IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 May 2023, accepted 6 June 2023, date of publication 14 June 2023, date of current version 26 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3286298

RESEARCH ARTICLE

Lenslet Image Coding With SAIs Synthesis via 3D CNNs-Based Reinforcement Learning With a Rate Reward

XIAODA ZHONG[1], TAO LU[2], (Member, IEEE), DIYANG XIAO[3], AND RUI ZHONG[3], (Member, IEEE)

[1]Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou 510640, China
[2]Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430073, China
[3]School of Computer Science, Central China Normal University, Wuhan 430079, China

Corresponding author: Diyang Xiao (xdiyang@163.com)

This work was supported in part by the Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology), in part by the National Natural Science Foundation of China under Grant 62002130, and in part by the Fundamental Research Funds for the Central Universities under Grant CCNU22QN014.

ABSTRACT The deep learning-based coding schemes for lenslet images combine coding standards and view synthesis through Deep Learning (DL) models, where the compression efficiency is heavily influenced by the coding structure and quality of synthesized views. To exploit the inter-view redundancy among Sub-Aperture Images (SAIs), this paper proposes a hybrid closed-loop coding system that uses a novel coding structure based on checkerboard interleaving at a frame level. The frame-wise checkerboard interleaving method partitions an Original SAIs' Set (OSS) of images into two mutually exclusive subsets, each consisting of alternating rows and columns of SAIs. We utilize the video coding standard Versatile Video Coding (VVC) to encode one subset while proposing a novel rate constraint-reinforced 3D Convolutional Neural Networks (CNNs) to predict the other subset, referred to as the complement subset. The rate constraint-reinforced 3D CNNs is newly designed with a gradient loss and reinforced rate cost to improve synthesized SAIs' image quality and bit cost saving simultaneously. Experimental results on the light field image dataset demonstrate that the proposed hybrid coding system outperforms both HEVC_LDP and the previous state-of-the-art (SOTA), achieving an average BD-Bitrate savings of 41.58% and 23.31%, respectively.

INDEX TERMS Lenslet image, compression, reinforcement learning, 3D CNNs, VVC.

I. INTRODUCTION

This paper introduces a novel approach for enhancing the quality of Lenslet (LL) images obtained through unfocused plenoptic cameras, such as the Lytro camera, as originally presented by Ng et al. [1]. The fundamental mechanism of these cameras is the primary lens that focuses the reflected light rays of an object onto the microlens plane. Subsequently, each microlens captures the converging light rays and directs them to the image plane, where the incoming light intensity from a discrete set of directions is recorded, leading to the creation of a macro-pixel [2], [3]. To improve the quality of LL images under a constant bitrate condition, we propose a hybrid closed-loop compression system.

To address the challenge of managing the considerable volume of data generated by modern cameras, it is crucial to develop compression systems that can efficiently store and transmit LL images. Recently, innovative compression techniques have been introduced, which leverage DL models [4], [5], [6] to overcome the challenges of compressing lenslet images.

Ionut [4] proposed a novel DL-based prediction model, coupled with a context modeling method to encode prediction errors. The approach exhibits superior performance over the previous SOTA methods due to the learning ability of deep learning models. However, this approach is designed for lossless image compression and incurs a relatively high bit cost compared to lossy compression techniques. This paper focuses on the development of a closed-loop lossy compression system that integrates DL-based prediction and video coding standards.

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

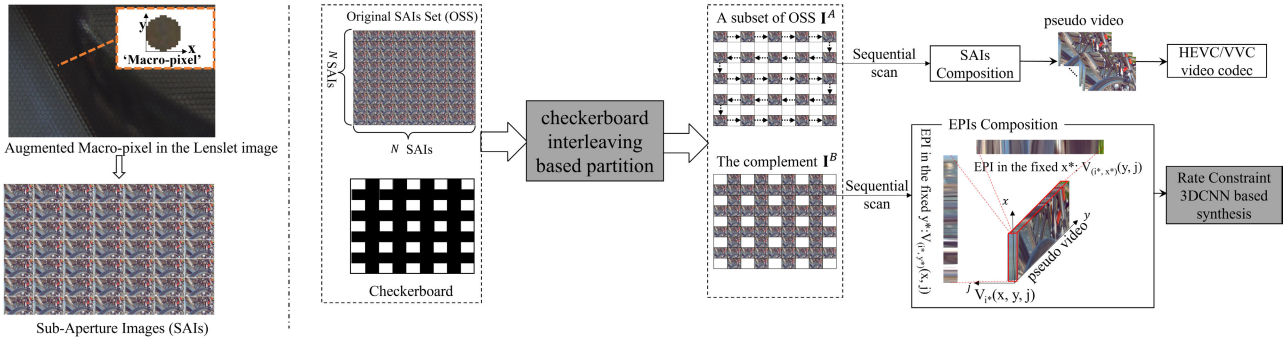

FIGURE 1. Flowchart of the hybrid coding system. OSS is the Original SAIs' Set. (a) The conversion from Lenslet (LL) image to multiple Sub-Aperture Images (SAIs) is carried out by collecting the pixels having the same coordinates (x, y) in a macro-pixel. (b) The contributions involve the checkboard interleaving based partition and the rate constraint 3D CNNs based synthesis. (c) The Epipolar Plane Image (EPI) is a 2-dimensional projection of a 4-dimensional light field. In this flowchart, EPI is obtained by selecting and fixing one angular dimension and one spatial dimension, where the optional angular dimensions are represented by i and j, and the spatial dimensions are represented by x and y.

In this paper, we discuss a DL-based hybrid compression system that utilizes the concept of pseudo video-sequence generation and inter-view redundancy to enhance compression efficiency. The approach involves converting Lenslet (LL) images to Sub-Aperture Images (SAIs) through the collection of pixels with the same coordinates in a macro-pixel (see Figure 1), as proposed in [7] and [8]. The hybrid compression system involves two critical elements: the efficiency of the hybrid coding structure, including scanning order and interleaving method, and the quality of the predicted images through deep-learning-based view synthesis.

To preserve the effectiveness of DL-based view synthesis in image prediction, Jia [5] proposed a view synthesis approach utilizing Generative Adversarial Networks (GANs). This method demonstrated exceptional performance in LL image synthesis. However, the unstable training of GAN can lead to unreliable image quality and affect the coding efficiency. To address the limitations of GAN-based methods, Bakir [9] proposed a Dual Discriminator GAN (D2GAN) that showed an improvement in the quality of the synthesized views. However, the adversarial components of the GAN still suffer from unstable training, which results in significant color differences among synthesized images.

To overcome the limitations, Hou introduced a compression framework that utilizes the angular super-resolution technique based on CNNs to generate synthesized SAIs [6]. The approach entails encoding a select subset of the Original SAIs' Set (OSS) using HEVC compression. The resultant sparse reconstructed SAIs are then fed into a CNNs model, which predicts the SAIs for the complementary subset of OSS. However, the degradation in the quality of synthesized views is directly proportional to their distance from the HEVC-encoded sparsely captured views, as our observations. In particular, the CNN-driven view synthesis technique results in lower image quality for views that are located farther from the HEVC-encoded sparsely captured views.

