Multi-Stream Dense View Reconstruction Network for Light Field Image Compression

Deyang Liu, Yan Huang, Yuming Fang, *Senior Member, IEEE*, Yifan Zuo and Ping An, *Member, IEEE*

***Abstract*—Recently, many view synthesis-based methods are proposed for high-efficiency light field (LF) image compression. However, most existing methods fail to recover more texture details on occlusion regions, which reduces the compression efficiency. In this paper, we propose a multi-stream dense view reconstruction network to further improve LF image compression performance. In our method, only sparsely-sampled LF views are transmitted and the rest of the views are reconstructed at the decoder side. During the reconstruction process, we firstly constitute a multi-disparity geometry (MDG) structure based on the decoded sparse LF views, which can reflect abundant disparity characteristics. Subsequently, a multi-stream view reconstruction network (MSVRNet) is put forward to reconstruct a high-quality dense LF image, which consists of a multi-scale feature fusion sub-network, a fusion reconstruction sub-network, and a detail refinement sub-network. The multi-scale feature fusion sub-network can implicitly lean abundant multiscale geometric structure features from the constituted MDG structure. The fusion reconstruction sub-network and the detail refinement sub-network are respectively utilized to fuse the learned multiscale geometric features and restore more texture details, especially for occlusion regions. Moreover, 3D convolutional operations are adopted in the whole reconstruction process, which allow information propagation among the learned multiscale geometric features. Comprehensive experimental results demonstrate the effectiveness of the proposed method. The perceptual quality of reconstructed views and application on depth estimation also demonstrate that the proposed method can keep structural consistency of the reconstructed LF image and recover more texture details.**

***Index Terms*—Light field image compression, dense view reconstruction, multi-stream network, deep learning**

I. Introduction

LIGHT field (LF) imaging can simultaneously capture spatial position and angular information of light rays in three dimensional (3D) scenes [1]. The additional angular information in dense LF images can facilitate many new applications [2], such as post-capture refocusing [3], 3D reconstruction [4], depth estimation [5] and virtual/augmented reality [6], *et al.*

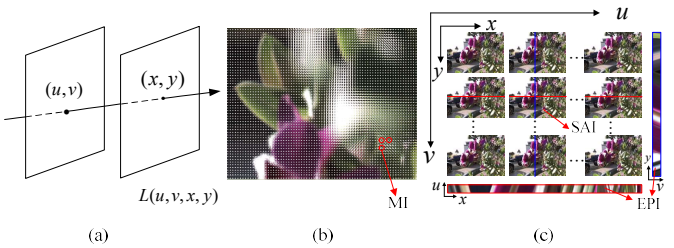


Fig. 1. Illustration of 4D LF data: (a) Parameterization of two-plane of LF imaging; (b) An example of LF lenslet image captured by plenoptic cameras; (c) An example of 4D LF SAI and EPI representation.

Normally, a two-plane parameterization model is used to describe a 4D light field $L = L(u, v, x, y)$ [7], where (x, y) represents two spatial dimensions and (u, v) describes two angular dimensions (see Fig.1 (a)). Based on the 4D representation, many acquisition devices have been proposed to capture the LF image. Especially, with the introduction of commercial and industrial plenoptic LF cameras (*e.g.*, Lytro [8]), the LF imaging enters a new era. For plenoptic LF cameras, an array of microlenses is embedded between the main lens and the camera sensor to capture light rays. Each microlens can capture one micro-image (MI) of the 3D scene at a slightly different angle to its neighbors. An example of the LF lenslet image captured by a plenoptic camera is shown in Fig.1 (b). An array of sub-aperture images (SAIs) can be further extracted from the LF lenslet image, where adjacent SAIs (also referred as views in this paper) contain the same 3D scene with a small disparity, which can be seen in Fig.1 (c). By fixing one angular dimension and one spatial dimension of the SAI array, an LF Epipolar Plane Image (EPI) can be achieved (see in Fig.1 (c)).

LF imaging facilitates many multimedia applications, however, the extremely large volume of LF data raises great challenges for both data transmission and processing. For example, each raw LF image captured by a plenoptic LF camera needs around 50 MB storage space. The bulky data cannot be efficiently transmitted. Thus, high efficiency LF data compression method is strongly demanded for LF applications.

Many efforts have been devoted to the LF image compression [52][53]. For example, the JPEG and MPEG committees have launched studies to standardize the compression of

This work was supported in part by the National Natural Science Foundation of China under Grant 62171002, 62132006, 61901197, in part by Shenzhen Municipal Science and Technology Innovation Council under Grant 2021Szvup051, and in part by STCSM under Grant SKLSFO2021-05. (*Corresponding author: Yuming Fang*).

Deyang Liu is with School of Computer and Information, Anqing Normal University, Anqing 246000, China. E-mail: (deyang.liu@hotmail.com).

Deyang Liu is also with School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330000, China, and Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, 200444 Shanghai.

Yuming Fang and Yifan Zuo are with School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330000, China. E-mail: (leo.fangyuming@foxmail.com, kenny0410@126.com).

Yan Huang is with National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing China. E-mail: (yan.huang@cripac.ia.ac.cn).

Ping An is with School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China. E-mail: (anping@shu.edu.cn).

LF images (referred to as JPEG Pleno [9][40] and MPEG-I [10]). For the LF compression, removing redundancy is the main concern to improve the performance of LF image compression. Based on the LF lenslet representation, many LF image compression methods are proposed to remove the spatial redundancies among MIs using existing image/video coding standards [11]-[14]. Since it is difficult to achieve a high prediction accuracy, SAI based LF compression methods are put forward to further remove LF image spatial and angular redundancies. A classic algorithm is to arrange all the SAIs into a pseudo-video sequence which is then encoded with high efficiency video coding (HEVC) standard [15]-[17]. With learning-based method has been progressing by substantial strides in image and video processing, learning based view synthesis methods are introduced into LF image compression [18]-[23]. Since this type of method can remove more LF redundancies, it becomes the mainstream technology for the LF image compression task.

The main idea of learning based view synthesis method is to encode a sparse set of LF SAIs and then used as the reference to reconstruct the rest of SAIs at the decoder side. During the reconstruction process, occlusions can break photo consistency assumption [24]. Therefore, artifacts are easily introduced on occlusion regions, which significantly influences the compression performance. Mitigating occlusion problem becomes the key issue in improving the LF image compression performance. Many existing learning based view synthesis solutions [25][35] try to decrease the sample rate and transmit more LF SAIs to the decoder side as the LF reconstruction priors to reduce the influence caused by occlusions. Although the occlusion can be suppressed to some extent, massive redundancies retains in sparsely-sampled SAIs. Moreover, since disparity can reveal geometric structure relations of adjacent SAIs [55], many researchers strive to transform the occlusion removal into a disparity estimation problem [26][27]. Disparities of unsampled SAIs are explicitly estimated firstly, and then disparity-based warping is used to reconstruct dense LF images. However, for sparsely-sampled SAIs, the disparity information is lost where occlusions occur. It is difficult to directly estimate accurate disparities, which degrades the dense LF reconstruction quality.

In order to further improve the LF reconstruction quality and compression performance, we follow the learning based view synthesis method and propose a multi-stream dense view reconstruction network for LF image compression. In the proposed method, we only need to encode fewer SAIs in the encoder side (see Fig.2), and reconstruct the rest of SAIs at the decoder side. In order to mitigate the occlusion problem, we implicitly explore the geometric structure relations of sparsely-sampled SAIs instead of explicitly estimating disparities. In the reconstruction procedure, we firstly construct a Multi-Disparity Geometry (MDG) structure based on the decoded sparse SAIs. The MDG structure can reflect abundant disparity characteristics, which benefits in recovering more high-frequency details. Based on the MDG structure, we further put forward a Multi-Stream View Reconstruction Network (MSVRNet) to reconstruct a dense LF image from decoded sparsely-sampled SAIs. In order to expand the receptive field

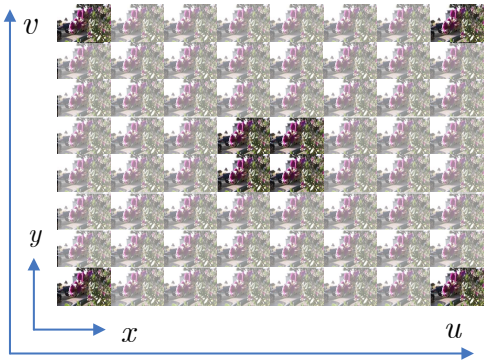


Fig. 2. Sparsely-sampling of dense 8×8 LF image in SAI representation with 8 reference SAIs needed to be encoded.

and implicitly learn abundant geometric structure feature, a multi-scale feature fusion sub-network is integrated into the MSVRNet in each branch, so as to learn multiscale geometric information for the subsequent LF reconstruction task. Moreover, we adopt 3D convolutional operations in the dense LF image reconstruct procedure, which allows information propagation among the learned multiscale geometric structure features. As a result, more high-frequency details can be recovered, especially for occlusion regions. The main contributions of this paper are as follows:

- Compared with previous works, this paper explores more abundant disparity characteristics for LF image compression by constructing a MDG structure based on decoded sparse SAIs. Moreover, in order to take full use of such disparity characteristics, we also put forward a multi-scale feature fusion sub-network to implicitly learn multiscale geometric structure feature from the MDG structure.
- In order to alleviate occlusion problem, a MSVRNet is put forward for view synthesis based LF image compression, in which a multi-stream framework with 3D convolution is adopted. This can allow information propagation among the learned multiscale geometric structure features so as to restore more texture details, especially for occlusion regions.
- Comprehensive experimental results demonstrate that the proposed method can achieve significant bits saving compared with some state-of-the-art solutions.

The remainder of this paper is organized as follows. Section II reviews the related works of LF image reconstruction and view synthesis based LF image compression. The proposed LF image compression method is introduced in Section III. Section IV gives the simulation results and analyses. Section V is devoted to conclusions.

II. Related work

LF imaging enables many applications. However, the bulky data raises a severe challenge for LF applications due to the limited transmission bandwidth and storage space. To address this problem, many researchers have dedicated to exploring efficient LF compression methods. Generally, LF compression methods can be roughly categorized into two groups based on the two representation forms of LF data [31][35], *i.e.,* lenslet

image and sub-aperture image. The lenslet-image-based LF compression method attempts to improve LF coding efficiency by exploring the high spatial correlations among adjacent MIs. While, the sub-aperture-image-based LF compression method achieves high coding performance by eliminating spatial and angular redundancies of LF SAI array. Especially, with learning based method achieving promising results in LF processing [28][29], learning based view reconstruction method becomes the mainstream technology for sub-aperture-image-based LF compression. This section will briefly revisit LF image compression methods, and the learning based LF reconstruction method.

