IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 12, 2019, accepted March 12, 2019, date of publication March 26, 2019, date of current version April 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2907572

Region-of-Interest Compression and View Synthesis for Light Field Video Streaming

BING WANG[1,2], QIANG PENG[1], ERIC WANG[2], KANG HAN[2], AND WEI XIANG[2], (Senior Member, IEEE)

[1]School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[2]College of Science and Engineering, James Cook University, Cairns, QLD 4878, Australia

Corresponding author: Wei Xiang (wei.xiang@jcu.edu.au)

This work was supported in part by the Beijing Natural Science Foundation under Grant L182032 and in part by the National Natural Science Foundation of China (Grant 61772436), Foundation for Department of Transportation of Henan Province (2019J-2-2). The work of B. Wang was supported by the China Scholarship Council (CSC) under Grant 201707000093.

ABSTRACT Light field videos provide a rich representation of real-world, thus the research of this technology is of urgency and interest for both the scientific community and industries. Light field applications such as virtual reality and post-production in the movie industry require a large number of viewpoints of the captured scene to achieve an immersive experience, and this creates a significant burden on light field compression and streaming. In this paper, we first present a light field video dataset captured with a plenoptic camera. Then a new region-of-interest (ROI)-based video compression method is designed for light field videos. In order to further improve the compression performance, a novel view synthesis algorithm is presented to generate arbitrary viewpoints at the receiver. The experimental evaluation of four light field video sequences demonstrates that the proposed ROI-based compression method can save 5%–7% in bitrates in comparison to conventional light field video compression methods. Furthermore, the proposed view synthesis-based compression method not only can achieve a reduction of about 50% in bitrates against conventional compression methods, but the synthesized views can exhibit identical visual quality as their ground truth.

INDEX TERMS Light field, video compression, region-of-interest, view synthesis, light field video dataset.

I. INTRODUCTION

Traditional applications of images and videos are limited to a single viewpoint of scenes. By contrast, a light field offers multiple viewpoints by sampling a huge number of light rays. Over the past decade, light fields have attracted tremendous attention due to their capability to represent 3D information of the environment. Light field technologies provide a rich representation of real-world scenes and have been popularly adopted by a wide range of industries. Light field data can be captured by light field cameras with a microlens array. These cameras can capture the distribution of light rays in free space, enabling exciting applications such as refocusing and viewpoint change. The most popular commercialized light field cameras are Lytro Illum [1] and Raytrix Plenoptic Camera [2]. In April 2014, Lytro Inc. announced Lytro Illum which uses microlens array technology to capture light field images in one camera. The Illum focuses on consumer markets and is designed to attract users with the concept of capturing first and refocusing later. Raytrix released Ratrix Plenoptic Camera which provides 3D high-speed video capture and enhanced depth of field. Light field data can also be captured with a camera array which is cumbersome and expensive compared with the microlens-based plenoptic light field cameras.

The existing light field datasets, [3]–[7], from real-world light fields captured with a plenoptic camera (e.g., Lytro Illum), real-world scenes captured with a camera array or a synthetic light field are light field image datasets. The only two exceptions are the light field video datasets recently proposed by Daḅała *et al.* [8] and Sabater *et al.* [9], which used camera arrays to capture light field videos. However, no work has yet attempted to create a light field video dataset captured with plenoptic cameras. In this paper, a light field video dataset captured with a Lytro Illum is created. Along with this paper, the proposed light field video dataset is published at [10] which we believe may be of special interest to the community.

The fifth generation mobile network (5G) is making a huge impact on multimedia applications. Although 5G certainly

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

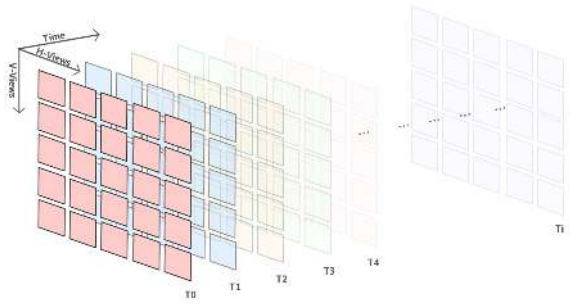


FIGURE 1. 3D matrix structure of a light field video.

provides the ability to have a larger bandwidth and higher throughput, there is still a huge demand for compression methods of larger-volume multimedia data, for example, light field data. A light field video sequence is a 3D matrix, with the three dimensions as horizontal views (H-Views), vertical views (V-Views) and time, respectively, as shown in Fig. 1. A light field video sequence records a scene with a set of streams from different viewpoints, thus exhibiting data redundancy in both spatial and temporal dimensions. In terms of applications, the sheer size of the data volume of a light field video brings new challenges to the efficient storage and transmission of this massive amount of complex data. Therefore, light field compression is a critical aspect of the practical usage of light field technologies. Raw data from light field cameras exhibit strong correlations, so there is a significant amount of research [11]–[14] that has been conducted on light field image compression by reducing these spatial redundancies. A light field video contains hundreds of times more pixels than a traditional monocular video sequence, and light field data exhibits a more complicated and unique structure. Therefore, despite the large volume of research in light field image coding and traditional video coding, using these methods to process light field videos is inadequate.

Light field videos not only have high spatial correlations in each light field image frame but also indicate strong correlations among continuous video frames. Thus inter micro-images correlations of each frame and inter-frame correlations of each view should be combined in order to acquire an efficient compression method for light field video coding. In the field of video processing, a light field video sequence can be represented as a two-dimensional multi-view sequence with both horizontal and vertical parallax. And because of this fact, there are many light field video coding methods proposed in the literature to date [15]–[18]. One direct coding method is to encode each view separately, but the strong correlations between views are ignored. Converting the two-dimensional multi-view video sequence into a one-dimensional video sequence with a horizontal zigzag order for light field video compression is proposed in [15]. In another paper [16], a rotary order which scans from the centre and revolves around the view until the final view of the last frame is proposed to improve the compression ratio for light field data. However, the two biggest shortcomings of these methods are that the vertical correlations between views and temporal correlations between frames are not fully utilized. Wang *et al.* [17] proposed a light field multi-view videos coding (LF-MVC) method by extending the inter-view prediction multi-view video coding (MVC) [18] structure into a two-directional parallel structure. However, this work just focused on analyzing the relationship of the prediction structure with its coding performance, and the biggest obstacle of LF-MVC is that this method only can be implemented in H.264-based multi-view coding standard (not compatible with HEVC).

