of Electrical Engineering and Computer Science. Since August 2003, he has been the Director of the Realistic Broadcasting Research Center, GIST, South Korea. His research interests include digital image and video coding, image analysis and image restoration, three-dimensional image modeling and representation, advanced source coding techniques, augmented reality (AR) and virtual reality (VR), three-dimensional television (3DTV), and realistic broadcasting technologies. He has served as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA (T-MM) and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS VIDEO TECHNOLOGY (T-CSVT).