A. Lenslet-image-based LF image compression

Due to the fact that the adjacent MIs correlate strongly, the lenslet-image-based methods aim to improve LF compression efficiency by exploring such correlations. Conti *et al.* [11] introduced a self-similarity (SS) prediction mode into HEVC standard to take full use of MI cross-correlation to achieve high coding efficiency. Monteiro *et al.* [12] proposed a two-stage high-order intrablock prediction method to predict each image block by utilizing the spatial redundancies in the lenslet images. Liu *et al.* [13] put forward a Gaussian process regression-based LF classification prediction scheme based on the LF lenslet image representation, and a high prediction accuracy was acquired by exploring the spatial correlations of MIs. Jin *et al.* [14] introduced three prediction modes into HEVC intra prediction framework, *i.e.*, multi-block weighted prediction mode, co-located single-block prediction mode, and boundary matching based prediction mode, to fully exploit spatial correlations among MIs. Schiopu *et al.* [30] proposed a LF lossless coding approach based on macro-pixel synthesis technique, where the entire LF image can be synthesized in one step.

The main idea of lenslet-image-based method is to remove spatial redundancies of MIs by exploring the high spatial correlations among adjacent MIs. Prediction accuracy is the main bottleneck to further improve LF coding efficiency. Moreover, since angular information is not fully considered, this kind of method cannot keep LF geometry consistency well [35], which inevitably reduces the perception quality of compressed LF image. Therefore, many researches focus on sub-aperture-image-based LF compression method, which will be reviewed in the next sub-section.

B. Sub-aperture-image-based LF compression

Sub-aperture-image-based LF compression intends to improve LF compression performance by removing more spatial and angular redundancies of LF SAI array. For example, the disparity-based view synthesis method is introduced into LF compression [31-35], where a subset of SAIs and corresponding disparity views are fed into the encoder and all texture views are reconstructed at the decoder side. This kind of method allows to transmit only a sparse set of LF SAIs and corresponding disparity maps to decoder side, and synthesize the rest of SAIs using disparity based reconstruction. However, the LF compression performance by using this kind of method is susceptible to the quality of estimated disparity. In order to further improve LF compression efficiency, many researches focus on prediction based LF compression method. For instance, Miandji *et al.* [36] proposed to train a multidimensional dictionary ensemble to sparsely represent LF data, which was then used to synthesize high-resolution LF image. Ravishankar *et al.* [37] presented an efficient layer-based representation method for lossy LF compression, which could approximatively predict dense LF. This method could also realize multiple bitrates flexibly by adjusting the ranks of BK-SVD representation and HEVC quantization. Mukati *et al.* [38] presented a distributed LF compression method, where side information was generated using a learning-based view synthesis method and then was used to predict scene geometry and inpaint occlusion. Ahmad *et al.* [39] proposed a shearlet transform based prediction scheme to reconstruct dense LF image at decoder side to improve LF compression performance under Low Bitrates.

With learning based LF reconstruction making rapid progress, many research efforts are carried out for LF compression by using learning based LF reconstruction. Jia *et al.* [18] introduced the Generative Adversarial Network (GAN) into LF image compression, where dense LF SAIs in certain positions were generated with decoded sparsely-sampled SAIs. However, this method needs the surrounding sampled SAIs of one target SAI to be fed into the multi-branch fusion network as priors to generate a high order approximation, causing the sparse sampling rate becomes low. Hou *et al.* [19] explored the LF image inter- and intra-view correlations, and proposed an Bi-level view compensation method. The non-key SAIs were synthesized by using a learning-based angular super-resolution network. Wang *et al.* [20] put forward a multi-branch spatial transformer network for LF image compression. In this method, the affine transformations between neighboring SAIs were learned to generate high-quality target SAIs. Bakir *et al.* [21][22] put forward a dual discriminator GAN for LF view synthesis, where a sparse set of reference views were encoded using versatile video coding standard and the non-reference views were reconstructed with the dual discriminator GAN. Although this method achieves a high coding performance, the disparity still needs to be estimated. By doing so, it may influence the reconstruction quality, especially for some occlusion regions. Our previous work [23] exploited a sparse representation and constructed a GAN based view synthesis network to enhance the quality of reconstructed SAIs. Chen *et al.* [41] tried to learn a global multiplane representation for LF compression, where two steps including disparity-based global representation prediction and view prediction with the multiplane translation were used to improve the compression performance.

LF compression with learning based LF reconstruction exploits LF sparsely-sampling at the encoder side and learning based LF reconstruction at the decoder side. Increasing the LF reconstruction quality is the key problem to improve the LF compression performance, and occlusion is the primary concern. Therefore, how to improve reconstruction quality by alleviating occlusion problem is the key issue to further improve LF image compression performance. So far, most existing

solutions prefer to mitigate the occlusion problem by explicitly estimating disparity or decreasing the sample rate, which is quite limited in exploring the geometric structure relations of sparsely-sampled SAIs to improve the reconstruction quality. Therefore, this paper follows the idea of LF compression with learning based LF reconstruction, and tries to implicitly explore the geometric structure relations of sparsely-sampled SAIs instead of explicitly estimating disparities to further increase LF reconstruction quality. In the next sub-section, we will briefly review the learning based LF reconstruction method.

C. Learning based LF reconstruction

Learning based LF view reconstruction aims to reconstruct dense LF SAIs with sparsely-sampled SAIs as input. As early as 2016, Kalantari *et al.* [42] proposed a learning-based LF view synthesis method, where the synthesis process consists of disparity and color estimation components. By training two sequential convolutional neural networks (CNNs), high-quality LF SAIs at arbitrary locations can be synthesized. Following the disparity based view synthesis method, Choi *et al.* [43] put forward a LF view synthesis method with few SAIs as inputs. Instead of exact depth estimation, a depth probability volume was estimated to alleviate occlusion problem followed by a learning based refinement network. Jin *et al.* [44] proposed a leaning-based method to reconstruct a dense LF image with a large baseline. The reconstruction procedure consisted of a depth estimation module, a physically-based warping module, and a LF blending module, which could model the LF geometry information for novel SAI synthesis. Shi *et al.* [45] used a light-weight optical flow estimation network to estimate depth maps, and reconstructed dense LF image in pixel and feature domains respectively. Although the depth based warping and refinement paradigm can reconstruct dense LF SAIs, the depth (also regarded as disparity) information is difficult to estimate where occlusion occurs. Consequently, artifacts are easily introduced in the reconstructed SAIs.

Many researchers start to explore non-depth based dense LF reconstruction method. Yeung *et al.* [46] proposed an end-to-end LF reconstruction network, where spatial-angular alternating convolutions were adopted to explore high-dimensional LF spatial-angular clues. Wu *et al.* [47] transformed the LF reconstruction into a angular restoration problem on EPI and constructed a "blur-restoration-deblur" framework to improve the reconstruction quality. This method was further improved by fusing a set of sheared EPIs to handle large disparity cases [48]. Wang *et al.* [49] introduced an extended pseudo 4DCNN for high-fidelity LF view synthesis, where an EPI structure preserving loss function was applied. To reconstruct dense LF image with arbitrary angular resolution using sparsely-sampled LF with irregular structures, Jin *et al.* [50] proposed a coarse-to-fine LF reconstruction network, where the end-to-end trainable network was divided into a coarse SAI synthesis module and a LF refinement module. Our previous work [51] also put forward a LF reconstruction network by taking advantage of the LF multi-angular epipolar geometry information.

The non-depth based dense LF reconstruction achieves high reconstruction quality, however, the abundant disparity characteristics are still ignored in LF reconstruction procedure. Since disparity can reveal geometric structure relations of adjacent SAIs [55], fully exploring such disparity characteristics is important to alleviate the occlusion problem and further improve the LF reconstruction quality. Therefore, in this paper, we construct a MDG structure to explore abundant disparity characteristics of sparsely-sampled SAIs. In addition, a multi-stream view reconstruction network is put forward to implicitly learn such abundant geometric structure features from the MDG structure to reconstruct a high-quality dense LF image. More details can be found in Section III.

III. Proposed method

In this section, we discuss our proposed multi-stream view reconstruction network for LF image compression. The entire flow graph of the proposed LF image compression framework is shown in Fig. 3. As a normal operation, the dense LF SAIs are firstly sparsely-sampled. In our method, only 8 SAIs are sampled from 8×8 dense LF including four corner SAIs and four center SAIs (see Fig. 2). Note that, the adopted sparse sampling method can eliminate more LF redundancies while preserving enough geometric structure information for a high-fidelity LF reconstruction. The sampled SAIs are re-organized into a pseudo-sequence with a specific scan order as shown in Fig. 3. At the encoder side, the HEVC standard [54] is utilized to compress the obtained pseudo-sequence by removing the intra- and inter-SAI redundancies. At the decoder side, the sampled-and-decoded pseudo-sequence is employed as the LF context prior to reconstruct a dense LF image.

In order to mitigate the occlusion problem and improve reconstruction quality, we constitute a MDG structure based on the sampled-and-decoded pseudo-sequence. Subsequently, the constituted MDG structure is fed into the proposed MSVRNet to reconstruct dense LF SAIs, where the MSVRNet consists of a multi-scale feature fusion sub-network (MSFNet), a fusion reconstruction sub-network (FRNet), and a detail refinement sub-network (DRNet). Note that, the proposed MSVRNet is only utilized to reconstruct luma component (Y), while the chrominance components (Cb and Cr) are reconstructed by using bilinear interpolations. The details of our MDG structure constitution and the proposed MSVRNet are introduced in the following sections.

A. Multi-Disparity Geometry Structure Construction

In Sec. I, we mention that the disparity can reveal geometric structure relations of adjacent SAIs [55] while further reflects occlusion relationships. Fully exploiting the geometric structure relations of sparsely-sampled SAIs might be advantageous to improve the reconstruction quality. However, it is difficult to accurately estimate the disparity information for occlusion regions. Thus, we propose to implicitly learn the geometric structure relations from the sparsely-sampled SAIs.

In order to learn the geometric structure relations, a MDG structure is constituted based on the disparity baseline of the sampled-and-decoded SAIs. Note that, the disparity baseline means the largest disparity range among the sparsely-sampled SAIs. In this paper, the MDG structure contains four

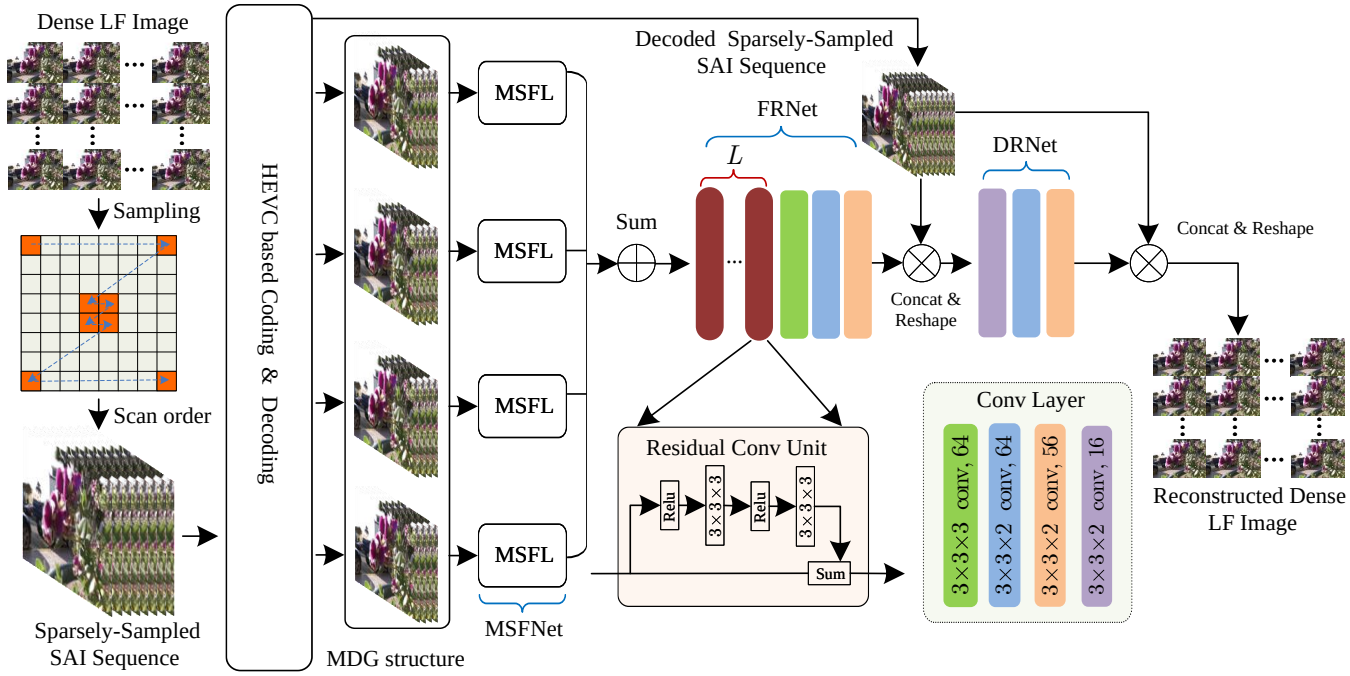


Fig. 3. Flowchart of the proposed compression method by using MSVRNet based LF view reconstruction.