All the above-mentioned compression methods for light field videos do not take the importance of some regions of a video scene into account. In most cases, a moving region of a scene attracts more attention than other areas. Therefore, considering motion characteristics of videos, a region-of-interest (ROI) based compression algorithm for light field videos is presented. Considering strong inter-view correlations of light field videos, a view synthesis algorithm is presented to reduce the costs of video processing, transmission and storage. Furthermore, the view synthesis algorithm can be used in some interactive or selective applications to achieve uneven views' transmission and compression. In summary, we make the following contributions:

- A dataset of four light field videos is captured and presented by a Lytro Illum camera. To our knowledge, this is the first light field video dataset which is captured with a plenoptic camera, and the proposed dataset can be downloaded from [10];
- Four popular light field video compression methods are implemented to benchmark the compression performance of the captured light field video sequences;
- A new ROI-based video compression method is proposed to improve light field video coding efficiency by 5%–7% compared with the other four popular light field compression methods;
- A novel view synthesis method for light field videos is proposed to realize low-latency real-time light field video streaming, which is able to achieve a significantly higher coding performance, and the synthesized views are close to their ground truth.

The rest of the paper is organized as follows. The proposed light field video dataset is introduced in Section II. The conventional coding methods and a new ROI-based compression method for light field videos are illustrated in Section III. The proposed view synthesis for steamed light field video is presented in Section IV. Experimental results are given in Section V. Concluding remarks and future work are provided in Section VI.

II. LIGHT FIELD VIDEO DATASET

Lytro and Raytrix are the two most important light field cameras manufacturers in the market today. Due to the complexity in light filed video processing, only the more expensive Raytrix light field camera can capture real-time light field videos, while the lower end Lytro light field camera only has

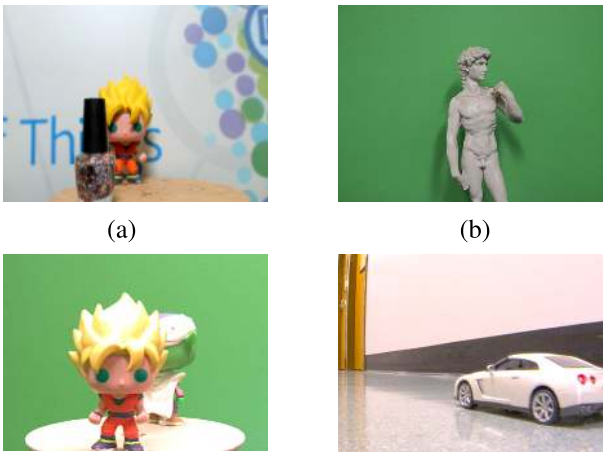


FIGURE 2. Example frames of the proposed light field video dataset: (a) *Bottle*, (b) *David*, (c) *Toys* and (d) *Car*.

the capability of capturing light field images. In consideration of real-world applications, we use a Lytro Illum camera to take several continuous still light field images and manually generate light field video sequences.

A. LIGHT FIELD VIDEO CAPTURE AND PRE-PROCESSING

Four scenes have been captured using a Lytro Illum camera (ISO is 3200, flicker reduction is 50Hz) to take several continuous still light field images and manually generate Light field video sequences. In order to generate light field videos, raw light field images captured by the Lytro camera are first decoded, calibrated and rectified by Matlab LFToolbox (Version 0.4) [19]. As is commonly carried out in video coding, each view sequence is then transformed to a YUV video sequence. Each light field image acquired with the Lytro Illum camera is represented by a 4D matrix of 15×15 sub-aperture images/views. However, the viewing quality of the border views are usually distorted and darksome, especially the three views from each side are usually black. To mitigate this problem, all the 15×15 views are refined and corrected by a frequency filter and a color correction method in this paper.

These videos are: *Bottle*, *David*, *Toys* and *Car*, as shown in Fig. 2. Our dataset has two green background close-ups sequences (*David*, *Toys*) that are interesting for some specific use cases such as realistic telepresence and face 3D recovering. We have also captured a rotary scene that includes two small objects with a complex background (*Bottle*), and the fourth light field video scene includes a car movement (*Car*).

B. DESCRIPTION OF THE CAPTURED LIGHT FIELD VIDEO SEQUENCES

The proposed light field video dataset can be downloaded from [10]. All light field video sequences in this dataset are provided as 15×15 views in YUV format. The frame rate of each light field video sequence in our dataset is 25 frames per second, and each view has 100 frames. The detailed description of each light field video sequence is as followed:

1) BOTTLE SEQUENCE

The video sequence *Bottle* contains one bottle and one toy on a turn table and with a poster as background. The resolution is 512×352.

2) DAVID SEQUENCE

The video sequence *David* contains one David sculpture on a turn table and with a green background. The resolution is 480×320.

3) TOYS SEQUENCE

The video sequence *Toys* contains 2 toys on a turn table and with a green background. The resolution is 480×320.

4) CAR SEQUENCE

The video sequence *Car* contains one car, and the car is moving from right to left at a constant speed. The resolution is 512×352.

III. VIDEO COMPRESSION METHOD FOR LIGHT FIELD VIDEOS

In the following part of this section, we first apply four conventional 2D/3D video coding methods to light field video coding, and then a novel ROI-based compression method for light field videos is proposed. The experimental results are shown in Section V.

A. CONVENTIONAL COMPRESSION METHODS FOR LIGHT FIELD VIDEOS

Taking 3×3 array as an example, four popular video compression methods are converted to light field array structure in order to further experimental analysis.

A straightforward way to compress a light field video sequence is to encode each view sequence separately, like a regular 2D video sequence. For example, every single view can be encoded using the hierarchical-B coding structure, in which the first frame of each view is encoded as I frames, and the remaining frames are predictively encoded as B frames. This method is straightforward and easy to implement, but only the inter-frame correlation between frames in each single view sequence is considered while the strong correlation between views is disregarded.

It is noted that a light field video can be considered as a 3D image matrix. By transposing this 3D image matrix, we can obtain a single video sequence and compress it with high-efficiency video coding methods. A horizontal zigzag transposed compression method is proposed in [15]. It transposes the first from top-left to bottom-right, then the second light field frame from bottom-right to top-left, and so on, and construct a single 2D video sequence with all the transposed images, as shown in Fig. 3. For example, a 3×3 views and 5 frames light field video sequence can be transposed into a 45 frames single view sequence. This method makes use of the inter-view correlation, but in contrast to the single view compression method, the inter-frame correlation no longer exists in the final reconstructed video sequence.