In general, the compression framework in [6] lacks efficient coding structure. Additionally, both the GAN and CNNs based view synthesis techniques have drawbacks in synthesizing high-quality images for LL images. Thus, our objective is to enhance the compression performance for LL images by presenting an appropriate coding structure and a highly efficient DL-based view synthesis model.

In this paper, we propose a novel closed-loop hybrid compression scheme for lenslet images inspired by the coding structure presented in [5]. Our scheme introduces VVC [9] to compress a subset of the original SAIs' set and utilizes rate constraint-reinforced 3D CNNs to predict the complement of the subset. Nonetheless, the division of the subset for conventional intra/inter or 3D CNNs-based prediction is a crucial component that has a significant impact on the coding efficiency of the hybrid compression framework. In summary, the novel contributions of this study are as follows:

1. To fully exploit the inter-view redundancy among SAIs, we propose a frame-wise checkerboard interleaving method and then design a hybrid coding structure. The frame-wise checkerboard interleaving method partitions an Original SAIs' Set (OSS) of images into two mutually exclusive subsets by alternating rows and columns of SAIs. More specifically, we assign the VCC's intra/inter prediction to the SAIs at the position with odd coordinates and the rate-constraint 3D CNNs prediction to the SAIs with even coordinates. However, the inputted LL images of the 3D CNNs is the reconstructed images of the VVC encoder rather than the original images. According to the coordinates of SAIs, we split the OSS into a subset I^A and its complement I^B (see Figure 1). The SAIs of the I^A are firstly composited into a pseudo video and then encoded by the VVC codec. The SAIs of the I^B are synthesized by the novel rate-constrained 3D CNNs.

2. The set $\hat{I}^A$ consisting of reconstructed SAIs is formed by decomposing the frames in the reconstructed pseudo video. The SAIs within $\hat{I}^A$ are then sequentially scanned to produce Epipolar Plane Images (EPIs), as illustrated in Figure 1. Subsequently, a rate constraint-reinforced 3D CNNs is applied to the EPIs to synthesize views. More specifically, the original 3D CNNs [10] cannot simultaneously achieve low bit cost and high image quality for view synthesis-based prediction. To address this issue, we introduce a new model

of rate constraint 3D CNNs with a gradient loss function, which employs the residuals' rate of all SAIs in I^B as a reward to supervise the reinforcement learning process of the 3D CNNs. This approach allows us to simultaneously optimize both the bit cost and image quality of the synthesized views.

II. RELATED WORK

In this section, we briefly survey the coding standards-based and deep-learning-based Lenslet Image Compression.

A. CODING STANDARDS-BASED LENSLET IMAGE COMPRESSION

JPEG-Pleno [11] was proposed as a standard for capturing, representing, converting formats, and compressing images for light fields. Although it offered a complete framework for plenoptic data, its encoder was not efficient in compressing light field images. The JPEG-Pleno [11] has been evaluated and found to be inefficient for LL image compression. To overcome this constraint, High Efficiency Video Coding (HEVC) [12] has been suggested as a compelling alternative due to its considerably enhanced compression performance in comparison to its forerunners. However, it's essential to note that HEVC was primarily designed to account for local spatial and temporal continuities in video data. Given that lenslet images exhibit systematic spatial discontinuities between microlens images, utilizing the HEVC standard to encode such data can result in inefficiencies in compression.

The field of lenslet image coding has witnessed the emergence of several techniques to tackle the challenges associated with compressing plenoptic data. These techniques include intra-prediction methodologies and wavelet compression methods designed to leverage the inherent intra-frame redundancies of the data. In a study conducted by [13], a 4D Discrete Wavelet Transformation (DWT) technique was proposed that was combined with the Set Partitioning into Hierarchical Trees (SPIHT) algorithm to encode the resultant wavelet subbands. This DWT compression system allows for progressive decoding of LL data. Furthermore, to reduce the redundancy within subbands, wavelet compression has been applied to SAIs. In another study [14], the low-frequency bands that were decomposed from reconstructed SAIs via a 2D DWT were coded by a 3D Discrete Cosine Transform (DCT) followed by Huffman coding, while the high-frequency bands were directly processed by arithmetic coding. These wavelet-based coding techniques provide quality scalability and a comprehensive framework to explore the intra-frame redundancies of LL images in the frequency domain.

In the domain of lenslet image coding, an alternative approach to minimize spatial redundancies is through intra prediction directly applied to microlenses. In the ICME 2016 Grand Challenge on LL Image Compression [15], Self-Similarity (SS) compensated intra-prediction [16] was proposed to exploit spatial redundancies for specific microlens arrangements in LL images. Bi-directional SS compensation based intra-prediction [17] was subsequently introduced to further minimize prediction errors for microlenses with slight view disparities. These SS-based intra-prediction methods were found to achieve high coding efficiency and low prediction error for the specific rectangular pattern of microlenses. Furthermore, a locally linear embedding method was proposed in [18] and integrated into specially designed HEVC directional intra prediction modes for rectangular microlenses. Recently, local redundancies were exploited using a Gaussian regression-based prediction, incorporated into directional intra prediction as a prediction mode [19]. To further explore the repetitive patterns of LL images, [20] proposed uni-directional and bi-directional SS search-based schemes for reference selection, which aim to minimize the prediction residuals under a Rate-Distortion Optimization (RDO) criterion.

In the realm of exploiting redundancies among neighboring viewpoints, a range of coding methods have been proposed over the years. One such approach is the inter-prediction coding method, which was presented in [21], where one of the views is used as a reference to predictively code the remaining views. Another popular approach involves the utilization of the Multiview Video Coding (MVC) extension [22] of the HEVC standard. To optimize the prediction residual for MVC, [23] employs a 2D warping-based disparity compensation. In [24], a joint motion and disparity estimation method is proposed to take advantage of inter-frame and inter-view predictions. Moreover, a hierarchical reference structure is designed for HEVC-based inter-coding of the pseudo video sequence in [8]. During ICME2016, Liu proposed a method to generate synthesized SAIs by collecting pixels corresponding to the same coordinates in a macro-pixel [7]. Alternatively, SAIs can be composited into a pseudo video, and HEVC_LDP can be utilized to exploit temporal redundancies.

In our previous research [25], we presented a method for macro-pixel prediction that involved a linear combination of the neighboring reconstructed macro-pixels. This approach used an L1-optimized prediction algorithm that exploited the spatial correlation of pixels with the same spatial coordinates within neighboring macro-pixels to reduce spatial redundancies. Building upon this method, our subsequent research [26] aimed to further improve coding efficiency by reducing spatial redundancies even further. To achieve this goal, we proposed a dictionary learning-based prediction method directly on macro-pixels, which has demonstrated promising performance on lenslet images [27].