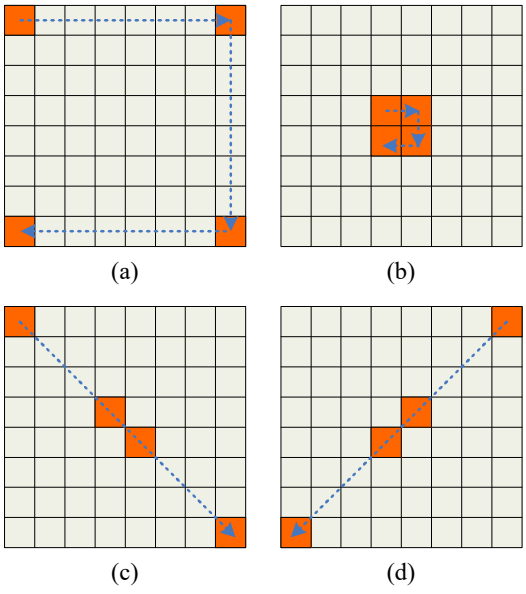


Fig. 4. MDG structure direction selection and constitution in each branch: (a) LF_a; (b) LF_b; (c) LF_c; (d) LF_d.

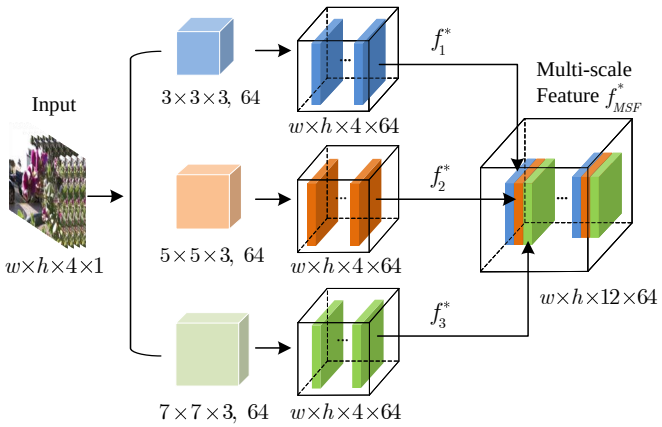


Fig. 5. Illustration of the multi-scale feature fusion layer (MSFL) in each branch.

branches. For each branch, four specific SAIs are selected and concatenated according to the arrow direction shown in Fig. 4. Let LF_a, LF_b, LF_c, and LF_d respectively represent each branch. The constructed MDG structure can be expressed as $LF_{MDG} = \{LF_a, LF_b, LF_c, LF_d\}$.

The constituted MDG structure contains abundant geometric structure relations and disparity characteristics. For example, the disparity range of LF_a is [-7, 7] while only [-1, 1] is presented in LF_b. Abundant disparity characteristics can provide enough geometric structure relations, which benefits in improving reconstruction quality. Based on the MDG structure, a MSVRNet is further put forward to implicitly learn the multiscale geometric structure feature to mitigate the occlusion problem. The details of the proposed MSVRNet are given in the next sub-sections.

B. Multi-scale feature fusion sub-network

Beside the constituted MDG structure, abundant feature representations are also important for high-fidelity dense LF reconstruction. Considering that receptive field expansion and multiscale feature extraction are conducive to obtain abundant

feature representation, in this paper, we propose a multi-scale feature fusion sub-network, which contains four multi-scale feature fusion layers (MSFL) for four branches of MDG structure. As illustrated in Fig. 5, three scales of convolution kernels, including $3\times3\times3$, $5\times5\times3$, and $7\times7\times3$, are adopted for each branch. Since each branch of MDG structure only contains four SAIs in angular dimension, the multiscale feature extraction are only conducted in two spatial dimensions. Thus, all the kernel sizes along the third dimension are set to 3 to preserve more geometry information. Note that, for each convolution kernel, the filter number is set to 64, which means that 64 feature maps can be produced in each branch. For a specified input f_{input}, where $f_{input} \in \{LF_a, LF_b, LF_c, LF_d\}$, the process to obtain the output features in each pipeline f_t^* can be expressed as:

$$f_t^* = C_k(f_{input}), t \in \{1, 2, 3\}; k = 2t + 1 \tag{1}$$

where t represents the t-th pipeline, $C_k(\cdot)$ represents the convolution operation, k is the size of the convolution kernel, f_t^* is the output feature of the t-th pipeline. Subsequently, the obtained features in each pipeline are cascaded together along the third dimension to constitute the multi-scale features f_{MSF}^*. The final multi-scale feature f_{MSF} is constituted by summing the obtained multi-scale features in all branches. In our case, suppose the size of f_{input} in each branch is $w \times h \times 4 \times 1$, the size of output feature of the t-th pipeline is $w \times h \times 4 \times 64$ after the multi-scale convolution operation, where $w \times h$ are two spatial dimensions of training patches. The size of the multi-scale feature f_{MSF}^* is $w\times h\times12\times64$. The final multi-scale feature f_{MSF} is with size of $w \times h \times 12 \times 64$, which is then fed into the subsequent fusion reconstruction sub-network to synthesize dense LF SAIs.

C. Fusion reconstruction sub-network

The fusion reconstruction sub-network (FRNet) aims to learn the geometric structure relations from the extracted multi-scale feature f_{MSF} and reconstruct a dense LF SAIs. The FRNet consists of one residual convolution unit and three 3D convolution layers, as illustrated in Fig. 3. The adopted residual convolution unit is a simplified version of ResNet proposed in [56], where the batch-normalization layers are removed. Moreover, instead of using 2D convolution, we utilize 3D convolution operation to better accommodate the obtained multi-scale feature. Two 3D convolutional layers with size of $3 \times 3 \times 3$ are designed to learn the geometric structure features. Leaky ReLU with $\alpha = 0.2$ is adopted as the activation function. For each convolutional layer, padding is utilized to compensate for the kernel size. This means that the output of each convolutional layer has the same size as the input. In our case, the input of residual convolution unit has the size of $w \times h \times 12 \times 64$, the size of output feature is also $w \times h \times 12 \times 64$.

The subsequent three 3D convolution layers is utilized to reconstruct the dense LF SAIs. In effect, the three 3D convolution layers serve as the downsampling operation along the third dimension of the input high-dimensional feature. For the first 3D convolution layer, the kernel size is $3\times3\times3$ with stride=3. The sizes of all the other 3D convolution layers are $3 \times 3 \times 2$ with stride=2. The downsampling operation used in this paper can not only increase the receptive field and synthesize a dense LF SAIs but also decrease the computational burden. The output sizes of each 3D convolution layer are $w \times h \times 4 \times 64$, $w\times h\times2\times64$, and $w\times h\times1\times56$, respectively. By concatenating the decoded sparsely-sampled SAI sequence with the output of FRNet according to the corresponding coordinates in the original LF image, we can obtain a dense LF SAIs. After a reshaping operation, the size of the acquired dense LF is $w \times h \times 8 \times 8$. The reconstructed dense LF is then fed into the detail refinement sub-network to further recover more high-frequency details.

D. Detail refinement sub-network

The detail refinement sub-network (DRNet) facilitates dense LF reconstruction quality by further exploring the geometric correlations in angular dimension. As shown in Fig. 3, the DRNet contains three 3D convolution layers, and the kernel sizes are all $3 \times 3 \times 2$ for all the three layers. Since the input size of DRNet is $w \times h \times 8 \times 8$, we pad zeros for the first two dimensions to keep the size of output feature maps the same as the input along the first two dimensions. In order to exploit the geometric correlations in angular dimension, downsampling operation is conducted by setting the stride of three convolution layers to 2 along the third dimension, while increasing the fourth dimension of input data by setting the number of filters for each layer to 16, 64, and 56, respectively. The output size of the DRNet is $w \times h \times 1 \times 56$. After a $Concat\&Reshape$ operation with the decoded sparsely-sampled SAI sequence, a dense LF SAIs can be reconstructed.

E. Loss function

Since the main goal of this paper is to reconstruct a high-fidelity dense LF, thus, we adopt the mean square error (MSE) loss to minimize the $\mathcal{L}_2$ distance between the reconstructed dense SAIs and their corresponding ground-truth to supervise our network, which is defined as

$$\mathcal{L}_2 = \sum_u \sum_v \sum_x \sum_y (LF_{HR}^Y(u, v, x, y) - LF_{GT}^Y(u, v, x, y))^2, \tag{2}$$

where $LF_{HR}^Y(u, v, x, y)$ is the synthesized SAI at coordinate position (u, v, x, y), while the $LF_{GT}^Y(u, v, x, y)$ represents its corresponding ground-truth.

IV. Experimental Results

In this section, we firstly give details of our experimental settings, including datasets, training details and evaluation metrics. Then we compare our method with some state-of-the-art works to demonstrate the effectiveness of our method. Ablation studies are conducted to validate the effectiveness of the MDG structure and the influence of MSFNet to overall compression performance. At last, application on depth estimation and the computational complexity of the proposed method are analysed.

Fig. 6. Central SAIs of the nine LF images used for testing: (a) Distance_View, (b) Orchid_White, (c) Vegetables, (d) Stone_Lion, (e) Branches, (f) Orchid_Purple, (g) Railing, (h) Bikes, (i) Aloe.