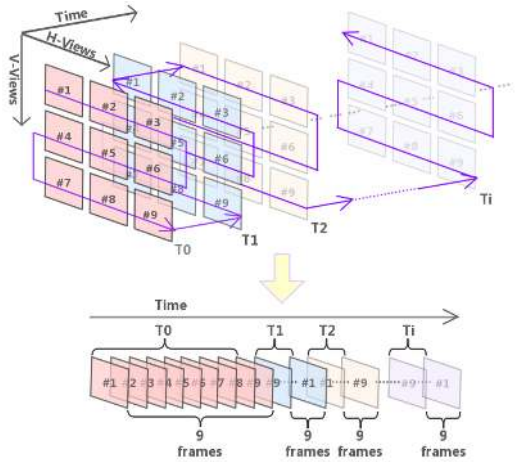


FIGURE 3. Horizontal zigzag transposed ordering coding structure for a light field video sequence.

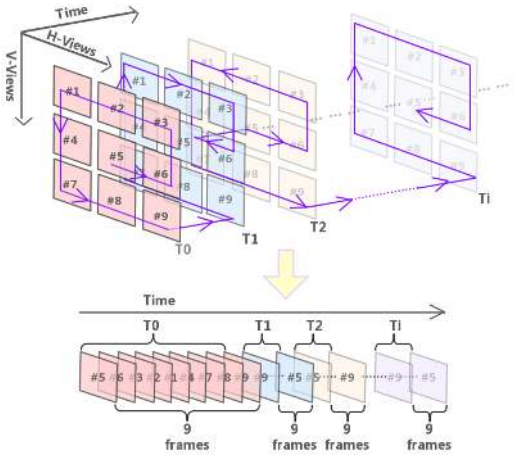


FIGURE 4. Rotary transposed ordering coding structure for a light field video sequence.

A rotary transposed ordering compression method for light field images is proposed in [16]. We carried out this method to compress light field videos by adding a time series into the structure. It transposed the first light field frame from the centre and revolved around the view until the final view of the last frame, as shown in Fig. 4. Rotary transposed ordering structure makes better use of the inter-view correlation when the number of views is small compared with horizontal zigzag transposed ordering structure, but more views will bring larger rotation and frames get father from their references. The primary problem of this compression method is the inter-frame correlation is still not considered in this structure.

A light field video can be seen as a multi-view video vector, and then the standard MVC coding structure can be implemented to compress a light field video sequence by making use of both inter-view and inter-frame correlations. To implement multi-view coding in light field video coding, a simple ordering operation can be applied to convert the light field video sequence array into a video sequence vector. For example, a two-dimensional (H-views and V-views) light field video sequence is converted to a one-dimensional multi-view video sequence, as shown in Fig. 5. MVC for light field video coding exploits both the spatial and temporal redundancy contained in a light field video sequence, which can achieve a much better compression ratio compared with the single view compression method and transposed ordering (horizontal zigzag and rotary) methods. However, through analyzing the extended MVC structure, we can find that the vertical correlation disappears after the image array-vector conversion. On the basis of the MVC prediction structure, a light field multi-view coding (LF-MVC) structure is proposed in [17]. This method makes use of inter-view correlations, inter-frame correlations and vertical correlations between frames, as shown in Fig. 6. However, the biggest obstacle of LF-MVC is that this method only can be implemented in the H.264-based multi-view coding standard.

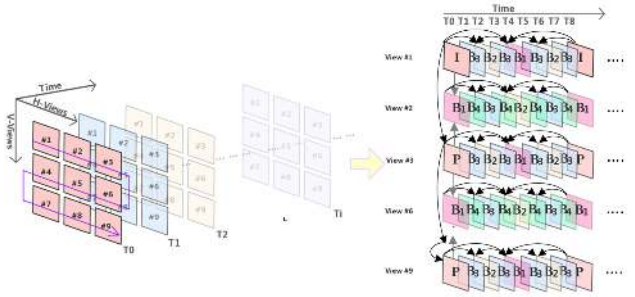


FIGURE 5. Multi-view video coding structure for a light field video sequence.

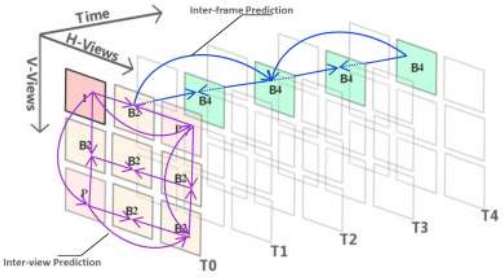


FIGURE 6. LF-MVC coding structure for a light field video sequence.

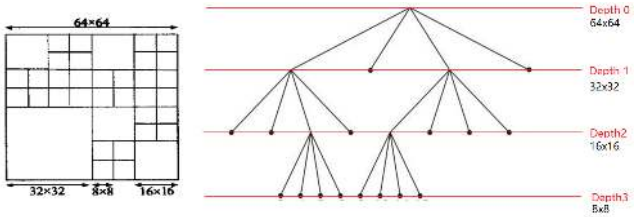


FIGURE 7. Flexible quadtree structure of HEVC standard.

B. A NEW ROI-BASED LIGHT FIELD VIDEO COMPRESSION ALGORITHM

ROI-based coding methods which consider the importance of some regions of videos are widely used to reduce bitrates

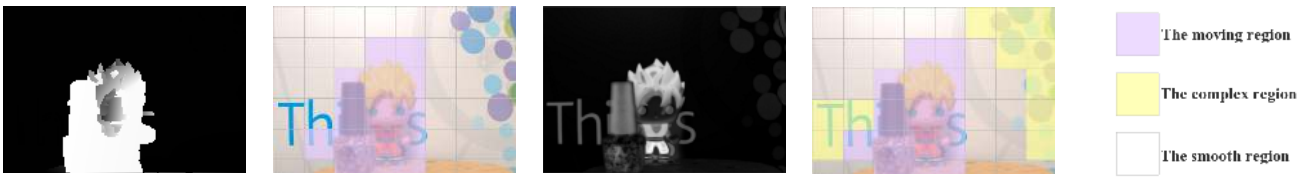


FIGURE 8. Results of the proposed region division method (The centre view's first frame of *Bottle* sequence, and the largest size of coding units is 64 × 64.): (a) Moving region distribution map, (b) Moving region division result, (c) Texture distribution map, (d) Non-ROI sub-division result.

and keep an excellent perceptual quality of videos. However, there are no works using ROI technology to guide light field video coding. In most cases, a moving region of a video frame attracts more attention than other areas. For example, the most important areas of the light field video sequences in the proposed dataset and other two light field video datasets [8], [9] are moving regions. Therefore, in order to further improve coding efficiency and compression ratio, we propose an ROI-based video compression method for light field videos that based on flexible quadtree structure of the high-efficiency video coding (HEVC) standard. The most important thing is that the proposed method can be combined with any HEVC-based compression methods.