B. DEEP-LEARNING-BASED LENSLET IMAGE COMPRESSION

When traditional compression methods fail to achieve breakthrough progress, researchers have turned to DL to improve the performance of light field image compression. In 2018, the first category is proposed based on macro-pixels as the basic unit, where macro-pixels are reconstructed using a CNNs prediction method, and CALIC encoding is applied to the residual of macro-pixels to create a lossless coding system [28]. However, this method consumes a relatively

high bit rate. Afterwards, Zhong proposed deep CNNs models based frame-intra prediction modes, which are embedded into 35 frame-intra prediction modes of HEVC. The optimal mode is selected using a rate-distortion optimization algorithm. This approach can save approximately 11% of the BD-Rate compared to the HEVC coding standard [29].

Besides the above mentioned macro-pixels level-based methods, the hybrid compression systems are alternatively designed on the SAIs obtained by collecting pixels with the same coordinates in a macro-pixel to achieve the conversion from a light field image. The hybrid deep-learning-based compression systems are categorized into two classes: the spatial synthesis based compression methods and angular view synthesis based compression methods.

In spatial synthesis-based compression method proposed by Ma et al. [30], the key idea is to encode low-resolution SAIs with coding standards and enable spatial super-resolution to synthesize the high-resolution SAIs. In contrast, the angular view synthesis based hybrid compression systems [6], [31] apply coding standards such as HEVC and VVC to a subset of the SAIs and then apply DL-based view synthesis to the complement of the SAIs. We fulfilled and compared the spatial synthesis and angular view synthesis-based compression methods, and noticed that the angular view synthesis-based method has more potential to perform better than the spatial synthesis-based method.

In angular view synthesis-based compression method, there are two critical elements: 1) the efficiency of the hybrid coding structure, including scanning order and interleaving method (how to partition the SAIs into the subset compressed with the coding standards or the complement subset synthesized by DL models), 2) the quality of the predicted images through deep-learning-based view synthesis. In 2016, Kalantari et al. proposed a light field image angular super-resolution algorithm based on two consecutive CNNs model of disparity and color estimation [32]. It synthesizes the complement subset of SAIs from only four corner sparse SAIs. The drawback of the interleaving strategy in Kalantari's method [32] is that the image quality of the synthesized views drop with the increasing distance between the synthesized view and the corner views. Experiments show that the maximum difference of view synthesis can reach more than 5dB, which demonstrates that the intervleaving strategy has a significant impact on the visual effect. In end-to-end 3D CNNs based view synthesis for LL images [10], the sparse SAIs are sampled with an interval. Thus, the complement subset of SAIs are synthesized from neighboring SAIs. Since the coding structure of 3D CNNs based view synthesis in [10] is more efficient for coding structure. We ultilize the interleaving method of 3D CNNs based view synthesis in [10] to partition the subset of view synthesis for hybrid coding structure.

In 2018, Zhao proposed a method that reconstructs light field images by encoding the base viewpoint obtained through sparse sampling and using CNNs to reconstruct the enhanced viewpoints based on the base viewpoint [33]. More specific, the CNNs is designed to characterize the nonlinear relationship among sub-views caused by the light intensity, angle displacement-induced parallax, and distortion. This method performs well, but it has not provided a complete coding scheme.

Afterwards, Jia [5] proposed a method that combines a view synthesis based on generative adversarial networks (GAN) with a hierarchical prediction structure in the light field image compression system. Jia's method explores the powerful performance of GAN in view synthesis and has been proven to perform well in low-frequency image synthesis. However, the instability of GAN training results in unreliable image quality, which ultimately affects the coding efficiency. Bakir proposes a double discriminator GAN (D2GAN) to improve the quality of synthesized views [9]. However, D2GAN suffers from the instability of GAN's adversarial components, which results in significant color differences among synthesized images and then damage the coding performance.

To overcome the limitations, Hou proposed a coding framework involves using HEVC to code part of the OSS, then using a CNNs model to predict the SAIs for the remaining part based on sparse reconstructed data [6]. However, the quality of the synthesized views decreases with increasing distance from the sparse views. Specifically, views further from the HEVC-coded sparse views correspond to lower image quality via the CNNs-based view synthesis.

In this paper, we investigate the angular view synthesis-based compression method, and then explore the efficiency of the hybrid coding structure and the image quality of the view synthesis. To enhance the performance of the coding framework, we proposed the checkerboard interleaving based partition for coding structure and the rate constraint 3D CNNs based synthesis to guarantee the synthesized views.

III. PROPOSED HYBRID LENSLET COMPRESSION SYSTEM

Our study introduces a new closed-loop lenslet image compression framework that uses checkboard interleaving based partition and 3D CNNs-based reinforcement learning with rate constraint for view synthesis. Prior research focused on combining coding standards and DL models to achieve higher performance, but our approach compares spatial [30] and angular view synthesis and chooses the latter as the better option. This addresses the challenge of partition methods and optimizing view synthesis and achieves efficient compression and high-quality rendering for LL image compression.

A. CHECKBOARD INTERLEAVING BASED PARTITION

The proposed lenslet image coding system is illustrated in Figure 2, which follows a closed-loop hybrid coding paradigm, takes the original SAI images of the OSS $I = I^A \cup I^B$ as input, and then performs VVC coding for all the SAIs within the subset $I^A = \{I(i,j) \in \mathbb{R}^{w \times h}\}$ (where w, h denote the width and height for each frame, and $i, j \in \{1, 3, \cdots, N\}$ denote the index of the SAI horizontally and vertically, and $N = 9$) as well as view synthesis for the complement $I^B = \{I(i,j)\}$, $I^A \cap I^B = \varnothing$, and $i, j \in \{2, 4, \cdots, N-1\}$.

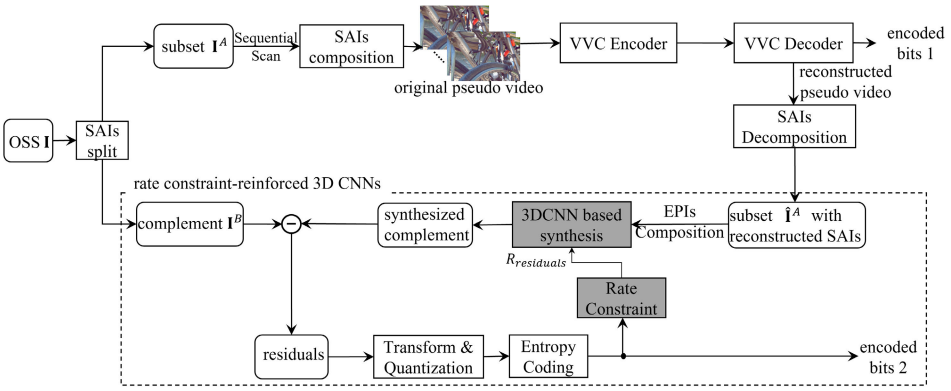

FIGURE 2. Framework of the hybrid coding system with the rate constraint-reinforced 3D CNNs-based view synthesis: OSS is the Original SAIs' Set.

