LIGHT FIELD IMAGE COMPRESSION BASED ON DEEP LEARNING

Zhenghui Zhao[1], *Shanshe Wang*[2], *Chuanmin Jia*[2], *Xinfeng Zhang*[3], *Siwei Ma*[2], *Jiansheng Yang*[1,4]

[1]LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, China
[2]Institute of Digital Media, School of EE & CS, Peking University, Beijing 100871, China
[3]Viterbi School of Engineering, University of Southern California
[4]Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048

ABSTRACT

In this paper, we propose a novel light field image compression scheme by exploiting the intrinsic similarity of light field images with deep learning. In particular, instead of conveying all LF sub-views, only sparsely sampled LF sub-views are compressed and the remaining sub-views are reconstructed from the coded sub-views in the neighbourhood with convolutional neural network (CNN). To jointly suppress the artifacts induced in compression and reconstruct the un-coded views with high geometric accuracy, a multi-view joint enhancement network is introduced to improve the coding performance. Extensive experiments show the superior compression performance of our scheme compared with the state-of-the-art methods.

***Index Terms*—** Light field, image coding, view reconstruction, deep learning

1. INTRODUCTION

Recently, the Light Field (LF) images have attracted tremendous attentions due to its capability in representing abundant information of the 3D environment. In particular, LF images record the intensity of the light ray at each direction as well as each spatial position. As such, the rich information about the light rays beyond the traditional imaging methods provides amazing imaging functionalities, such as digital refocusing and viewpoint changing. Moreover, numerous regular image processing methods [1] can also benefit from the depth or other geometric information derived from LF images.

However, one major obstacle regarding the application of LF images is that the recorded information of the light rays in the LF images is represented in an inefficient way. Two major LF imaging devices are the camera array [2] and the lenslet-based cameras(e.g., Lytro [3]) where micro-lens are introduced to capture individual lights rays. These imaging devices store plenty of redundant information about the light rays in the same scene, leading to strong inter-view correlation. This is creating a grand challenge for the storage and transmission of the LF images. Hence, the high efficiency compression methods of LF images are highly desired.

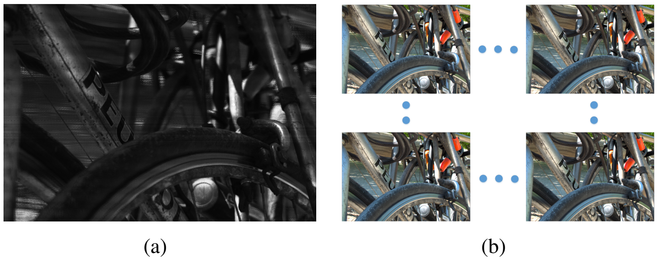


Fig. 1. Illustration of the original lenslet image and the decomposed sub-views.

With the standardization progress of JPEG Pleno [4], there is a tremendous interest in developing the compression algorithms of the LF images. As shown in Fig. 1, the captured LF images can be decomposed into multiple sub-views to better express the intrinsic structure of the LF images. Liu *et al.* [5] reordered the sub-views into a pseudo sequence which can be efficiently compressed by the advanced video compression methods such as HEVC [6]. To fully exploit the intrinsic geometry between the LF sub-views, Chen *et al.* [7] proposed a disparity guided sparse coding method. Zhao *et al.* [8] proposed a sparse sampling method with the linear approximation prior to utilize the similarity between the sub-views, which achieves significant compression performance improvement. However, the linear prior is not accurate enough to describe the correlation between the viewpoints, such that better characterization of the relationship between the sub-views is desired to further improve the compression performance.

Recently, deep learning has presented amazing advantages in dealing with the complex non-linear tasks such as image classification [9], image super-resolution [10], etc. Moreover, the deep neural network has also been recognized as a

This work was supported in part by National Natural Science Foundation of China (61632001, 61421062, 61571017, 61520106004), the National Basic Research Program of China (973 Program, 2015CB351800), National Program for Support of Top-Notch Young Professionals, which are gratefully acknowledged.

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

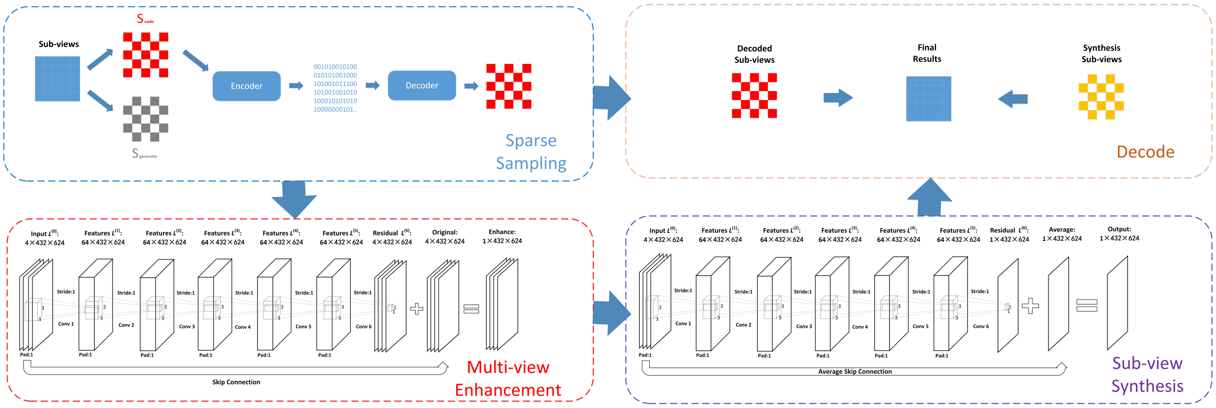


Fig. 2. The architecture of the proposed deep learning based light field images compression.