As shown in Fig. 7, different from the fixed size macroblock of H.264, HEVC standard uses quadtree-based variable-size coding units. The largest size of coding units is 64 × 64 pixels, and the smallest size is 8 × 8. There are four depth levels (*0, 1, 2, 3*) to divide the size of coding units by doing in a recursive manner. A depth level of zero means that the size of the current coding unit is 64 × 64, and a few bits are allocated to this unit. On the contrary, a depth level of three means that the current coding unit is partitioned into four 8 × 8 sub-units, and many bits are allocated to ensure the quality of this unit. The size of a coding unit in HEVC is various depending on the complexity of the video content. In order to decide the most optimal partition mode, HEVC standard compares the coding cost of units at current depth level and the sum of four coding costs of units at other depth levels. The process time of partitioning coding units accounts for around 40% of the whole coding time. However, the HEVC standard allows users to set the depth level for each coding unit. We can use this option to save a lot of coding costs by limiting depth level of coding units in non-ROI regions which are not important in a video frame. According to the human visual system, humans can not notice all details in non-ROI regions, so the loss of some details caused by the limitation of depth level in non-ROI regions have only a small impact on users' viewing experience.

In our ROI-based video coding method, moving areas served as ROI regions, and non-moving areas served as non-ROI regions. Considering the small difference between views due to the dense and narrow-baseline of light field videos captured with a Lytro Illum camera, ROI detection methods only are implemented on the centre view of a light field video sequence in order to avoid extra computation time. We firstly use frame differencial method to detect motion regions and generate moving region maps. The difference between frames for each pixel is calculated. The largest size of coding units in our algorithm is 64 × 64, and the coding unit is marked as a moving region if it contains moving pixels. Results of a generated moving region distribution map and a moving region division result are shown in Figs. 8 (a) and (b). Moreover, through our observations, the characteristics of coding units inside a non-ROI region may not be uniform. Therefore, non-ROI regions in our research are further divided into smooth regions and complex regions. Smooth regions are easy to encode and only needs a few numbers of bits, and complex regions need more bits to ensure their perspective quality. In order to complete the division of non-ROI, a texture detection method is used to obtain texture maps. The essential idea behind the method is to calculate Euclidean distance between the *Lab* pixel vector in a Gaussian filter for each frame with the average *Lab* vector for each frame. Results of a texture map and non-ROI sub-division are shown in Figs. 8 (c) and (d).

The generated moving region maps and texture maps are then used to guide coding units partition and bitrate allocation. On the basis of these maps, we add a constraint on coding unit partition strategy, and the partition size is determined by (1).

$$depth(cu) = \begin{cases} \text{standard partition, } cu \in \text{ROI} \\ 1, \; cu \in \text{complex region of non-ROI} \\ 0, \; cu \in \text{smooth region of non-ROI} \end{cases} \quad (1)$$

where *cu* is the current coding unit. The standard recursive manner is used to determine the unit partition if the unit belongs to the ROI region. Otherwise, a constraint on the depth level of the coding unit is added in the non-ROI region. The depth level is one (two 32 × 32 sub-units) if the unit belongs to the complex region, or the depth level is zero (one 64 × 64 sub-unit) if the unit belongs to the smooth region.

IV. VIEW SYNTHESIS ALGORITHM FOR STREAMED LIGHT FIELD VIDEO

There is an inherent trade-off between angular and spatial resolutions in light field data because of the limited resolution of sensors. The light field videos captured by camera

array systems have a high spatial resolution but usually do not have many available views. For example, the light field videos in [8] and [9] only have 3 × 3 views and 4 × 4 views, respectively. On the other hand, the light field videos captured by plenoptic cameras have many available views at the cost of reducing the spatial resolution. For instance, the light field videos in our dataset are captured with a Lytro Illum camera have 15 × 15 available views, while the spatial resolution is only 434 × 625 and a limited range of field-of-view (FOV). To mitigate this problem, a learning-based approach is proposed in [20] to improve the light field spatial resolution. This algorithm synthesizes novel views from a sparse set of input views captured by the Lytro Illum camera. However, the primary issue in this system is that the authors designed disparity features to represent appearance flow instead of learning them from original light field images. Computer vision researchers have shown that learned features are generally much better than hand-designed features [21]. Niklaus *et al.* [22] propose an adaptive separable convolution approach for video frame interpolation. The authors approximate adaptive convolutional kernels by separating them to the vertical and horizontal kernels. This approximation greatly reduces the cost of computer memory so that the network can handle large motion between video frames. This deep neural network is fully convolutional and can be trained end-to-end using widely available video data without any difficult-to-obtain meta data like optical flow. However, this approach is designed for dealing with frame interpolation of 2D monocular video sequences.

Inspired by the above-mentioned algorithms, a view synthesis algorithm for light field video compression is proposed to improve coding efficiency. The essential idea behind combining the proposed view synthesis algorithm with light field video coding described in this section is that only a sparse set of light field views are encoded, transmitted and stored instead of processing all views. Obviously, compressing and transmitting only a sparse set of views and synthesizing other views according to the actual application is an effective way to improve coding efficiency. Fig. 9 shows the proposed view synthesis algorithm for streamed light field videos. There are 10 × 10 pink color views which need to be encoded in a 15 × 15 light field video sequence, and the other views (gray and white color views) can be synthesized at the decoder according to the actual application requirements. Traditional compression methods need coding 15 × 15 views for each light field video sequence, while the proposed view synthesis method can directly cut down approximately 53% of views to increase the compression ratio.