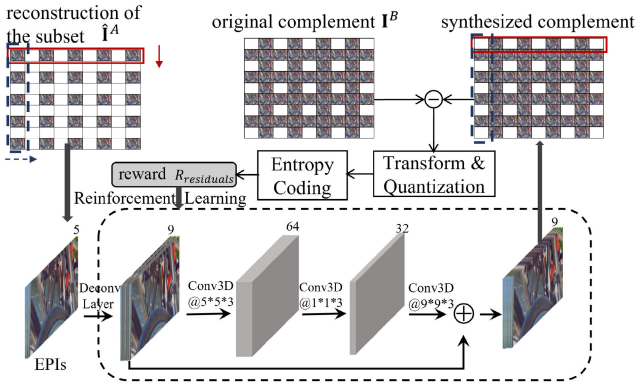

FIGURE 3. A novel rate constraint-reinforced 3D CNNs for view synthesis.

As illustrated in Figure 3, the SAIs within a red solid rectangle are sequentially scanned to form horizontal EPIs. EPIs are slices extracted from the pseudo video $V \in \mathbb{R}^{w\times h\times N^2}$ at a constant angular dimension. More specific, an Epipolar Plane Image (EPI) is a 2D slice projected from a 4D Light field by selecting and fixing two dimensions, which involves three kinds of choices: two angular dimensions, two spatial dimensions, or one angular dimension and one spatial dimension.

In Figure 1, i and j represent the two optional angular dimensions, while x and y are the two spatial dimensions. $V_{(i^*)}(x, y, j)$ is a 3D pseudo video obtained by fixing one angular dimension i^* in 4D Light field. $V_{(i^*,x^*)}(y, j)$ and $V_{(i^*,y^*)}(x, j)$ are two kinds of EPIs by fixing x^* and y^*. Moreover, $I(i, j)$ is also one kind of EPI by fixing two angular dimensions (i and j), formulated in (1):

$$I(i, j) = L_{(i^*=i,j^*=j)}(x, y), \tag{1}$$

where $L \in \mathbb{R}^{w\times h\times N\times N}$ denotes the 4D LL images.

B. 3D CNNs-BASED REINFORCEMENT LEARNING WITH RATE CONSTRAINT FOR VIEW SYNTHESIS

The input of the view synthesis is the EPIs extracted from the reconstructed SAIs of the subset $\hat{I}^A = \{\hat{I}(i, j)\}$. Afterwards, we generate the synthesized SAIs $\hat{I}_s(i, j)$ for the complement I^B in two phases, including the upsampling interpolation via Long's fractionally stridden convolution [34], as well as the novel 3D CNNs-based residual prediction with rate constraint. The residuals of SAIs within the complement I^B are transformed, quantized, and entropy coded to generate the encoded bits. The rate of the entropy-coded residuals is added on the 3D CNNs [10] to reinforce the learning of the model. Moreover, the gradient difference is also considered. The total loss is formulated in (2):

$$E_{loss} = \alpha_D E_D + \alpha_G E_G + \alpha_R E_R, \tag{2}$$

where α_D, α_G, and α_R are the parameters that denote the weight of each regularization term. In the following, we will mathematically describe each term.

Residual. The novel 3D CNNs-based prediction residual E_D is given by (3):

$$E_D = \sum_{i,j\in\{2,4,\cdots,N-1\}} \{\left\|\hat{I}(i, j) - \hat{I}_s(i, j)\right\|^2\}, \tag{3}$$

where $\hat{I}(i, j)$ represents the reconstruction of the SAI at the index denoted by (i, j) (the i-th row and the j-th column), and $\hat{I}_s(i, j)$ is the predicted image yielding via the rate-reinforced 3D CNNs.

Rate reward. Moreover, the rate reward E_R is formulated in (4):

$$E_R = \sum_{i,j\in\{2,4,\cdots,N-1\}} \{\lambda R_{residual}(i, j)\}, \tag{4}$$

where $R_{residual}(i, j) = Rate(\hat{I}(i, j) - \hat{I}_s(i, j))$, $Rate()$ is the rate of the entropy coded residual after the processes of transformation and quantization, and $\lambda = 0.85 \cdot 2^{(QP-12)/3}$ is set on the selected QPs originally presented in [35].

More specifically, the frame-wise residual $I_r(i, j) = \hat{I}(i, j) - \hat{I}_s(i, j)$ is partitioned into blocks. The nonzero coefficients within the block-wise matrix are transformed via DCT [12], quantized, and entropy coded by CABAC [12] with quantization parameters QPs. Two components determining the bits cost of coding the residual are the position of

the nonzero coefficients, denoted by $I_p(i,j)$, and the nonzero coefficients $I_r(i,j)$ calculated by (5):

$$R_{residual}(i,j) = Rate(I_p(i,j)) + Rate(I_r(i,j)), \tag{5}$$

where $Rate(I_p(i,j))$ and $Rate(I_r(i,j))$ are the bits of encoding the position and the coefficients, which are formulated as (6) and (7):

$$Rate(I_p(i,j)) = a_1 \cdot \sum_{k\in\{1,2,\cdots,K\}} N(k), \tag{6}$$

$$Rate(I_r(i,j)) = \sum_{k\in\{1,2,\cdots,K\}} (a_2 \cdot QP + b_2), \tag{7}$$

where $k \in \{1,2,\cdots,K\}$ represents the index of the block in a frame, K is the total number of blocks. a_1 is a parameter computed as $a_1 = c \cdot log_2^d$ [35], depending on the block-wise sparsity level c and the length of the coefficient vectors d, and $N(k)$ is the number of vectors from nonzero coefficients which are transmitted to the decoder. $(a_2, b_2) = (-0.023, 1.576)$ is the pair of parameters used to encode the coefficients calculated via the least-squares regression line [36].

Gradient loss. We also include the gradient loss to preserve the structure consistency of the image. The gradient loss E_G is formulated in (8):

$$E_G = \sum_{i,j\in\{2,4,\cdots,N-1\}} \{\left\|\nabla\hat{I}(i,j) - \nabla\hat{I}_s(i,j)\right\|^2\}, \tag{8}$$

In our system, the Sobel operator is used to compute the horizontal and vertical gradient components for the gradient operator denoted by ∇.

We firstly minimize the term $(E_D + E_G)$ to optimize the networks Θ^t of the 3D CNNs via the standard back propagation [10]. Furthermore, we adopt the DDPG [37] to train the rate reward-reinforced 3D CNNs by maximizing the reward R at the $t-th$ iteration, formulated in (9):

$$R = \frac{1}{1+E_R}, \tag{9}$$

where $\Theta^t \leftarrow \arg\max\{R_{t-1}(\Theta^{t-1}), R_t(\Theta^t)\}$ represents the process of updating the network Θ^t at the t-th iteration.