A. Experimental Settings

In this paper, 100 LF images from the publicly available dataset [42] are selected for training and validation. Nine other LF images are adopted for testing, where central SAIs of the nine test images are shown in Fig. 6. All the selected LF images are captured by Lytro Illum cameras, from which 14×14 SAIs with spatial resolution 376×541 can be extracted. In our method, in order to avoid vignetting and optical distortion, we only select the central 8×8 SAIs in our experiment. The spatial resolution of each SAI is set to 332×496 by cropping the boundary pixels. Moreover, in order to expediently apply HEVC standard, extra pixels with zero values are added around the boundary of each SAI. Therefore, the final spatial resolution of each SAI that used in testing dataset is 352×512. The dense 8×8 SAIs are firstly sparsely-sampled and rearranged into a pseudo-sequence. Before being fed into the HEVC standard, the obtained pseudo-sequence is converted into YUV 420 format, where HEVC Test Model (HM) reference software version 14 is applied to encode the pseudo-sequence with Low Delay P (LDP) coding structure.

Regarding the training process, patches with spatial resolution 64×64 are extracted from each SAI with stride 1 to construct the training dataset. The proposed network model is optimized by the stochastic gradient descent method with batch size of 1. MatConvNet toolbox [57] is utilized to implement the proposed network, and the Gaussian distribution [58] with standard deviation $\sqrt{2/N}$ is used to initialize the weights of the proposed network. Note that, all the biases are initialized to zero. The learning rate is initialized to $1e-6$, and decreased by a factor of 2 for every 1000 epochs.

In order to verify the efficiency of the proposed LF compression method, we integrate the view reconstruction method with HEVC standard. Four state-of-the-art LF compression methods are used for the comparison to evaluate the rate-distortion (R-D) performance of our method. The corresponding abbreviations are listed below:

1) LSM: Proposed in [15], where all SAIs are firstly arranged into a pseudo sequence with Line Scan Mapping (LSM) and then encoded with HEVC standard;

2) FLFRNet: Proposed in [46], in which the dense SAIs are firstly sparsely-sampled. The sampled SAIs are rearranged into a pseudo sequence, and then encoded with HEVC. The rest of SAIs are reconstructed by using the Fast Light Field Reconstruction Network (FLFRNet) with the sampled-and-decoded SAIs as priors at the decoder side;

3) MALFRNet: Proposed in [51], where the sampling and compression process at the encoder side is similar to FLFRNet. However, at the decoder side, the rest of SAIs are reconstructed by using the Multi-Angular LF Reconstruction Network (MALFRNet) with the sampled-and-decoded SAIs as priors;

4) GCCM: Proposed in [35], where the sparse SAIs and the corresponding disparity maps are transmitted and the rest of SAIs are reconstructed by using a low bitrate light field compression method by considering the Geometry and Content Consistency.

Note that, FLFRNet, MALFRNet and GCCM are all intended to reconstruct dense LF images by using sparsely-sampled SAIs at the decoder side. FLFRNet employs spatial-angular alternating convolutions to accurately characterize high-dimensional spatial-angular clues of LF data to improve the LF reconstruction quality, while, MALFRNet aims to increase the reconstruction quality by exploring rich LF angular information and constitutes a multi-angular reconstruction network for high-efficiency LF compression. However, abundant disparity characteristics are all ignored in these two methods. Unlike these methods, our method constructs a MDG structure to explore such abundant disparity characteristics of sparsely-sampled SAIs. Moreover, we also construct a multi-stream view reconstruction network to implicitly learn such abundant geometric structure features from the MDG structure to reconstruct a high-quality dense LF image. As for GCCM, it needs to transmit sparse SAIs and their corresponding disparity maps to the decoder side. Conversely, our method does not need to estimate disparity maps. Abundant disparity characteristics can be implicitly learned by constructed multi-stream view reconstruction network, which is beneficial to restore more texture details.

The BD-PSNR and BD-Rate [59] are utilized as objective quality metric to evaluate the RD performance, while the average structural similarity index (SSIM) is adopted as the perceptual quality metric. The average PSNR and SSIM in Y channel can be calculated as follows [60]:

$$PSNR_Y^{Avg} = \frac{1}{8 \times 8} \sum_{m=1}^{8} \sum_{n=1}^{8} PSNR[m][n] \tag{3}$$

$$SSIM_Y^{Avg} = \frac{1}{8 \times 8} \sum_{m=1}^{8} \sum_{n=1}^{8} SSIM[m][n] \tag{4}$$

where $PSNR[m][n]$ and $SSIM[m][n]$ respectively are the PSNR and SSIM value of reconstructed dense SAIs at angular coordinate (m, n).

B. Comparison Results and Analyses

This paper proposes a multi-stream dense view reconstruction network for LF image compression. The primary purpose

TABLE I
Y-RATE-DISTORTION PERFORMANCE OF THE PROPOSED METHOD COMPARED WITH FOUR STATE-OF-THE-ART METHODS

LF Images	Proposed vs. LSM		Proposed vs. FLFRNet		Proposed vs. MALFRNet		Proposed vs. GCCM	
	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)	BD-PSNR (dB)	BD-Rate (%)
Distance_View	1.50	-40.72	0.54	-14.00	0.53	-14.02	0.62	-20.64
Orchid_White	2.32	-47.62	0.94	-16.57	0.63	-9.69	1.87	-46.66
Vegetables	1.52	-36.87	0.43	-8.58	0.29	-5.15	0.19	-5.51
Stone_Lion	0.90	-25.09	0.17	-6.62	0.10	-3.64	0.41	-14.55
Branches	2.28	-49.61	0.82	-16.03	0.32	-4.49	1.15	-31.17
Orchid_Purple	2.23	-45.28	0.70	-12.10	0.33	-4.08	2.17	-49.03
Railing	1.23	-27.96	1.66	-32.92	1.20	-24.63	0.61	-17.89
Bikes	2.09	-45.04	0.76	-14.68	0.55	-9.35	0.52	-14.37
Aloe	2.07	-40.15	0.91	-19.31	0.61	-13.79	0.64	-15.56
Average	1.79	-39.82	0.77	-15.65	0.51	-9.87	0.91	-23.93

is to reconstruct a high-quality dense LF image to further improve LF compression performance. In this subsection, the effectiveness of our method is verified from five aspects.

1) Compression Performance and Analyses: The Y-RD-performance comparison between the proposed method and four state-of-the-art methods is shown in Table I, which verifies the superiority of the proposed method. From Table I, one can find that the proposed method achieves an average of 1.79 dB BD-PSNR gain and up to 2.32 dB BD-PSNR gain for LF scene *Orchid_White* as compared to LSM method. The main reason lies in two aspects. One is that the proposed method can remove more LF image spatial and angular redundancies, and the other is that the proposed method can reconstruct a high-quality dense LF SAIs. When compared with FLFRNet and MALFRNet, the average gains are 0.77 dB and 0.51 dB, respectively. The reason is twofold. First, FLFRNet focuses on exploring LF spatial-angular clues in LF reconstruction, while MALFRNet places more emphasis on exploring LF angular information in LF reconstruction. The disparity characteristics are ignored, which take important role in mitigating the occlusion problem during LF reconstruction. The proposed method constructs a MDG structure to explore abundant disparity characteristics of sparsely-sampled SAIs, and the abundant geometric structure features can be implicitly learned from the MDG structure, which results in a high-quality LF reconstruction. Second, 3D convolution is adopted in the whole reconstruction network of our method, which allows information propagation among the learned multiscale geometric structure features so as to restore more texture details. Moreover, the proposed multi-scale feature fusion sub-network can expand receptive field and extract multi-scale features, which also benefits in mitigating the occlusion problem and reducing artifacts in reconstructed LF SAIs. Compared with GCCM, one can find from Table I that the average gain is 0.91 dB. Since the GCCM has to estimate depth information and use depth-based warping to synthesize dense SAIs, artifacts are easily introduced where occlusion occurs. Even though a geometry consistency improvement algorithm and a content-similarity-based prediction algorithm are put forward to improve the quality of estimated depth, the reconstructed LF quality is still not very high. Conversely, our method can reconstruct high-quality LF SAIs by circumventing depth information with the help of the proposed end-to-end multi-stream LF reconstruction network. Therefore, a high LF compression performance can be achieved by our method.

From Table I, we can also observe that the average BD-PSNR gain achieved by our method is not very high compared with other four methods for some LF scenes with consistent texture structure. For instance, for LF scene *Stone_Lion*, the average gains obtained by the proposed method over other methods are 0.90 dB, 0.17 dB, 0.10 dB, and 0.41 dB, respectively. The main reason lies in that the LF scene *Stone_Lion* has a consistent texture structure with few occlusion situation. An accurate intra and inter predictions can be achieved by LSM method. FLFRNet and MALFRNet can also reconstruct a high-quality dense LF SAIs. Regarding the GCCM, since the texture structure of *Stone_Lion* is consistent, high-quality depth information can be estimated and, subsequently, high-quality LF SAIs can be synthesized. However, for LF scenes with occlusion, *i.e. Branches*, *Railing* and *Aloe*, our method has an enormous advantage than the other compared methods, which further verifies that the MDG structure and MSVRNet can mitigate occlusion problem and reconstruct high-quality LF SAIs.

Fig. 7 gives the rate-distortion curves of all the test LF images, which further demonstrates the superiority of the proposed method. The result is consistent with Table I. In Fig. 7, it can be observed that the superiority of the proposed method is limited at low bitrate compared with other methods. The main reason is that the compression noises are not considered during the training procedure of MSVRNet, and more compression noises remain in the decoded sparse SAIs for low bitrate, which influences the LF reconstruction quality. Fig. 8 gives the perceptual quality comparison of reconstructed SAIs for high and low bitrates by using the proposed MSVRNet. From Fig. 8, we see that the perceptual qualities of reconstructed SAIs become better for high bitrate, which is also consistent with Fig. 7.