Considering narrow baseline structure and strong inter-view correlations exist in light field videos, we extend the time series of video frames interpolation [22] to the spatial series of light field views synthesis. The overview of the neural network architecture is shown in Fig. 10. The learning-based framework contains three convolutional neural networks. Firstly, a multi-scale convolutional neural network is used to estimate optical flow implicitly. Then the second

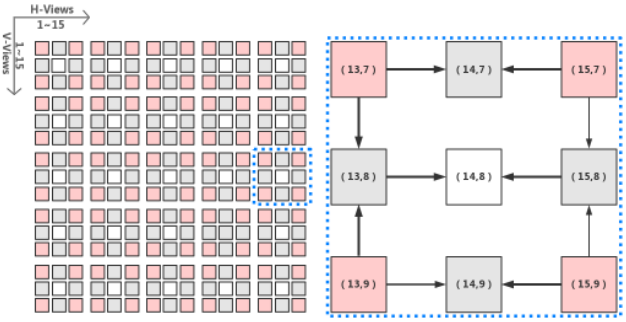


FIGURE 9. Proposed view synthesis method for light field videos. There are 15 × 15 views in the left picture. The pink color views (10 × 10) are key views which are encoded using a conventional HEVC codec, and the remaining views (gray and white color views) are synthesized in the decoder to improve coding efficiency. Gray color indicates this view is generated by key views, and white color indicates this view is yielded by synthesized views. The coordinates in the right picture show the locations of views in a light field video.

convolutional neural network is used to generate pixel-wise adaptive kernels. The generated kernels are then doing the convolution with the existing views to synthesize novel views. Finally, we adopt the third convolutional neural network to refine stacked synthesized views from adaptive convolution to get the final synthesized views. This deep neural network is fully convolutional and can be trained end-to-end. Since the objective of the model is to synthesize novel views, optical flow ground truth is not required to train the model. This objective also brings a benefit that accurate optical flow estimation is not necessary. Pixel-wise adaptive convolution is used to synthesize novel views, so the problem of view wrapping in large motion places can be avoided. Adaptive kernels contain optical flow information thus they does the view wrapping implicitly. Furthermore, these kernels have common patterns which mean common features of images, so the learning-based framework can synthesize more real and accurate novel views.

Compared with traditional methods which compress all views in light field videos, our approach reconstructs dense light field video from sparse views. The high correlations between neighboring views allow a deep neural network to reconstruct original light field videos accurately. The most significant difference between our approach and traditional methods is that the deep neural network extracts a common pattern from light field videos and leverages it to reconstruct dense light field videos for all the possible videos. The common pattern is a more efficient way to compress the data and can be seen as a kind of intelligence. With the great improvement of GPU's performance and the property of highly parallel of the deep neural network, compressed light field videos can be efficiently reconstructed at the receiver.

In some interactive or selective applications (e.g., [23]) where users can choose some views that interest them to be displayed, compressing and transmitting the whole light field video generate not only a very significant computation load but also an enormous encoding/decoding and transmission latency. The proposed view synthesis compression method

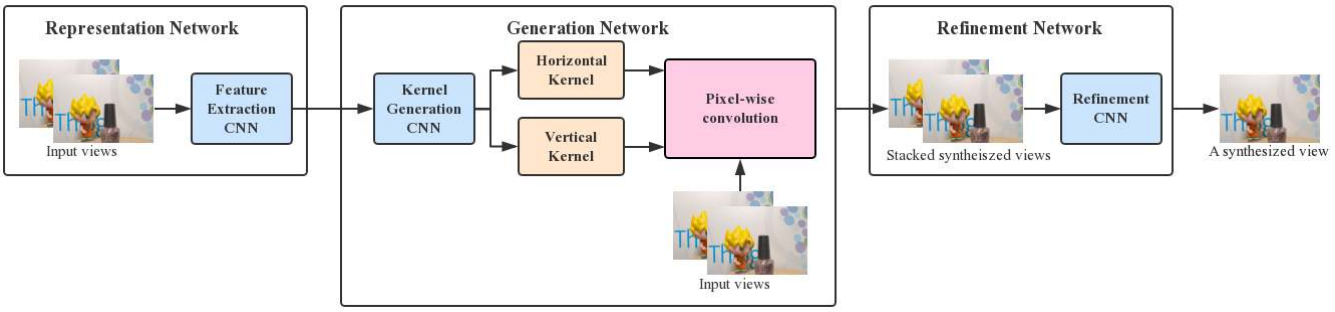


FIGURE 10. An overview of our neural network architecture. Given input frames, a multi-scale convolutional neural network extracts features that are given to horizontal and vertical kernels that each estimates one of four 1D kernels for each output pixel in a dense pixel-wise manner. The estimated pixel-wise kernels are then adaptive convolution with input views to yield initial stacked synthesized views. A refining process is used to refine the synthesized results from adaptive convolution to obtain an optimized synthesized view. The whole model is fully convolutional and can be trained end-to-end.

TABLE 1. Rate-Distortion Performance of Four Conventional Compression Methods.

Videos	MVC		LF-MVC		SV		HZT		RT	
	Bitrates [kbps]	PSNR [db]	Bitrates [kbps]	PSNR [db]	Bitrates [kbps]	PSNR [dB]	Bitrates [kbps]	PSNR [db]	Bitrates [kbps]	PSNR [db]
Bottle	620.046	34.87	507.267	34.88	706.281	35.32	261.390	37.09	260.480	37.08
David	786.115	35.41	634.588	35.59	968.324	35.57	296.506	37.54	289.615	37.48
Toys	660.817	34.97	532.077	35.06	775.911	35.37	269.583	37.37	266.175	37.37
Car	410.553	33.85	368.291	33.88	383.825	35.03	228.861	35.53	228.731	35.52
Average	**619.383**	**34.78**	**510.556**	**34.85**	**708.585**	**35.32**	**264.085**	**36.88**	**261.250**	**36.86**

can be used to solve this problem by allowing users to generate novel views from a sparse set of decompressed views at the receiver according to their interactive applications. Therefore, the proposed view synthesis method for light field video compression can significantly improve the coding efficiency and apply to low-latency real-time interactive light field streaming applications.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. EXPERIMENTAL RESULTS

Taking 3 × 3 light field video sequences as examples, experiments for above compression methods are carried out. The frame rate is 25 frames per second and quantization parameter (QP) 37 is adopted. To evaluate the compression efficiency, we use conventional rate-distortion and Bjøntegaard Delta (BD) rate [24] to evaluate the performance of compression methods.

Firstly, the rate-distortion performance of the four popular compression methods: Single view (SV) compression method, Horizontal zigzag transposed (HZT) ordering compression method, Rotary transposed (RT) ordering compression methods and LF-MVC for four light field videos in our dataset are given in Table 1. According to the results, the bitrate savings of LF-MVC relative to MVC under the same objective quality about 19.527% on average, because the vertical correlations of views are used in LF-MVC. The performance of HZT is similar to RT due to the strong inter-view correlations. However, the objective performance of RT is affected by the number of views because more views will bring larger rotation and frames get farther from their references.