The DDPG is a united algorithm of the actor-critic deterministic policy gradient algorithm [37], which contains two kinds of models: the actor and the critic. The actor takes action according to the environment state, and the critic evaluates the actor's action and provides action-value. In this work, the actor is the rate reward-reinforced 3D CNNs. The proposed method serves as the actor to learn a policy μ_{θ^μ} by maximizing the expected actor-value, written as (10), and then a linear regression network is developed to play as a critic.

$$\max_{\theta^\mu} J(\mu) = E_{s\sim p^\beta}\left[Q^{\mu_{\theta^\mu}}\left(s, \mu_{\theta^\mu}(s)\right)\right], \tag{10}$$

where θ^μ is the parameter of the policy, p^β is the state visitation distribution, s is the hidden state in the network, and $Q^{\mu_{\theta^\mu}}(\cdot)$ is the actor-value function, modelled by the critic.

IV. EXPERIMENTAL SETUP

The same evaluation procedure as in [38] is followed in the experimental evaluation of the proposed coding system, i.e., we use the EPFL test set [39] consisting of 12 LL images to evaluate the coding performance. Each raw image has a resolution of $15 \times 15 \times 434 \times 625$ pixels. Due to the dark artifacts around the macro-pixels, we extract $9 \times 9 \times 434 \times 625$ pixels to form SAIs ($N = 9$). The raw images are first demosaiced, devignetted, clipped from 10-bit to 8-bit representation, color calibrated, and converted to the YCBCR4:2:0 color-map representation.

Besides the 12 images for the test, the rest 106 images from the dataset presented in [39] were used to train the upsampling learnable kernel and 3D CNNs. To demonstrate the advantages of the proposed compression method, we compare the PSNR and the codelength of the encoded LL images against the following HEVC-involved coding systems:

1) HEVC operating in low delay P [12], denoted here HEVC_LDP;
2) the pseudo-sequence-based compression of [7], denoted here Liu2016;
3) the work with disparity-guided sparse coding described in [40], denoted here as Chen2018;
4) the CNNs-based angular super-resolution approach within the light field compression system in [6], denoted here as Hou2019.

The experiments are performed using a set of QPs, which includes 4 QPs, namely (18, 24, 30, 36). We empirically set $\alpha_D = 0.5$, $\alpha_G = 0.25$, and $\alpha_R = 0.25$. The PSNR of the decoded LL image denoted as *PSNR*, is computed on the raw 8-bit image as $PSNR = 10 \times \log_2 \frac{255^2}{MSE}$, where

$$MSE = \frac{1}{N^2} \cdot \sum_{i,j=1}^{N} (I(i,j) - \hat{I}(i,j))^2. \tag{11}$$

The transmitted bits consist of the entropy coded residuals of the prediction for both VVC and views synthesis, as well as prediction modes information within the VVC encoder. The 3D CNNs and upsampling kernels are learned and optimized during training. During the test, we directly adopt the networks stored in the designed hybrid codec. Thus, the 3D CNNs and upsampling kernels are not transmitted and counted in encoded bits.

V. EXPERIMENTAL RESULTS

The comparisons are given in Table 1 and illustrated in Figure 4. Table 1 reports the BD-PSNR (Bjontegaard Delta Peak Signal-to-Noise Ratio) and BD-Bitrate (Bjontegaard Delta Bitrate) computed using Bjontegaard's evaluation tools [42] for the 12 LL images from the EPFL dataset. More specific, BD-PSNR and BD-Bitrate are commonly used metrics to evaluate the performance of image and video compression methods, where BD-PSNR measures the quality of the compressed image by computing the peak signal-to-noise ratio between the compressed and original

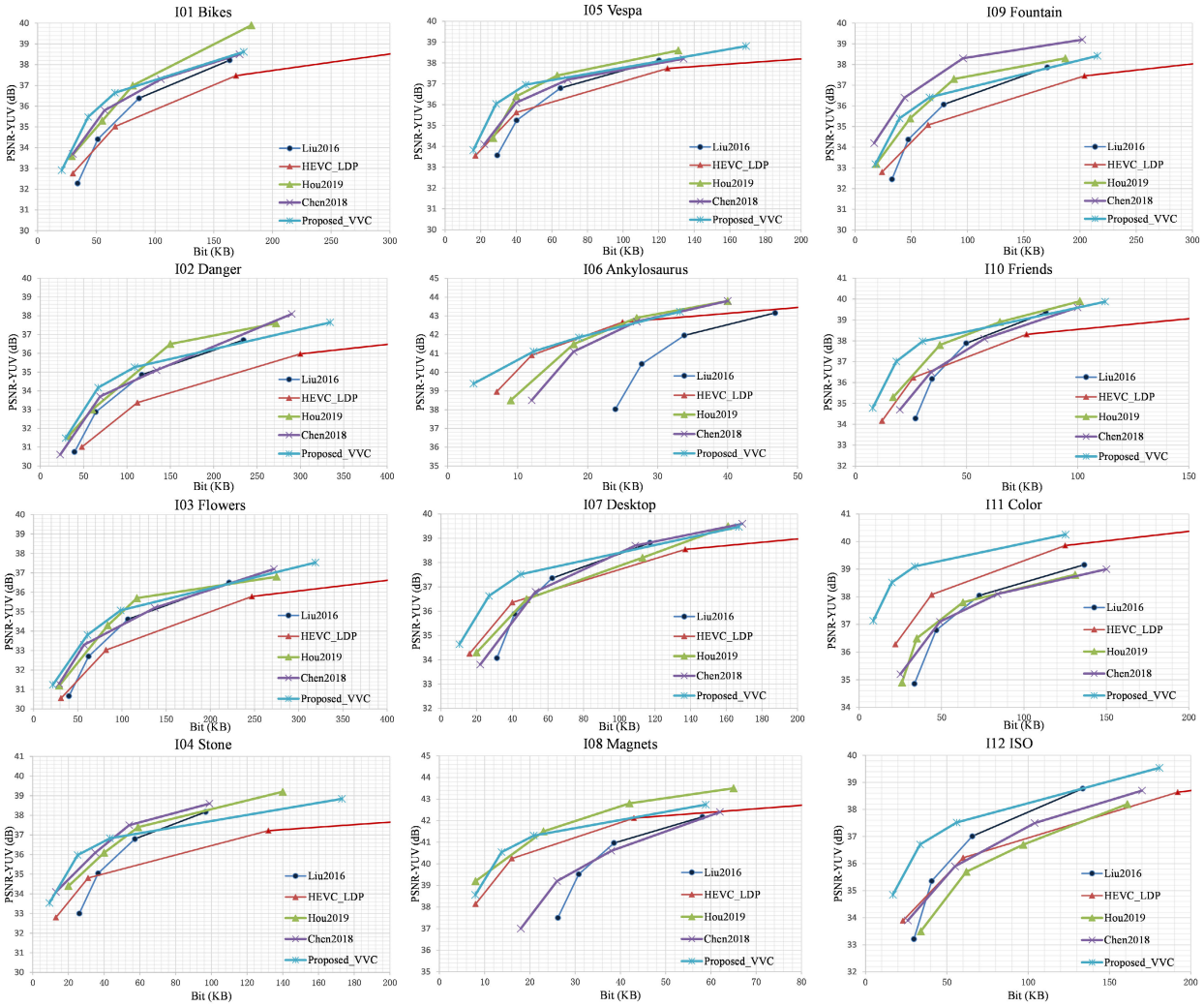

FIGURE 4. Rate distortion curves of the proposed method and the references.