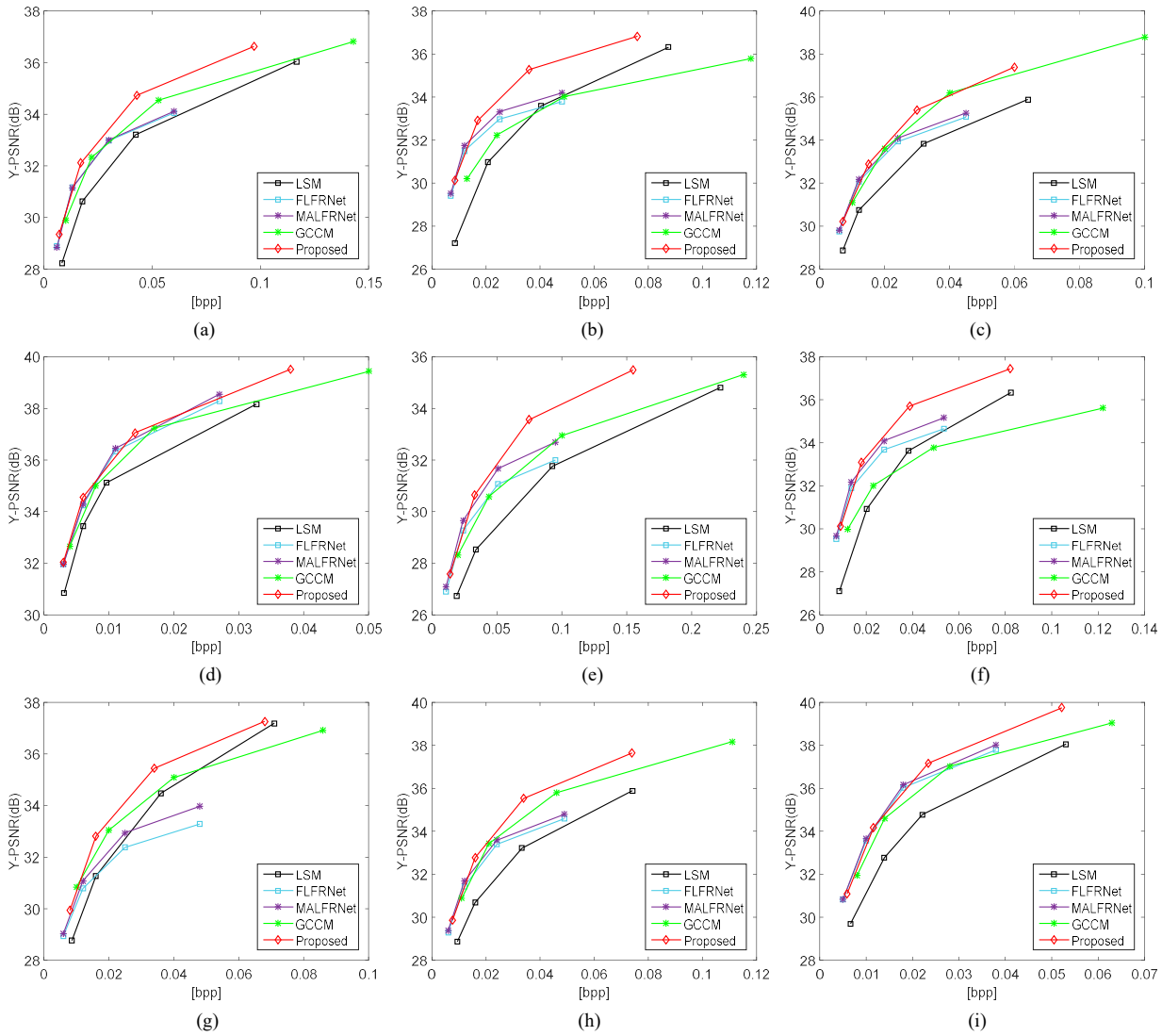


Fig. 7. The rate-distortion curves for test LF images: (a) Distance_View, (b) Orchid_White, (c) Vegetables, (d) Stone_Lion, (e) Branches, (f) Orchid_Purple, (g) Railing, (h) Bikes, (i) Aloe.

TABLE II
BD-RATE (%) SAVING BY THE PROPOSED METHOD VERSUS MULE PROPOSED IN [40], D2GAN PROPOSED IN [21] AND LFC-GAN PROPOSED IN [23].

LF Images	vs. MuLE	vs. D2GAN	vs. LFC-GAN
Ankylosaurus Dip1	-8.1%	-3.3%	-15.1%
Bikes	-20.2%	-9.5%	-17.7%
Danger De Mort	-14.3%	-4.8%	-13.8%
Flowers	-7.6%	-4.3%	-21.6%
Average	-12.55%	-5.48%	-17.05%

In order to further verify the superiority of the proposed method, we compare our method with three other state-of-the-art methods, *i.e.*, MuLE [40], D2GAN [21] and LFC-GAN [23], adopting four LF images from EPFL dataset [62], *i.e.*, *Ankylosaurus Dip1*, *Bikes*, *Danger De Mort* and *Flowers*, for testing. The BD-Rate saving by the proposed method versus these three methods is shown in Table II.

MuLE [40] is the standard LF coding method provided by JPEG-Pleno, where 4D-Transform mode is used to explore LF redundancies across four dimensions. When compared with MuLE, the proposed method can save average 12.55% BD-rate over the four test LF images. The main reason is that the proposed MDG structure can remove more LF redundancies, and the proposed MSVRNet can ensure high-quality LF reconstruction. D2GAN proposed in [21] intends to drop a sub-set of SAIs at the encoder side and generate them at decoder side using a dual discriminator generative adversarial network. From Table II, we find that the proposed method outperforms the D2GAN by 5.48% BD-rate reduction. This is because that the D2GAN has to estimate disparities of dropped SAIs, and reconstruct dropped SAIs through depth-assisted synthesis. However, since occlusion breaks photo consistency assumption, it is hard to estimate an accurate disparity for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3175023, IEEE Transactions on Multimedia

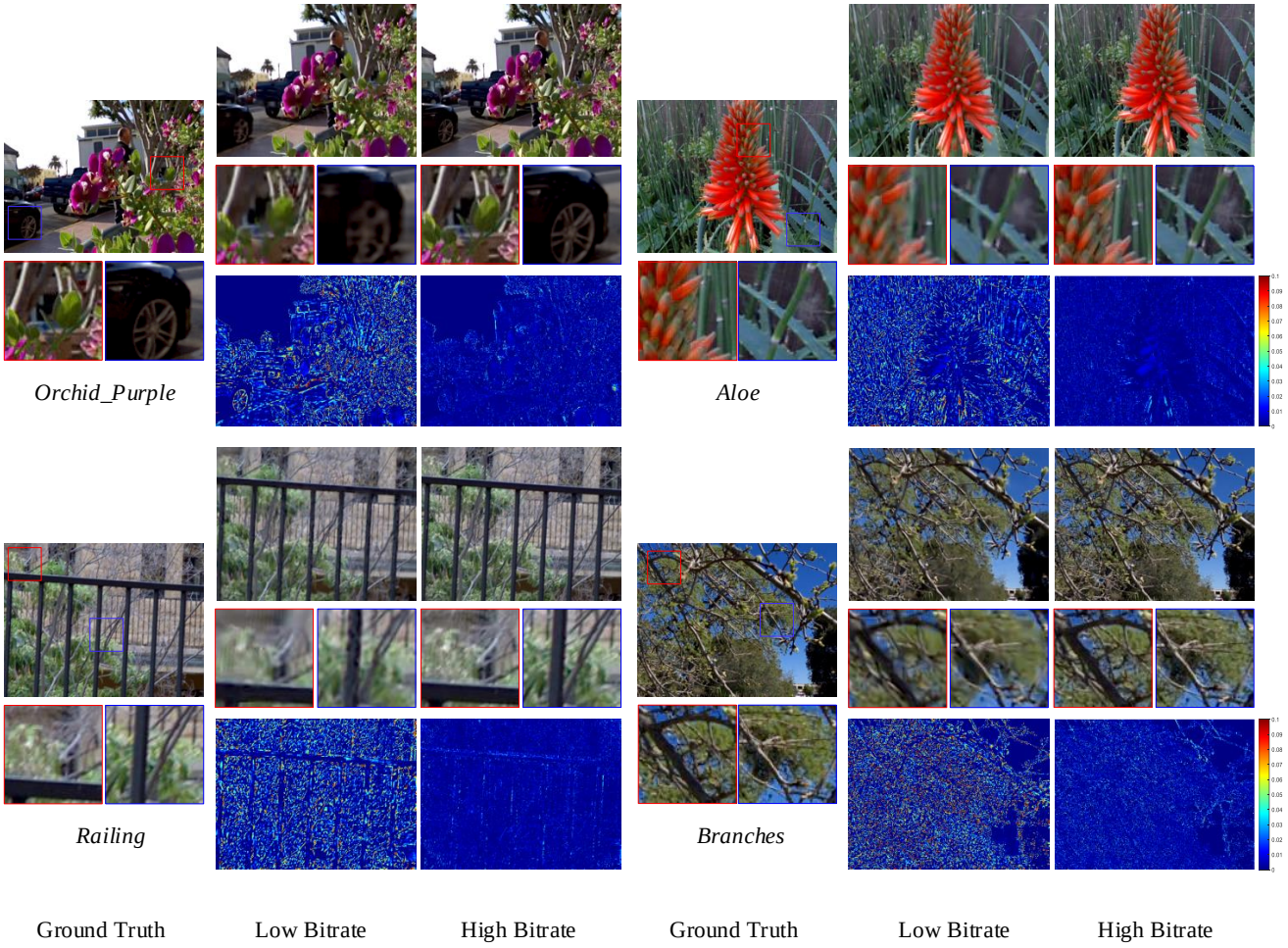


Fig. 8. Perceptual quality comparison of reconstructed SAIs for high and low bitrates by using the proposed MSVRNet.

a dropped SAI, which influences the reconstruction quality. Our previous work [23] proposes a view synthesis-based LF Compression by using Generative Adversarial Network (LFC-GAN). Similar to D2GAN, the LFC-GAN sparsely samples SAIs at the encoder side and tries to reconstruct dense LF using GAN at the decoder side. Since the reconstruction procedure also contains a disparity estimation process, the reconstruction quality is also affected. Conversely, the proposed MSVRNet can implicitly explore abundant disparity characteristics from constructed MDG structure and synthesize dense LF SAIs by circumventing depth information. Therefore, an average 17.05% BD-rate can be saved by the proposed method compared with LFC-GAN, which can be seen in Table II.

2) Perceptual quality of reconstructed SAIs: In order to further verify the effectiveness of the proposed method, Fig. 9 gives the perceptual quality comparison of reconstructed SAIs at the decoder side with angular coordinate (2,3) at a similar bpp. From Fig. 9, we find that the proposed method achieves a higher perceptual quality than the other methods. The LSM method only uses interpicture prediction for LF compression, which remains a high degree of LF spatial-angular redundancies in compression procedure. Therefore, more bit stream is needed and the perceptual quality of decoded LF image is not high. Compared with FLFRNet and GCCM, our method can restore more texture details, especially for some occlusion regions, which can be seen in the areas enclosed within the red boxes and blue boxes in Fig. 9. Regarding for the MALFRNet, LF multi-angular epipolar geometry information is explored in reconstruction process to mitigate occlusion problem. Thus, a high SAI reconstruction quality can be obtained. Compared with MALFRNet, we constitute MDG structure, and design a multi-scale feature fusion sub-network to learn abundant feature representations from MDG structure. Combining with the multi-stream reconstruction framework, our method is superior to MALFRNet, which can also be find in Fig. 9.