TABLE 2. BD-Rate Reductions Obtained by Combining ROI Technology in Comparison to Conventional Methods.

Videos	BD-rate reduction [%] relative to		
	SV	HZT	RT
Bottle	-8.312	-6.530	-6.724
David	-5.438	-3.829	-3.176
Toys	-6.303	-4.297	-4.324
Car	-8.102	-5.231	-5.255
Average	**-7.039**	**-4.972**	**-4.869**

Moreover, Table 2 shows the BD-rate reductions obtained by the proposed ROI-based compression method to the SV, HZT and RT methods. According to Table 2, the proposed compression method can achieve a reduction of 4.869%-7.039% in bitrates against the conventional compression methods.

Table 3 shows the BD-rate reductions obtained by the proposed view synthesis-based compression strategy to the above four commonly used compression methods. According to the results, the proposed view synthesis-based compression strategy can achieve a reduction of about 50.811% in bitrates against the conventional compression methods.

B. EXPERIMENTAL RESULTS OF VIEW SYNTHESIS ALGORITHM FOR STREAMED LIGHT FIELD VIDEO

Compressing, transmitting and storing only a sparse set of views instead of processing all views is an effective way to improve the performance of light field video compression. Therefore, a learning-based view synthesis algorithm for streamed light field videos is proposed in Section IV. The precondition of combining view synthesis algorithm with

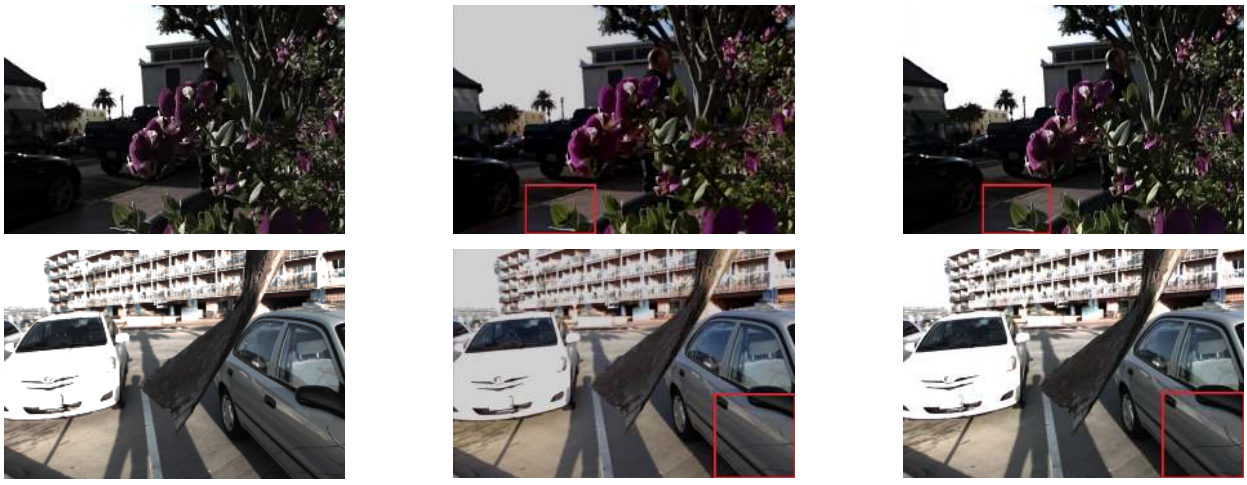


FIGURE 11. Comparison of our algorithm against the resent method of Kalantari *et al.* [20] on the *Flower1* and *Cars* scenes. The contents in the red boxes obviously show our synthesized views are better than Kalantari *et al.*'s synthesized views in terms of the quality of edges and local details.

TABLE 3. BD-Rate Reductions Obtained by the Proposed View Synthesis-Based Compression Strategy in Comparison to Conventional Methods.

Videos	BD-rate reduction [%] relative to			
	LF-MVC	SV	HZT	RT
Bottle	-48.269	-53.000	-52.092	-52.768
David	-47.312	-53.000	-51.775	-52.018
Toys	-47.694	-53.000	-53.167	-52.932
Car	-49.662	-53.000	-52.284	-53.007
Average	**-48.234**	**-53.000**	**-52.330**	**-52.681**

TABLE 4. Objective Comparison of Our View Synthesis Algorithm Against the State-of-the-Art Method [20].

Images	Kalantari *et al.* [20]		Ours	
	SSIM	PSNR	SSIM	PSNR
Flower 1	0.969	33.31	0.974	34.01
Flower 2	0.959	31.93	0.974	34.35
Cars	0.966	31.65	0.971	32.64
Rock	0.970	34.67	0.969	34.85
Leaves	0.936	27.80	0.948	30.14
Average	**0.960**	**31.87**	**0.967**	**32.20**

video coding is to make sure novel views can be synthesized correctly by a few input views. In order to verify the effectiveness of the proposed view synthesis algorithm, comparison experiments are done by using the existing popular view synthesis algorithm [20] and the proposed algorithm.

1) OBJECTIVE EXPERIMENTAL RESULTS FOR LIGHT FIELD VIEW SYNTHESIS ALGORITHM

We test the proposed light field view synthesis algorithm on microlens-based light field image dataset [20] in order to compare the proposed algorithm against the state-of-the-art algorithm available in the literature [20]. The experimental results are evaluated numerically, in terms of the PSNR and structural similarity (SSIM) [25]. SSIM produces a value between 0 and 1, where 1 indicates perfect perceptual quality with respect to the ground truth.

Table 4 shows the PSNR and SSIM values for these two methods on different test scenes. As can be seen in the table, the results of our algorithm are better than another algorithm.

2) SUBJECTIVE EXPERIMENTAL RESULTS FOR LIGHT FIELD VIEW SYNTHESIS ALGORITHM

We first test our algorithm on *Flower1* and *Cars* scenes which are offered in microlens-based light field image dataset [20] to compare the subjective quality of the proposed view synthesis algorithm against Kalantari's algorithm [20], and the results are shown in Fig. 11. Ground truth is shown in Fig. 11 (a) as the judging criteria, and the contents within red boxes in Figs. 11 (b) and (c) show obviously subjective improvement of our synthesized views compared with another algorithm. In the *Flower1* scene, the leaves in the red box of the right image (our synthesized view) have better quality than the same area in the medium image (Kalantari *et al.*'s synthesized view). In the *Cars* scene, the tree branches in the red box of the right image are much clearer than the same area in the medium image. Therefore, this subjective contrast shows our algorithm is able to synthesize more details and much closer to the ground truth.