TABLE 1. Different methods compared to the Liu2016 [7]: BD-PSNR (dB) and BD-Bitrate (%) on YUV channels.

	HEVC_LDP [12]		Chen2018 [40]		Hou2019 [6]		Zhang2022 [41]		Proposed_HEVC_LDP		Proposed_VVC	
	(dB)	(%)	(dB)	(%)	(dB)	(%)	(dB)	(%)	(dB)	(%)	(dB)	(%)
I01	-0.35	13.61	0.74	-18.5	0.99	-21.57	0.64	-14.80	0.92	-26.27	1.17	-34.22
I02	-1.24	52.04	0.39	-13.8	0.72	5.97	0.18	-10.71	0.59	-16.26	0.82	-27.16
I03	-0.63	23.08	0.39	-13.6	0.63	-7.91	0.17	-5.20	0.72	-20.11	0.9	-30.07
I04	-0.45	16.13	1.04	-33.6	0.74	-20.72	0.27	-15.26	0.79	-31.75	0.9	-44.2
I05	-0.05	-7.89	0.50	-17.6	0.76	-27.23	0.34	-9.64	0.78	-33.4	0.93	-40.49
I06	1.68	-56.65	2.01	-33.7	2.16	-68.87	0.76	-37.60	1.77	-43.58	2.38	-60.16
I07	0.11	-11.94	0.11	-4.6	-0.09	-3.98	0.97	-14.15	0.66	-34	0.98	-45.41
I08	1.34	-53.42	0.06	-11.5	2.01	-69.98	0.53	-38.74	0.95	-48.14	1.47	-61.25
I09	-0.32	12.44	1.95	-66	0.95	-24.50	0.38	-12.84	0.8	-24.88	0.9	-35.11
I10	0.01	-11.68	0.08	-6.3	0.82	-35.9	0.47	-9.57	0.87	-37.33	1.26	-52.93
I11	1.23	-42.53	0.05	-3	0.19	-14.13	0.78	-9.24	1.56	-55.07	2.16	-80.07
I12	-0.32	15.19	-0.38	24.9	-1.12	41.29	0.12	-5.26	0.26	-13.7	1.06	-39.46
Average	0.09	-4.30	0.58	-16.44	0.73	-20.63	0.47	-15.25	0.89	-32.04	1.24	-45.88

images, as well as BD-Bitrate measures the efficiency of the compression method by quantifying the reduction in bit-rate achieved relative to a reference method. Moreover, Figure 4 illustrates the rate distortion curves for the reference methods and the proposed approach.

In this study, we conducted an evaluation of both existing reference methods and our proposed approach through a comparative analysis with Liu2016 method [7]. The proposed method, named Proposed_VVC in Table 1 and Figure 4, was compared with several reference methods,

including Liu2016 [7], JPEG-Pleno [11], HEVC_LDP [12], Chen2018 [40], and Hou2019 [6], in terms of BD-PSNR gain and BD-Bitrate savings. The JPEG-Pleno [11] corresponds to nearly -1.89 dB BD-PSNR loss against Liu2016 [7]. However, HEVC_LDP [12], Chen2018 [40], Hou2019 [6], and the proposed method achieve an average BD-PSNR gain of 0.09 dB, 0.58 dB, 0.73 dB, and 1.24 dB against Liu2016 [7] respectively. These BD-PSNR gains correspond to BD-Bitrate savings of 4.3%, 16.44%, 20.63%, and 45.88%. Overall, the proposed method outperforms the reference methods, achieving 1.15 dB, 0.66 dB, and 0.51 dB BD-PSNR gain and 41.58%, 29.44%, and 25.25% BD-Bitrate saving over HEVC_LDP [12], Chen2018 [40], and Hou2019 [6], respectively.

Moreover, we conduct a comparative analysis between our proposed method and Zhang's approach [41]. Zhang's method [41] demonstrates an average BD-PSNR improvement of 0.47 dB and BD-Bitrate savings of 15.25% when compared to Liu2016 [7]. Although Zhang's technique does not surpass the performance of Chen2018 [40], Hou2019 [6], and our proposed method, it exhibits robustness across all 12 EPFL test images. Additionally, we investigate the findings presented in Shi2023 [43]. The results from Shi2023 [43] indicate that their approach slightly outperforms JPEG-Pleno [11] but falls short when compared to Liu2016 [7].

The present study demonstrates that the proposed method exhibits superior robustness compared to the reference methods. Notably, the proposed method outperforms Liu2016 [7] in all tested LL images. In contrast, the reference methods [6], [12], [40] exhibit limited effectiveness in improving all images. For instance, HEVC_LDP [12] shows a BD-PSNR loss in seven out of twelve images, including 'I01', 'I02', 'I03', 'I04', 'I05', 'I09', and 'I12'. Furthermore, Chen2018 [40] reports that the image 'I12' demonstrates a BD-PSNR loss of -0.38 dB. Similarly, Hou2019 [6] reports that the images 'I07' and 'I12' demonstrate a BD-PSNR loss of -0.09 dB and -1.12 dB, respectively. The proposed method shows substantial improvement in all test images, thereby demonstrating its robustness.

We also compare our method with Bakir's work, LL image compression with the dual discriminator GAN-based view synthesis and Versatile Video Coding (VVC), denoted as Bakir2020 [9]. The average BD-PSNR gain and BD-Bitrate saving of Bakir2020 [9] against Liu2016 [7] is 0.68 dB and 22.57%. Given the excellent performance of the VVC, our method yields an average BD-PSNR gain of 0.56 dB and rate saving of 23.31% compared to Bakir2020 [9].

To confirm the effectiveness of the proposed hybrid coding structure, we list the performance of the proposed method with the HEVC_LDP [12] codec in Table 1. The proposed method, termed Proposed_HEVC_LDP, replaces the VVC codec with HEVC operating in low delay P, thereby foregoing the advantages offered by the high coding efficiency of VVC. Despite being unable to perform as well as the Proposed_VVC, the Proposed_HEVC_LDP outperforms the reference methods, yielding 0.8 dB, 0.31 dB, and 0.16 dB BD-PSNR gain and 27.74%, 15.6%, and 11.41% BD-Bitrate saving over HEVC_LDP [12], Chen2018 [40], and Hou2019 [6], respectively. The experimental results of the Proposed_HEVC_LDP demonstrate that the hybrid coding structure perform superior than the SOTA.