In order to demonstrate the high perceptual quality of reconstructed SAIs achieved by the proposed method, Table III gives the perceptual quality comparisons of reconstructed SAIs in terms of the average Y-SSIM metric at a similar bpp for tested LF images. The results shown in Table III are consistent with Fig. 9. The proposed method achieves the highest average Y-SSIM value. Average Y-SSIM values obtained by FLFRNet, MALFRNet, and proposed method all outperform 0.9. However, the average Y-SSIM value of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3175023, IEEE Transactions on Multimedia

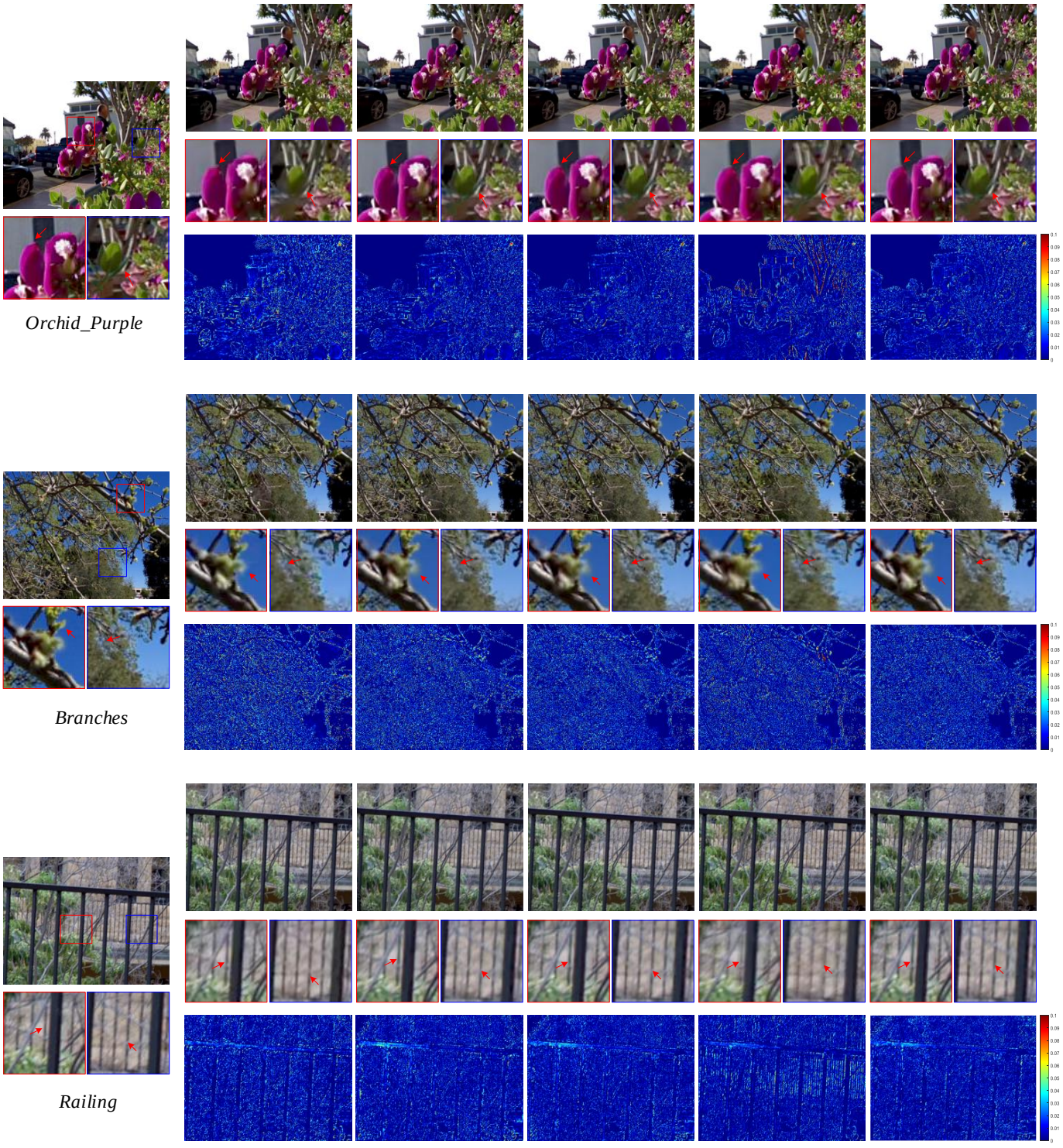


Fig. 9. Perceptual quality comparison of reconstructed SAIs at the decoder side at angular coordinate (2,3) at a similar bpp for three LF images, (*e.g.*, 0.038 bpp for *Orchid_Purple*, 0.09 bpp for *Branches* and 0.036 bpp for *Railing*). The results illustrate the ground truth SAI, reconstructed SAIs by utilizing five different methods, close-up versions of the image portions in blue and red boxes and error maps of reconstructed SAI in Y channel.

TABLE III
PERCEPTUAL QUALITY COMPARISONS OF RECONSTRUCTED SAIS IN TERMS OF THE AVERAGE Y-SSIM METRIC AT SIMILAR BITRATES (BPP) FOR TESTED LF IMAGES (*i.e.*,0.017 BPP FOR DISTANCE_VIEW AND ORCHID_WHITE,0.016 BPP FOR RAILING AND BIKES, 0.015 BPP FOR VEGETABLES, 0.006 BPP FOR STONE_LION, 0.018 BPP FOR ORCHID_PURPLE, 0.075 BPP FOR BRANCHES, AND 0.012 BPP FOR ALOE).

LF Images	LSM	FLFRNet	MALFRNet	GCCM	Proposed
Distance_View	0.8808	0.9121	0.9164	0.7616	0.9200
Orchid_White	0.8874	0.9206	0.9215	0.7952	0.9323
Vegetables	0.8909	0.9271	0.9286	0.8279	0.9304
Stone_Lion	0.8864	0.9062	0.9022	0.5046	0.9072
Railing	0.8701	0.8803	0.9047	0.8343	0.9144
Orchid_Purple	0.9068	0.9323	0.9353	0.8007	0.9425
Branches	0.9136	0.9220	0.9337	0.8413	0.9458
Bikes	0.8630	0.8988	0.9084	0.7908	0.9120
Aloe	0.8881	0.9156	0.9180	0.8148	0.9205
Average	0.8875	0.9128	0.9188	0.7746	0.9250

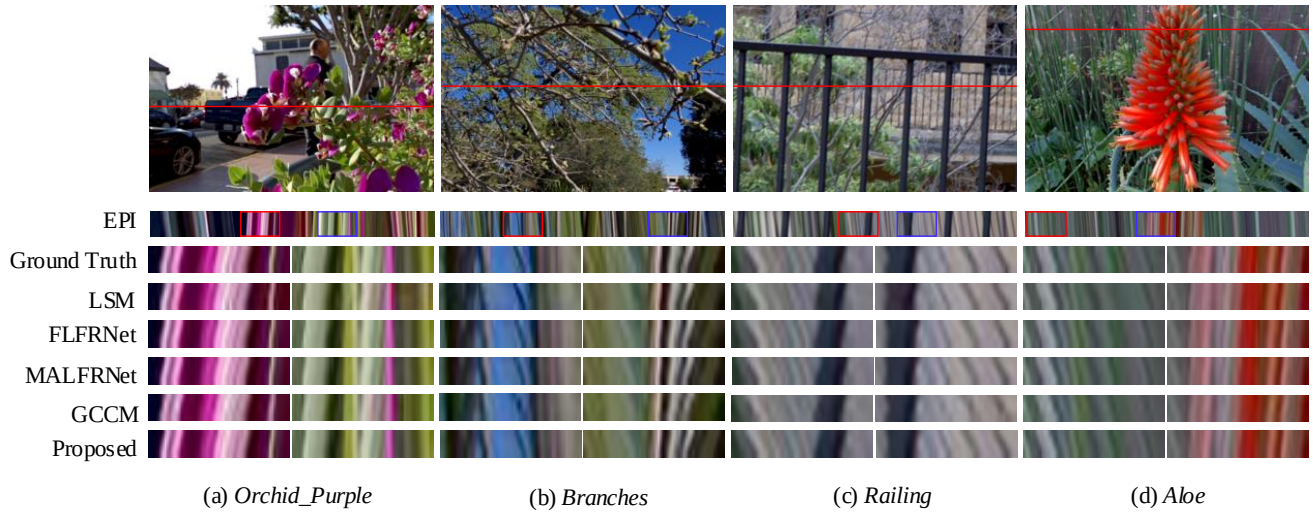


Fig. 10. Structural consistency comparison of reconstructed dense LF image at the decoder side by using various compression methods.

GCCM method is only 0.7746. This is because the depth based warping paradigm easily introduces artifacts and reduces the SAI reconstruction quality.

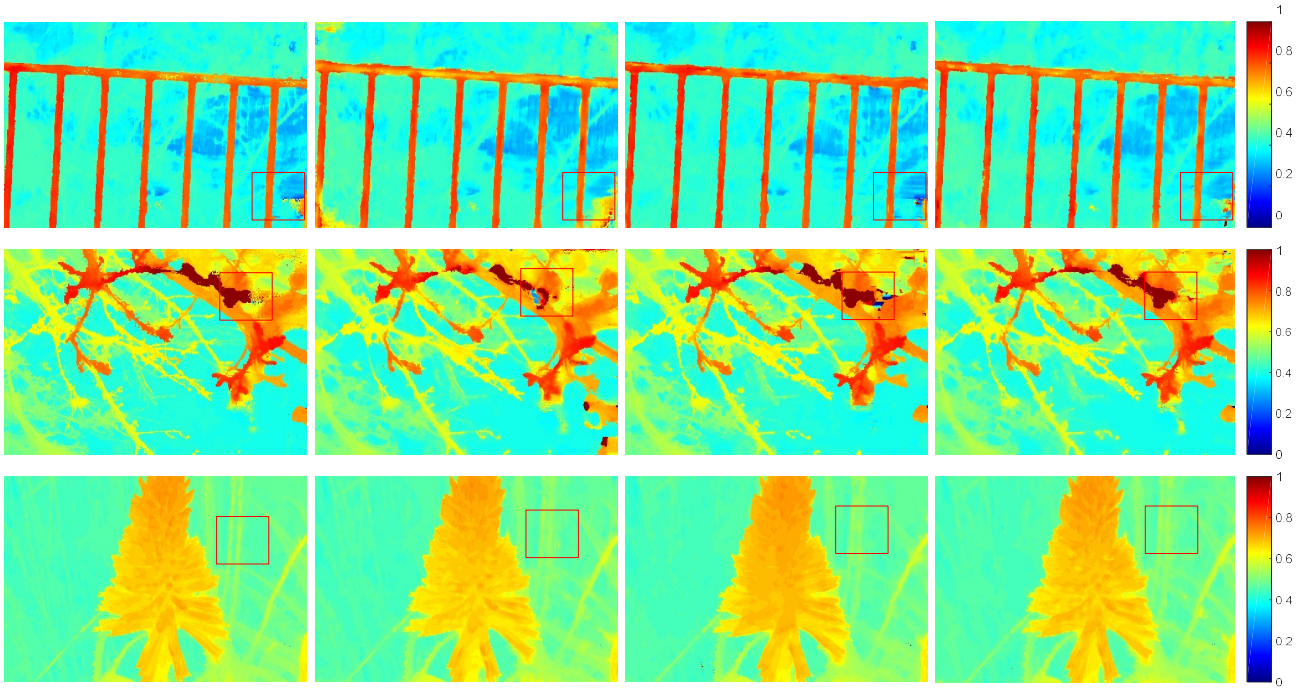
Structural consistency of reconstructed dense LF image is crucial for LF application. Since EPI can reveal the structural consistency of a LF image, we extract EPIs from reconstructed LF images by using the considered compression methods at the decoder side to intuitively demonstrate the superiority of our method. Fig. 10 gives the structural consistency comparison of reconstructed dense LF images at the decoder side by using various compression methods. From Fig. 10, one can find that the proposed method can achieve a better EPI performance. GCCM tries to synthesize dense LF SAIs by using depth based warping paradigm. Since the depth is difficult to estimate on occlusion regions, the structural consistency is destroyed. Compared with FLFRNet and MALFRNet, our method can also achieve a better reconstructed quality, thereby a better EPI performance can be achieved.