Then we implement our algorithm in the proposed light field video dataset to demonstrate its superiority and effectiveness. Taking four 3×3 light field video sequences which are belong to the proposed light field video dataset as examples, the subjective results are shown as Fig. 12. The results show no detectable differences between the ground truth and the synthesized view. Therefore, the proposed view synthesis-based light field video compression strategy is feasible, and views can be flexibly synthesized according to the specific application.

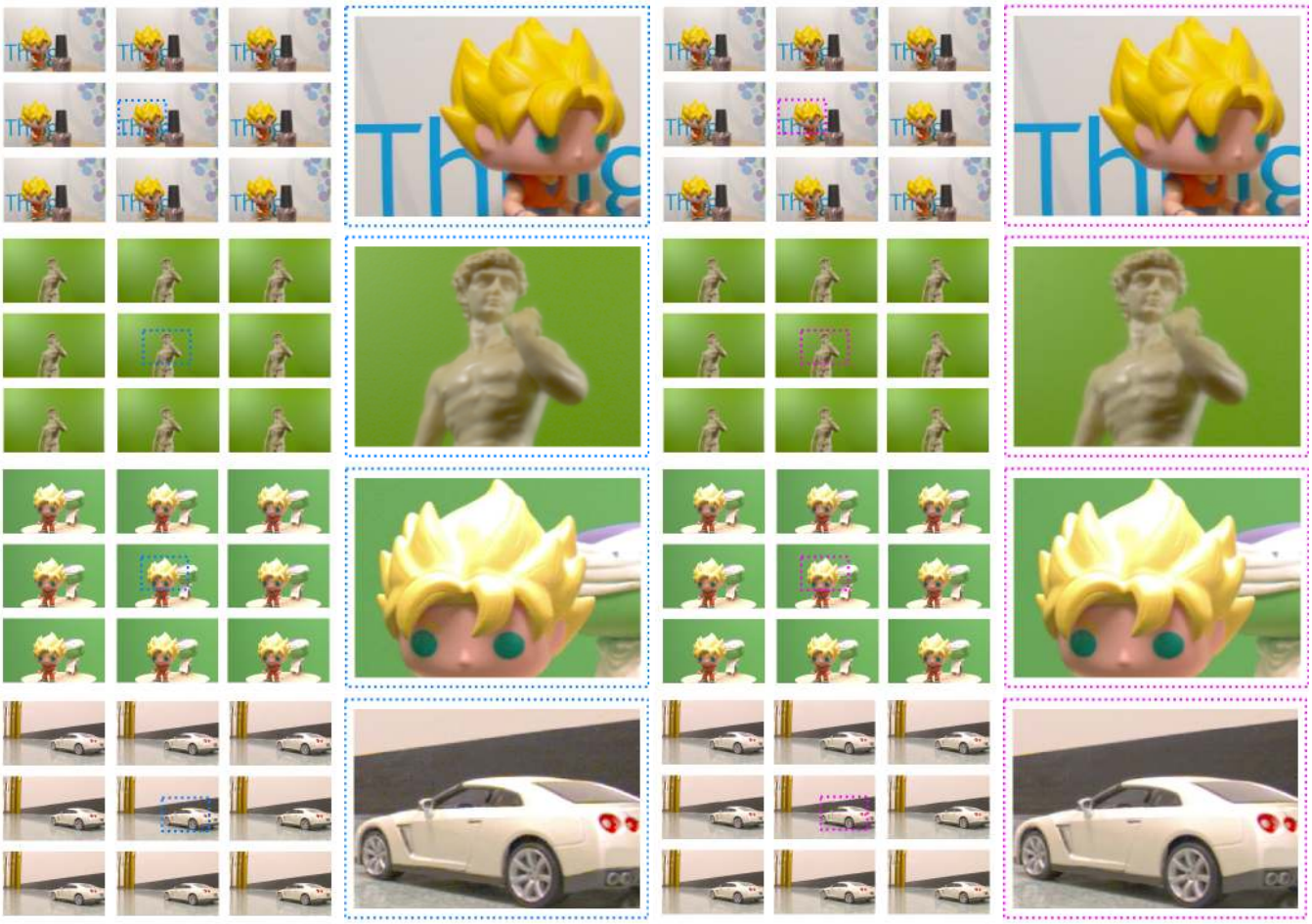


FIGURE 12. Subjective results of synthesized views. (a) Ground truth of *Bottle* (the 36th frame), *David* (the 23rd frames), *Toys* (the 5th frame) and *Car* (the 36th frame) sequences, (b) Magnified details of the central ground truth, (c) Synthesized views of the four sequences, (d) Magnified details of the central synthesized views.

VI. CONCLUSION

In this paper, we presented a light field video dataset with four scenes. To our knowledge, this is the first light field video dataset captured with a plenoptic camera (Lytro Illum). Four popular compression methods were implemented in the proposed dataset to set a benchmark of light field video compression performance. In order to further improve the compression performance, a new ROI-based light field video compression method which considers motion characteristics was proposed. Instead of using the same way to encode the whole frame, every frame in our compression method is divided into an ROI region, a complex non-ROI region and a smooth non-ROI region to be processed differently. Then a novel fview synthesis method for light field video compression was presented to reduce bitrates further. Experimental results show that the proposed ROI-based compression method can save 5%–7% in bitrates compared with traditional HEVC-based light field video compression methods, and a reduction of about 50% in bitrates can be achieved by using the proposed view synthesis-based compression method. Furthermore, subjective results show that our view synthesis algorithm yields high-quality views that are superior to the state-of-the-art synthesis algorithm.

In the future, we would like to improve coding efficiency by utilizing depth estimation [26] from light field videos. Another interesting approach will be to design an interactive light field video compression method that will efficiently encode uneven views by predicting users' viewing trajectories and gestures.