TABLE 2. Ablation results of the proposed 3D CNNs-based reinforcement learning with rate constraint.

Mode	E_D [10]	E_G	E_R	12 LL images (Average PSNR dB)
1	✓			40.3
2	✓	✓		40.53
3	✓		✓	40.57
4	✓	✓	✓	41

VI. ABLATION STUDY

We conduct the ablation study for the LL prediction based on the 3D CNNs-based reinforcement learning with rate constraint. In Table 2, we utilize mode 1 to mode 4 to denote the baseline method 3D CNNs [10] (mode 1), the proposed models (mode 2, 3, 4). The proposed method comprises the pixel-wise loss E_D, the gradient loss E_G, and the rate constraint E_R.

The ablation results of the proposed method on 12 LL images justify that both the gradient loss E_G can slightly enhance performance of synthesized images. Mainly, for the 12 LL images from the new EPFL test set [39], the combination (mode 2) of the pixel-wise loss E_D and the gradient loss E_G perform at 0.23 dB PSNR gain compared to the baseline method of the proposed method (mode 1). Moreover, the combination of the E_D and the E_R (mode 3) raises the 0.27 dB PSNR against mode 1.

Furthermore, the proposed method outperforms the 3D CNNs [10] (mode 1) significantly. The combination (mode 4) arrives at 0.7 dB PSNR gain against mode 1. The superior performance of the gradient loss (mode 2), the the rate contraint E_R (mode 3), and the combination (mode 4) demonstrates that the proposed method effectively enhance the image quality of the view synthesis while maintaining the bitrate cost. The improvements verify that the 3D CNNs-based reinforcement learning with rate constraint is more robust than the 3D CNNs [10] (mode 1) by enhancing the gradient and rate constraint. Meanwhile, we see that the image enhancement and the rate constraint are compatible in enhancing the performance.

VII. TIME COMPLEXITY COMPARISON

The proposed framework was implemented on a machine equipped with an Intel® Core^{TM} i9-10900K CPU, NVIDIA GeForce RTX 3090, and 32GB of RAM, running a 64-bit Ubuntu 18.04.05 LTS Operating System. We evaluate the complexity of the proposed method. The proposed method comprises two components, the standard VVC to encode one subset followed by the novel rate constraint-reinforced 3D CNNs to predict the complement subset.

In Versatile Video Coding (VVC), the time complexity of video compression is determined by various processing steps, including intra prediction, inter prediction, transform

coding, quantization, and entropy coding. Due to its increased complexity and improved compression efficiency, the time complexity of VVC is expected to be higher than that of previous video coding standards. Specifically, the time complexity of intra and inter prediction in VVC can be expressed as $O(M_v \cdot k_v \cdot \alpha_{cu})$ and $O(M_v \cdot f_v \cdot \beta_{cu})$, respectively. Here, $M_v = 6$ represents the number of Coding Unit (CU) sizes (4×4, 8×8, 16×16, 32×32, 64×64, 128×128), $k_v = 65$ denotes the number of intra directions, and $f_v = 5$ denotes the average number of motion vector candidates. The specific factors α_{cu} and β_{cu} are determined by various factors such as method implementation, code optimizations, execution platform, and memory allocation of CU.

In comparison, HEVC has a time complexity of $O(M_h \cdot k_h \cdot \alpha_{cu})$ and $O(M_h \cdot f_h \cdot \beta_{cu})$, where $M_h = 4$ and $k_h = 35$ represent the number of CU sizes and intra directions, and $f_h = 3$ represents the average number of motion vector candidates. It should be noted that the actual time complexity of both VVC and HEVC is dependent on the implementation and hardware used. Nevertheless, the enhanced compression efficiency and additional processing steps of VVC are expected to result in a higher time complexity than that of HEVC.

For the training of the rate constraint 3D CNNs model, an offline approach is used, while the online view synthesis component solely employs 3D CNNs. The time complexity of view synthesis is expressed in terms of 3D CNNs as $O(\sum_{l=1}^{L} w \cdot h \cdot f \cdot D_t(l) \cdot D(l))$. Here, $l \in [1, L]$ denotes the number of convolutional kernels, with $L = 3$ in our case. Furthermore, $D_t(l)$ refers to the number of convolutional layers, where $D_t(1) = 64$, $D_t(2) = 16$, and $D_t(3) = 1$. Additionally, $D(l)$ denotes the dimensions of each convolutional layer, where $D(1) = 9 \times 9 \times 3$, $D(2) = 1 \times 1 \times 3$, and $D(3) = 5 \times 5 \times 3$. The values of $w = 624$, $h = 432$, and $f = 5$ represent the width, height, and frame number of the pseudo video adopted for view synthesis, respectively. Therefore, the time complexity of the view synthesis is determined by the summation of the product of the number of convolutional kernels, dimensions of each convolutional layer, and the pseudo video's width, height, and frame number.

VIII. FAILURE CASE

The proposed method exhibits weaknesses in high bitrate coding scenarios. To illustrate this issue, we have provided rate distortion curves as shown in Figure 4. Our framework for coding SAIs has shown that the coding performance, particularly the PSNR values, is highly dependent on the image quality of the rate constraint 3D CNNs model-based view synthesis. However, we have observed that our method performs better in low bitrate scenarios than in high bitrate scenarios during the training procedure of the rate constraint 3D CNNs. To improve the coding performance in high bitrate cases, we plan to augment the training datasets in our future work.

IX. CONCLUSION

We propose a closed-loop hybrid coding system with a novel prediction structure, which designs rate constraint-reinforced 3D CNNs to synthesize the views for image reconstruction. Firstly, we adopt the VVC codec to encode and reconstruct a subset. A novel 3D CNNs with a rate constraint reinforcement learning method is proposed to synthesize the complement of the subset on the stacked EPIs. The residual of the 3D CNNs prediction is also transformed, quantified, and entropy-coded to generate the transmitted bits. The proposed coding system achieves significantly higher PSNR and rate savings compared to reference codecs, with impressive rate savings going as high as 41.58% and 23.31% against HEVC_LDP and Bakir2020 in lenslet image coding.