3) Ablation Investigation: MDG structure and multi-stream view reconstruction network contribute to the high compression performance of the proposed method. In this subsection, we will illustrate the influence of these two parts on the compression performance. Three other situations are under consideration: 1) MSVRNet-1, where MDG structure is constituted only with two branches (*i.e.* LF_a and LF_b, which is written as $LF_{MDG} = \{LF_a, LF_b\}$); 2) MSVRNet-2, where MDG structure is constituted only with three branches (*i.e.* LF_a, LF_b and LF_c, which is written as $LF_{MDG} = \{LF_a, LF_b, LF_c\}$); 3) MSVRNet-3, where multi-scale feature fusion sub-network is removed.

Table IV gives the RD-performance of the proposed method compared with three other situations with LSM as the anchor. From Table IV, we find that the proposed method achieves the highest RD-performance. Compared with MSVRNet-1 and MSVRNet-2, around 0.44 dB and 0.27 dB average PSNR gains are achieved by the proposed method. This illustrates that the reconstruction quality increases with the addition of

TABLE IV
Y-RATE-DISTORTION PERFORMANCE OF THE PROPOSED METHOD COMPARED TO THREE OTHER SITUATIONS WITH LSM AS THE ANCHOR AND THE AVERAGE Y-SSIM ACHIEVED BY THE FOUR METHODS ACROSS ALL QPS.

LF Images	MSVRNet-1			MSVRNet-2			MSVRNet-3			Proposed		
	BD-PSNR (dB)	BD-Rate (%)	Y-SSIM	BD-PSNR (dB)	BD-Rate (%)	Y-SSIM	BD-PSNR (dB)	BD-Rate (%)	Y-SSIM	BD-PSNR (dB)	BD-Rate (%)	Y-SSIM
Distance_View	1.12	-33.84	0.9241	1.28	-36.99	0.9244	1.24	-31.67	0.9250	1.46	-40.14	0.9253
Orchid_White	1.56	-41.33	0.9345	1.66	-42.83	0.9352	1.78	-38.88	0.9362	1.93	-45.82	0.9362
Vegetables	0.55	-21.66	0.9369	1.39	-35.22	0.9372	1.28	-33.7	0.9370	1.59	-38.54	0.9371
Stone_Lion	0.88	-26.44	0.9183	0.90	-26.88	0.9182	0.91	-26.59	0.9178	0.98	-29.60	0.9170
Branches	2.09	-46.77	0.8998	2.16	-47.66	0.9005	2.11	-49.28	0.9026	2.36	-49.88	0.9020
Orchid_Purple	1.59	-39.34	0.9445	1.71	-40.97	0.9452	1.77	-39.55	0.9456	1.96	-43.83	0.9459
Railing	0.63	-18.76	0.9176	0.70	-20.19	0.9180	0.81	-17.29	0.9188	1.11	-27.19	0.9201
Bikes	1.94	-43.52	0.9199	2.01	-44.39	0.9203	1.90	-38.50	0.9207	2.16	-46.25	0.9206
Aloe	1.85	-37.66	0.9253	1.99	-39.61	0.9259	2.00	-39.77	0.9257	2.62	-48.46	0.9260
Average	1.36	-34.37	0.9245	1.53	-37.19	0.9250	1.53	-35.03	0.9255	1.80	-41.08	0.9256



(a) Ground Truth (b) FLFRNet (c) GCCM (d) Proposed

Fig. 11. Quality comparisons of depth maps estimated from reconstructed LF image by using different compression methods at decoder side. The LF images from top to bottom are Railing, Branches, and Aloe. The ground truth depth maps are extracted from original 8×8 SAIs.

branches within the MDG structure, which also demonstrates the advantage of the MDG structure in improving the dense LF reconstruction quality. Regarding the MSVRNet-3, the average PSNR gain of the proposed method outperforms MSVRNet-3 by 0.27 dB. The main reason is that the multi-scale feature fusion layer can learn abundant feature representation from each branch with expansive receptive field. Abundant features benefit the restoration of texture details, especially for occlusion regions, which further demonstrates the effectiveness of the proposed multi-scale feature fusion sub-network. Table IV also gives the average Y-SSIM achieved by the four LF compression methods across all QPs. Our method obtains the highest average Y-SSIM value, which further verifies the effectiveness of the MDG structure and the multi-scale feature fusion sub-network in keeping the structural consistency of reconstructed dense LF.

4) Application on depth estimation: Depth estimation from LF image becomes a rising research topic because of its capacity to capture both spatial and angular information of 3D scenes. High-quality reconstructed LF image can improve the

TABLE V
COMPUTATIONAL TIME COMPARISONS OF ENCODER AND DECODER SIDE BY USING DIFFERENT LF COMPRESSION METHODS (IN SECOND)

Methods	Computational Time		
	Encoding	Decoding	Reconstruction
LSM	310	0.55	0.00
FLFRNet	31.3	0.08	0.52
MALFRNet	35.3	0.11	1.02
Proposed	42.7	0.16	2.61

quality of extracted depth maps. Therefore, in order to verify that the proposed method can reconstruct a high-quality dense LF image, Fig. 11 gives the quality comparison of extracted depth maps from reconstructed LF images by using different compression methods at the decoder side. Note that, the robust occlusion-aware depth estimation method put forward in [61] is utilized to extract scene depth maps. From Fig. 11, we find that the proposed method can achieve a better quality of estimated depth map compared with FLFRNet and GCCM, especially for some occlusion regions. For example, for the *Branches* case, since occlusions exist in the leaves and twigs (as shown in the red boxes in Fig. 11), texture details are difficult to restore in the reconstructed LF image by using FLFRNet and GCCM. As a result, some estimated errors are occurred in depth maps. Conversely, our method can reduce artifacts in occlusion regions and recover more details, which can produce high-quality depth maps. The same results can be achieved for *Railing* and *Aloe* cases, which further validate the advantage of our method on restoring texture details by keeping structural consistency in reconstructed dense LF images at the decoder side.

5) Complexity Analysis: Table V gives the computational time comparisons of encoder and decoder side by using different LF compression methods. Note that, FLFRNet, MALFRNet, and the proposed method are implemented by using CPU and GPU jointly. Specifically, the codec including encoding and decoding by using HEVC standard runs on a CPU, while the reconstruction process based on deep learning runs on a GPU. In our implement, we use Intel i7-7700 CPU with HQ 2.80 GHz and NVIDIA Quadro P5000 GPU.

For LSM, since all SAIs are encoded and decoded with HEVC standard, it consumes more time than the other LF compression methods on *Encoding* and *Decoding* processed. Although, FLFRNet, MALFRNet, and the proposed method only need to encode and decode sparsely-sampled SAIs, *Reconstruction* procedures are needed to synthesize dense LF SAIs. The computational complexity is normally high. For instance, the computational time consumed by FLFRNet, MALFRNet, and the proposed method on *Reconstruction* processes is 0.52, 1.02, and 2.61, respectively, even with GPU acceleration. Due to the usage of 3D convolutional operations and multi-scale feature fusion layer at each branch in the proposed network to learn abundant feature representations, the time consumption of our method is more than 2 times that of MALFRNet, and more than 5 times than that of FLFRNet. In the future, lightweight networks will be explored to better achieve this task.

V. CONCLUSION

In this paper, a multi-stream dense view reconstruction network is proposed for light field image compression. Based on LF 4D representation, the dense LF SAIs are firstly sparsely-sampled, and only sparse SAIs are transmitted. At the decoder side, the dense LF SAIs are reconstructed by using the proposed multi-stream view reconstruction network (MSVRNet). In order to mitigate occlusion problem and improve reconstruction quality, the decoded SAIs are firstly rearranged to constitute a MDG structure before being fed into the MSVRNet. Multi-scale feature fusion sub-network is utilized to learn abundant feature representations from constituted MDG structure. Moreover, 3D convolutional operations are used in the whole reconstruction procedure to allow information propagation among the learned multiscale geometric structure features to further recover more texture details, especially for occlusion regions.

Experimental results demonstrate the superiority of the proposed LF compression method. It outperforms state-of-the-art methods in improving average BD-PSNR and reducing average BD-Rate. Moreover, the perceptual quality of reconstructed SAIs and applications on depth estimation also demonstrate that the proposed method can keep structural consistency of reconstructed dense LF images and recover more texture details.

REFERENCES

[1] I. Ihrke, J. Restrepo, and L. Mignard-Debise, "Principles of light field imaging: Briefly revisiting 25 years of research," *IEEE Signal Processing Magazine*, vol. 33, no. 5, pp. 59-69, Sep. 2016.
[2] G. Wu *et al*, "Light Field Image Processing: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926-954, 2017.
[3] J. Fiss, B. Curless, and R. Szeliski, "Refocusing plenoptic images using depth-adaptive splatting," *IEEE International Conference on Computational Photography*, 2014, pp. 1-9.
[4] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 73:1-73:12, 2013.
[5] J. Peng, Z. Xiong, Y. Wang, Y. Zhang and D. Liu, "Zero-shot depth estimation from light field using a convolutional neural network," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 682-696, 2020.
[6] R. S.Overbeck,D. Erickson,D. Evangelakos,M. Pharr, and P.Debevec, "A system for acquiring, processing, and rendering panoramic light field stills for virtual reality," *ACM Transactions on Graphics*, vol. 37, no. 6, pp. 197:1-197:15, 2018.
[7] C. Michael, G. Steven, J., S. Richard, G. Radek, and S. Rick, "The lumigraph," *SIGGRAPH*, 1996.
[8] Lytro, (2011). [Online]. Available: https://www.lytro.com/.
[9] T. Ebrahimi, JPEG Pleno Abstract and Executive Summary, ISO/IEC JTC1/SC 29/WG1 N6922, Sydney, Australia, 2015.
[10] "Working Draft 0.1 of TR: Technical Report on Immersive Media," ISO/IEC JTC1/SC29/WG11/N16718, Geneva, Jan. 2017.
[11] C. Conti, L. D. Soares and P. Nunes, "Light Field Coding With Field-of-View Scalability and Exemplar-Based Interlayer Prediction," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 2905-2920, 2018.
[12] R. J. S. Monteiro et al., "Light field image coding using high order intra block prediction," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1120-1131, Oct. 2017.
[13] D. Liu, P. An, R. Ma, W. Zhan, X. Huang, and A. A. Yahya, "Contentbased light field image compression method with gaussian process regression," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 846-859, 2020.

[14] X. Jin, H. Han, and Q. Dai, "Plenoptic Image Coding Using Macropixel Based Intra Prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3954-3968, 2018.

[15] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenselet image compression scheme based on subaperture images streaming," *IEEE International Conference on Image Processing*, pp. 4733-4737, 2015.

[16] D. Liu, L. Wang, L. Li, X. Zhiwei, W. Feng, and Z. Wenjun, "Pseudosequence-based light field image compression," *IEEE International Conference on Multimedia and Expo Workshops*, pp. 1-4, 2016.

[17] 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 Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1107-1119, 2017.

[18] C. Jia, X. Zhang, S. Wang, S. Wang, and S. Ma, "Light field image compression using generative adversarial network-based view synthesis," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 177-189, 2018.