REFERENCES

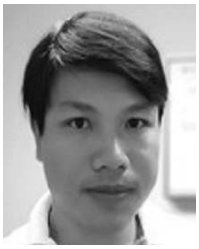
[1] I. Lytro. (2014). *Technical Specifications Lytro*. [Online]. Available: http://www.lytro.com/cameara/specs
[2] A. Raytrix. (2017). *3D Light Field Camera Technology*. [Online]. Available: http://raytrix.de/products
[3] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, Jul. 2005.
[4] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.*, vol. 32, no. 4, p. 46, 2013.
[5] A. Mousnier, E. Vural, and C. Guillemot. (Mar. 2015). "Partial light field tomographic reconstruction from a fixed-camera focal stack." [Online]. Available: https://arxiv.org/abs/1503.01903
[6] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–12, Jul. 2013.

[7] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int. Conf. Qual. Multimedia Exper.*, Lisbon, Portugal, 2016, pp. 1–2.
[8] Ł. D bała *et al.*, "Efficient multi-image correspondences for on-line light field video processing," *Comput. Graph. Forum*, vol. 35, no. 7, pp. 401–410, Oct. 2016.
[9] N. Sabater *et al.*, "Dataset and pipeline for multi-view light-field video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1743–1753.
[10] *The Proposed Light Field Video Dataset*. Accessed: Dec. 16, 2018. [Online]. Available: http://sites.google.com/view/lightfieldvideodataset
[11] I. Ihm, S. Park, and R. K. Lee, "Rendering of spherical light fields," in *Proc. 5th Pacific Conf. Comput. Graph. Appl.*, Seoul, South Korea, Oct. 1997, pp. 59–68.
[12] C. L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, Apr. 2006.
[13] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Efficient intra prediction scheme for light field image compression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 539–543.
[14] X. Huang, P. An, L. Shen, and R. Ma, "Efficient light field images compression method based on depth estimation and optimization," *IEEE Access*, vol. 6, pp. 48984–48993, 2018.
[15] U. Fecker and A. Kaup, "H.264/AVC-compatible coding of dynamic light fields using transposed picture ordering," in *Proc. 13th Signal Process. Conf.*, Antalya, Turkey, Sep. 2005, pp. 1–4.
[16] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenselet image compression scheme based on subaperture images streaming," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 4377–4737.
[17] G. Wang, W. Xiang, M. Pickering, and C. W. Chen, "Light field multi-view video coding with two-directional parallel inter-view prediction," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5104–5117, Nov. 2016.
[18] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H. 264/MPEG-4 AVC standard," *Proc. IEEE*, vol. 99, no. 4, pp. 626–642, Apr. 2011.
[19] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 1027–1034.
[20] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, p. 193, 2016.
[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[22] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 261–270.
[23] B. Wang, Q. Peng, X. Wu, E. Wang, and W. Xiang, "An interactive light field video system with user-dependent view selection and coding scheme," in *Proc. Pacific Rim Conf. Multimedia*, Hefei, China, 2018, pp. 727–736.
[24] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," ITU-T Video Coding Experts Group, Austin, TX, USA, Tech. Rep. VCEG-M33, Apr. 2001.
[25] 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.
[26] Y. Wu, "Research on depth estimation method of light field imaging based on big data in Internet of Things from camera array," *IEEE Access*, vol. 6, pp. 52308–52320, 2018.

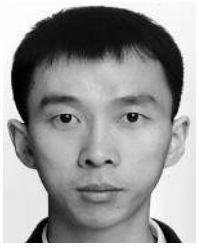
BING WANG received the B.Eng. degree in network engineering from Southwest Jiaotong University, Chengdu, China, in 2014, where she is currently pursuing the Ph.D. degree with the School of Information Science and Technology.

She is currently a Visiting Scholar with the College of Science and Engineering, James Cook University, Cairns, Australia. Her research interests include multimedia, video coding, and transmission.

QIANG PENG received the B.Eng. degree in automatic control from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.Eng. degree in computer application technology and the Ph.D. degree in traffic information engineering and control from Southwest Jiaotong University, Chengdu, China, in 1987 and 2004, respectively, where he is currently a Professor with the School of Information Science and Technology. His current research interests include digital video coding and transmission, image/graphics processing, and multimedia systems.



ERIC WANG received the B.Eng. degree in mechanical engineering and the M.Eng. degree in mechatronic engineering from the University of Science and Technology, Beijing, China, in 2006 and 2009, respectively, and the Ph.D. degree in telecommunications engineering from the University of South Queensland, Australia. He was a Postdoctoral Research Fellow with the School of Electrical and Mechanical, USQ, for two and a half years. He has authored more than 20 high quality papers, and participated in many national and international research projects. He received the runner-up in The Asia-Pacific Robot Contest (ABU Robocon) 2006, and the Best Paper Award from the IEEE Wireless Communications and Networking Conference, Cancun, Mexico, in 2011.



KANG HAN received the B.Eng. degree in electronic information engineering from Anhui Polytechnic University, Anhui, China, in 2014, and the M.Eng. degree in circuits and system from Shanghai University. He is currently pursuing the Ph.D. degree with the College of Science and Engineering, James Cook University, Cairns, Australia. His research interests include computer vision and machine learning.

WEI XIANG (S'00–M'04–SM'10) received the B.Eng. and M.Eng. degrees in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2000, respectively, and the Ph.D. degree in telecommunications engineering from the University of South Australia, Adelaide, Australia, in 2004. He is currently a Foundation Professor and the Head of discipline Internet of Things engineering with the College of Science and Engineering, James Cook University, Cairns, Australia. From 2004 to 2015, he was with the School of Mechanical and Electrical Engineering, University of Southern Queensland, Toowoomba, Australia. In 2008, he was a Visiting Scholar with Nanyang Technological University, Singapore. From 2010 to 2011, he was a Visiting Scholar with the University of Mississippi, Oxford, MS, USA. From 2012 to 2013, he was an Endeavour Visiting Associate Professor with The University of Hong Kong. He has authored more than 170 papers in peer-reviewed journal and conference papers. His research interests include the broad area of communications and information theory, particularly coding and signal processing for multimedia communications systems. He is an IET Fellow and a Fellow of Engineers Australia. He was named a Queensland International Fellow by the Queensland Government of Australia, from 2010 to 2011, an Endeavour Research Fellow by the Commonwealth Government of Australia, from 2012 to 2013, a Smart Futures Fellow by the Queensland Government of Australia, from 2012 to 2015, and a JSPS Invitational Fellow jointly by the Australian Academy of Science and the Japanese Society for Promotion of Science, from 2014 to 2015. He was a co-recipient of the Three Best Paper Awards at 2009 ICWMC, 2011 IEEE WCNC, and 2015 WCSP. He received several prestigious fellowship titles. He is an Editor of the IEEE COMMUNICATIONS LETTERS.