REFERENCES

[1] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 2005.
[2] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.
[3] 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.*, Jun. 2013, pp. 1027–1034.
[4] I. Schiopu and A. Munteanu, "Deep-learning-based lossless image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1829–1842, Jul. 2020.
[5] C. Jia, X. Zhang, S. Wang, S. Wang, and S. Ma, "Light field image compression using generative adversarial network-based view synthesis," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 177–189, Mar. 2019.
[6] J. Hou, J. Chen, and L. Chau, "Light field image compression based on bi-level view compensation with rate-distortion optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 517–530, Feb. 2019.
[7] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–4.
[8] L. Li, Z. Li, B. Li, D. Liu, and H. Li, "Pseudo-sequence-based 2-D hierarchical coding structure for light-field image compression," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1107–1119, Oct. 2017.
[9] N. Bakir, W. Hamidouche, S. A. Fezza, K. Samrouth, and O. Déforges, "Light field image coding using dual discriminator generative adversarial network and VVC temporal scalability," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[10] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, "End-to-end view synthesis for light field imaging with Pseudo 4DCNN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 333–348.
[11] P. Schelkens, P. Astola, E. A. Da Silva, C. Pagliari, C. Perra, I. Tabus, and O. Watanabe, "JPEG Pleno light field coding technologies," *Proc. SPIE*, vol. 11137, pp. 314–324, Sep. 2019.
[12] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[13] M. A. Magnor, A. Endmann, and B. Girod, "Progressive compression and rendering of light fields," in *Proc. VMV*, 2000, pp. 199–204.
[14] A. Aggoun and M. Mazri, "Wavelet-based compression algorithm for still omnidirectional 3D integral images," *Signal, Image Video Process.*, vol. 2, no. 2, pp. 141–153, Jun. 2008.
[15] M. Rerabek, T. Bruylants, T. Ebrahimi, F. Pereira, and P. Schelkens, "ICME 2016 grand challenge: Light-field image compression," Call Proposals Eval. Procedure, EPFL, Switzerland, 2016.
[16] C. Conti, J. Lino, P. Nunes, L. D. Soares, and P. L. Correia, "Improved spatial prediction for 3D holoscopic image and video coding," in *Proc. 19th Eur. Signal Process. Conf.*, Aug. 2011, pp. 378–382.
[17] 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)*, May 2014, pp. 539–543.

[18] L. F. R. Lucas, C. Conti, P. Nunes, L. D. Soares, N. M. M. Rodrigues, C. L. Pagliari, E. A. B. da Silva, and S. M. M. de Faria, "Locally linear embedding-based prediction for 3D holoscopic image coding using HEVC," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2014, pp. 11–15.

[19] D. Liu, P. An, R. Ma, C. Yang, and L. Shen, "3D holoscopic image coding scheme using HEVC with Gaussian process regression," *Signal Process., Image Commun.*, vol. 47, pp. 438–451, Sep. 2016.

[20] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Coding of focused plenoptic contents by displacement intra prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1308–1319, Jul. 2016.

[21] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp. 338–343, Apr. 2000.

[22] S. Shi, P. Gioia, and G. Madec, "Efficient compression method for integral images using multi-view video coding," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 137–140.

[23] S. Kundu, "Light field compression using homography and 2D warping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1349–1352.

[24] S. Adedoyin, W. A. C. Fernando, and A. Aggoun, "A joint motion & disparity motion estimation technique for 3D integral video compression using evolutionary strategy," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 732–739, May 2007.

[25] R. Zhong, S. Wang, B. Cornelis, Y. Zheng, J. Yuan, and A. Munteanu, "L1-optimized linear prediction for light field image compression," in *Proc. Picture Coding Symp. (PCS)*, 2016, pp. 1–5.

[26] R. Zhong, S. Wang, B. Cornelis, Y. Zheng, J. Yuan, and A. Munteanu, "Efficient directional and L1-optimized intra-prediction for light field image compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1172–1176.

[27] R. Zhong, I. Schiopu, B. Cornelis, S.-P. Lu, and A. Munteanu, "Dictionary learning-based, directional and optimized prediction for lenslet image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1116–1129, Apr. 2018.

[28] I. Schiopu and A. Munteanu, "Macro-pixel prediction based on convolutional neural networks for lossless compression of light field images," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 445–449.

[29] T. Zhong, X. Jin, L. Li, and Q. Dai, "Light field image compression using depth-based CNN in intra prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8564–8567.

[30] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7766–7775.

[31] N. Bakir, W. Hamidouche, O. Déforges, K. Samrouth, and M. Khalil, "Light field image compression based on convolutional neural networks and linear approximation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1128–1132.

[32] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–10, Nov. 2016.

[33] Z. Zhao, S. Wang, C. Jia, X. Zhang, S. Ma, and J. Yang, "Light field image compression based on deep learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[34] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[35] B. Bross, "High efficiency video coding (HEVC) text specification draft 9 (SODIS)," in *Proc. 11th JCT-VC Meeting*, 2013.

[36] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Anal. Chim. Acta*, vol. 185, pp. 1–17, 1986.

[37] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[38] M. Rerabek, L. Yuan, L. A. Authier, and T. Ebrahimi, *EPFL Light-Field Image Dataset*, Standard ISO/IEC JTC 1/SC 29/WG1 Contribution, 2015.

[39] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, 2016.

[40] J. Chen, J. Hou, and L. Chau, "Light field compression with disparity-guided sparse coding based on structural key views," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 314–324, Jan. 2018.

[41] Y. Zhang, L. Wan, Y. Mao, X. Huang, and D. Liu, "Geometry-aware view reconstruction network for light field image compression," *Sci. Rep.*, vol. 12, no. 1, p. 22254, Dec. 2022.

[42] G. Bjontegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document SG16 VCEG-M33, ITU, Geneva, Switzerland, 2001.

[43] J. Shi and C. Guillemot, "Light field compression via compact neural scene representation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

XIAODA ZHONG is currently pursuing the bachelor's degree with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology (SCUT). His research interests include video compression and image coding and natural language processing.

TAO LU (Member, IEEE) received the Ph.D. degree in communication and information systems from Wuhan University, in 2013. He is currently a professor and a doctoral supervisor. He completed his postdoctoral fellowship with the Department of Electronic and Computer Engineering, Texas A&M University, from March 2015 to March 2017. His research interests include intelligent systems and multimedia signal processing. He is also a member of various associations, including the Image and Video Communication Committee of the China Society of Image and Graphics, the China Artificial Intelligence Society, and the China Computer Society. He was selected for the Hubei Province Youth Morning Light Program. He has organized various international academic conferences and served as a reviewer for authoritative journals, such as IEEE TRANSACTIONS ON MULTIMEDIA and IEEE TRANSACTIONS ON IMAGE PROCESSING.

DIYANG XIAO is currently pursuing the Graduate degree with the School of Computer Science, Central China Normal University (CCNU). His primary research interests include light field compression, 3D image synthesis, and super resolution.

RUI ZHONG (Member, IEEE) received the bachelor's degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2008, and the Ph.D. degree in computer science from Wuhan University, in 2014. She has been an Associate Professor with the School of Computer Science, Central China Normal University (CCNU), since 2018. Prior to this, she was a Postdoctoral Researcher with the Electronics and Informatics (ETRO) Department, Vrije Universiteit Brussel (VUB), Belgium. Her research interests include video and 3D graphics coding and multimedia transmission over networks.