[19] J. Hou, J. Chen, and L. Chau, "Light field image compression based on bi-level view compensation with rate-distortion optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 517-530, 2019.

[20] J. Wang, Q. Wang, R. Xiong, Q. Zhu, and B. Yin, "Light field image compression using multi-branch spatial transformer networks based view synthesis," *Data Compression Conference (DCC)*, pp. 397-397, 2020.

[21] N. Bakir, W. Hamidouche, S. A. Fezza, K. Samrouth and O. Deforges, "Light Field Image Coding Using Dual Discriminator Generative Adversarial Network And VVC Temporal Scalability," *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, 2020.

[22] N. Bakir, W. Hamidouche, S. A. Fezza, K. Samrout and O. Deforges, "Light Field Image Coding Using VVC standard and View Synthesis based on Dual Discriminator GAN," *IEEE Transactions on Multimedia*, 1-14, 2021.

[23] D. Liu, X. Huang, W. Zhan, L. Ai, X. Zheng, S. Cheng, "View synthesis-based light field image compression using a generative adversarial network," *Information Sciences*, vol. 545, pp. 118-131, 2021.

[24] H. Sheng, P. Zhao, S. Zhang, J. Zhang, D. Yang, "Occlusion-aware depth estimation for light field using multi-orientation EPIs," *Pattern Recognition*, vol. 74, pp. 587-599, 2018.

[25] Z. Zhao, S. Wang, C. Jia, X. Zhang, S. Ma, J. Yang, "Light field image compression based on deep learning," *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, 2018.

[26] X. Huang, P. An, L. Shan, R. Ma, L. Shen, "View synthesis for light field coding using depth estimation," *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, 2018.

[27] M. Singh and R. M. Rameshan, "Learning-Based Practical Light Field Image Compression Using A Disparity-Aware Model," arXiv, pp. 1-5, 2021, Available: https://arxiv.org/abs/2106.11558.

[28] G. Wu, et al., "Light Field Image Processing: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926-954, 2017.

[29] K. Ko, Y. J. Koh, S. Chang and C. -S. Kim, "Light Field Super-Resolution via Adaptive Feature Remixing," *IEEE Transactions on Image Processing*, vol. 30, pp. 4114-4128, 2021.

[30] I. Schiopu, A. Munteanu, "Deep-learning-based macro-pixel synthesis and lossless coding of light field images", *APSIPA Transactions on Signal and Information Processing*, vol. 8, pp. 1-12, 2019.

[31] P. Astola and I. Tabus, "Coding of Light Fields Using Disparity-Based Sparse Prediction," *IEEE Access*, vol. 7, pp. 176820-176837, 2019.

[32] X. Jiang, M. Le Pendu and C. Guillemot, "Light field compression using depth image based view synthesis," *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 19-24, 2017.

[33] X. Huang, P. An, F. Cao, D. Liu, and Q. Wu, "Light-field compression using a pair of steps and depth estimation," *Opt. Express*, vol. 27, pp. 3557-3573, 2019.

[34] T. Senoh, K. Yamamoto, N. Tetsutani and H. Yasuda, "Efficient Light Field Image Coding with Depth Estimation and View Synthesis", *European Signal Processing Conference (EUSIPCO)*, pp. 1840-1844, 2018.

[35] X. Huang, P. An, Y. Chen, D. Liu and L. Shen, "Low Bitrate Light Field Compression with Geometry and Content Consistency," *IEEE Transactions on Multimedia*, vol. 24, pp. 152-165, 2022.

[36] E. Miandji, S. Hajisharif, J. Unger, "A Unified Framework for Compression and Compressed Sensing of Light Fields and Light Field Videos", *ACM Transactions on Graphics*, vol. 38, no. 23, pp. 1-18, 2019.

[37] J. Ravishankar, M. Sharma, P. Gopalakrishnan, "A Flexible Coding Scheme Based on Block Krylov Subspace Approximation for Light Field Displays with Stacked Multiplicative Layers", *Sensors*, vol. 21, no. 13, p. 4574, 2021.

[38] M. U. Mukati, M. Stepanov, G. Valenzise, F. Dufaux and S. Forchhammer, "View Synthesis-based Distributed Light Field Compression," *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1-6, 2020.

[39] W. Ahmad, S. Vagharshakyan, M. Sjostrom, A. Gotchev, R. Bregovic and R. Olsson, "Shearlet Transform-Based Light Field Compression Under Low Bitrates," *IEEE Transactions on Image Processing*, vol. 29, pp. 4269-4280, 2020.

[40] G. De Oliveira Alves et al., "The JPEG Pleno Light Field Coding Standard 4D-Transform Mode: How to Design an Efficient 4D-Native Codec," *IEEE Access*, vol. 8, pp. 170807-170829, 2020.

[41] Y. Chen, P. An, X. Huang, C. Yang, D. Liu, and Q. Wu, "Light field compression using global multiplane representation and two-step prediction," *IEEE Signal Processing Letter*, vol. 27, pp. 1135-1139, 2020.

[42] N. K. Kalantari, T. C. Wang, and R. Ramamoorthi, "Learning based view synthesis for light field cameras," *ACM Transactions on Graphics*, vol. 35, no. 6, pp. 193:1-193:10, 2016.

[43] I. Choi, O. Gallo, A. Troccoli, M. H. Kim and J. Kautz, "Extreme View Synthesis," *IEEE International Conference on Computer Vision (ICCV)*, pp. 7780-7789, 2019.

[44] J. Jin, J. Hou, H. Yuan, S. Kwong, "Learning Light Field Angular Super-Resolution via a Geometry-Aware Network," *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pp. 11141-11148, 2020.

[45] J. Shi, X. Jiang and C. Guillemot, "Learning Fused Pixel and Feature-Based View Reconstructions for Light Fields", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2552-2561, 2020.

[46] W. F. H. Yeung, J. Hou, J. Chen, Y. Ying Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues," *Europeon Conference on Computer Vision (ECCV)*, pp. 137-152, 2018.

[47] G. Wu, Y. Liu, L. Fang, Q. Dai, T. Chai, "Light field reconstruction using convolutional network on EPI and extended applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1681-1694, 2019.

[48] G. Wu, Y. Liu, Q. Dai, and T. Chai, "Learning sheared EPI structure for light field reconstruction," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3261-3273, Jul. 2019.

[49] Y. Wang, F. Liu, K. Zhang, Z. Wang, Z. Sun and T. Tan, "High-fidelity View Synthesis for Light Field Imaging with Extended Pseudo 4DCNN," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 830-842, 2020.

[50] J. Jin, J. Hou, J. Chen, H. Zeng, S. Kwong and J. Yu, "Deep Coarse-to-fine Dense Light Field Reconstruction with Flexible Sampling and Geometry-aware Fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-17, 2020.

[51] D. Liu, Y. Huang, Q. Wu, R. Ma and P. An, "Multi-Angular Epipolar Geometry Based Light Field Angular Reconstruction Network," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1507-1522, 2020.

[52] C. Conti, L. D. Soares, and P. Nunes, "Dense light field coding: A survey," *IEEE Access*, vol. 8, pp. 49244-49284, 2020.

[53] C. Brites, J. Ascenso and F. Pereira, "Lenslet Light Field Image Coding: Classifying, Reviewing and Evaluating," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 339-354, Jan. 2021.

[54] G. J. Sullivan, J. R. Ohm, W. J. Han, T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, 2012.

[55] S. Zhang, Y. Lin and H. Sheng, "Residual Networks for Light Field Image Super-Resolution," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11038-11047, 2019.

[56] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016,

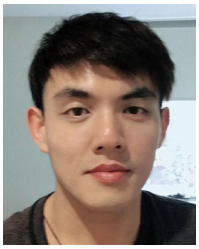
[57] A. Vedaldi, K. Lenc, "Matconvnet C convolutional neural networks for matlab," *ACM International Conference on Multimedia*, pp. 689-692, 2015.

[58] K. He, X. Zhang, S. Ren, J. Sun, "Delving deep into rectifiers: Surpassing humanlevel performance on imagenet classification," *IEEE International Conference on Computer Vision (ICCV)*, pp. 1026-1034, 2015.

[59] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves document," VCEG-M33 (2001).

[60] ISO /IEC JTC 1/SC 29 /WG 1, JPEG, "JPEG PLENO LIGHT FIELD CODING COMMON TEST CONDITIONS V3.3," Doc. N804025, Brussels, Belgium, July 2019.
[61] H. Sheng, P. Zhao, S. Zhang, J. Zhang, D. Yang, "Occlusion-aware depth estimation for light field using multi-orientation EPIs," *Pattern Recognition*, vol. 74, pp.587-599, 2018.
[62] M. Rerabek and T. Ebrahimi, "New light field image dataset," in https://mmspg.epfl.ch/EPFL-light-field-image-dataset, 2016.

Deyang Liu received the B.S. degree from Anqing Normal University, Anqing, China, in 2011, and the M.S. and Ph.D. degrees from Shanghai University, Shanghai, China, in 2014 and 2017, respectively. He is currently an Associate Professor with School of Computer and Information, Anqing Normal University. He is also a Postdoctoral Fellow with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China. From 2019 to 2020, he was a visiting scholar with University of Technology Sydney, Sydney, NSW, Australia. His research interests include 3-D video processing, light field image processing and video coding.



Yan Huang received B.S. and M.S. degrees from the School of Computer Science and Engineering in Sichuan University (SCU), China, in 2013 and Beihang University (BUAA), China, in 2016, respectively, and the Ph.D. degree from the University of Technology Sydney (UTS), Australia, in 2021. He is currently a Postdoctoral Fellow under the International Exchange Talent Introduction Program with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include deep learning, object re-identification and recognition.

Yuming Fang (Senior Member, IEEE) received the B.E. degree from Sichuan University, Chengdu, China, the M.S. degree from the Beijing University of Technology, Beijing, China, and the Ph.D. degree from Nanyang Technological University, Singapore. He is currently a Professor with the School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China. His research interests include visual attention modeling, visual quality assessment, computer vision, and 3D image/video processing. He serves on the Editorial Board of IEEE Transactions on Multimedia and Signal Processing: Image Communication.

Yifan Zuo received the Ph.D. degree from the University of Technology Sydney, Ultimo, NSW, Australia, in 2018. He is currently an associate professor with the School of Information Management, Jiangxi University of Finance and Economics. His research interests include computer vision and image processing.

Ping An received the B.S. and M.S. degrees from Hefei University of Technology, Hefei, China, in 1990 and 1993, respectively, and the Ph.D. degree from Shanghai University, Shanghai, China, in 2002. In 1993, she joined Shanghai University, where she is currently a Professor with Video Processing Group, School of Communication and Information Engineering. From 2011 to 2012, she joined the Communication Systems Group, Technische University of Berlin, Berlin, Germany, as a Visiting Professor. She has finished more than 10 projects supported by the National Natural Science Foundation of China, the National Science and Technology Ministry, and the Science and Technology Commission of Shanghai Municipality. Her research interests include image and video processing, with a focus on immersive video processing. She was the recipient of the Second Prize in Natural Sciences of the Ministry of Education in 2016, and the Second Prize in Natural Sciences of the Chinese Institute of Electronics in 2018.