powerful tool for image and video compression. Yu *et al.* [11] proposed a deep learning based method in order to reduce various compression artifacts, which achieves significant visual quality improvement. In [12], the performance of the loop filter is further boosted using deep neural network. Regarding LF images, Kalantari *et al.* [13] presented a CNN-based approach for view synthesis from a sparse set of input views, which shows great potential for the application of deep learning in LF sub-views rendering. Wang *et al.* [14] presents a temporal interpolation method of the LF sub-views. However, in these methods the influence of the coding distortion [15, 16] introduced by video compression has not been taken into consideration, such that it is impractical to directly apply these approaches in the LF image compression.

In this paper, we propose a deep-learning based coding scheme to improve the performance of the LF images compression. By taking advantage of the similarity between neighbouring viewpoints, the sparsely sampled sub-views are rearranged into a pseudo sequence and encoded by the video codec. The other un-coded sub-views are reconstructed non-linearly via convolutional neural networks. Considering only the reconstructed sub-views are obtained at the decoder, a multi-views joint enhancement network is applied to suppress the compression artifacts. The proposed enhancement network and the view synthesis network are combined into an end-to-end learning-based LF image compression scheme. Experimental results show the superior compression performance compared with the state-of-the-art algorithms.

The rest of the paper is organized as follows. Section 2 describes the proposed scheme in details. Experimental results are presented in the Section 3. Finally, we conclude the paper in Section 4.

2. PROPOSED LF COMPRESSION FRAMEWORK

In this section, we detail the proposed light field image compression framework. As shown in Fig. 2, the proposed LF images compression algorithm is composed of three major modules: light field sparse sampling, deep learning based sub-view synthesis and multi-view joint enhancement, which are elaborated respectively in the following subsections.

2.1. Light Field Sparse Coding

In order to better express the intrinsic characteristic of the light field images, the original light field image with the resolution 7728×5328 are firstly decomposed into a 5-D LF structure with the MATLAB toolbox [17]. The decomposed LF images are composed of 15×15 sub-views with the resolution $432 \times 624 \times 3$. Each sub-view embodies the light intensity in one direction.

Due to the similarity of these sub-views, it may not be necessary to compress all of them equivalently. Therefore, we divide all of the sub-views into two categories, *i.e.,* S_{code} and $S_{generate}$, as described in [8], where the sub-views in S_{code} are compressed while the others are generated from the reconstructed sub-views in S_{code}. Firstly, the sub-views in S_{code} are rearrange into a pseudo sequence and a video codec is applied to compress them. Then, the corresponding decoding operation is applied to these compressed sub-views to obtain the decoded sub-views in S_{code}. Finally, each sub-view in $S_{generate}$ is generated from the neighbouring views, which are decoded views in S_{code} to reconstruct all the decomposed lenset images.

2.2. Deep Learning based Sub-view Synthesis

2.2.1. Network Architecture

In this paper, we design a six-layer convolutional neural network $S(x)$ to generate the skipped views in coding process from the neighboring reconstructed frames as shown in Fig. 2, where the i-th layer is denoted as $L^{(i)}$. The first layer $L^{(0)}$ accepts a multi-channel tensor where each channel corresponds to the luma component of a neighbouring sub-view. It is worth noting that the channel amount of the input is variant, which is determined by the number of surrounding accessible sub-views. All the accessible decoded sub-views are arranged into a queue according to the clockwise order from current viewpoint. The last layer outputs an non-linear approximation of the current un-coded view.

In our network, each convolutional layer consists of 64 convolutional kernels with the spatial resolution 3×3. The rectified linear unit (ReLU) [18] $f(x) = \max\{x, 0\}$ is adopted as the non-linear activator for these convolutional layers except for the last layer. Each layer extracts multiple features by convoluting these features in the following non-linearly strategy:

$$\begin{aligned} L^{(i+1)} &= \max\{W^{(i)} \otimes L^{(i)} + b^{(i)}, 0\}, \\ i &= 0, 1, \cdots, 4, \end{aligned} \tag{1}$$

where the $W^{(i)}$ and $b^{(i)}$ correspond to the convolutional kernel and bias in the i^{th} convolutional layer. The symbol $\otimes$ represents the convolution operation.

Some recent research [8] has utilized the linearity of light field images in angular domain to approximate the un-coded sub-views with the average of the surrounding viewpoints. However, the relationship among different sub-views is obvious nonlinear due to the illuminance changes, the disparity caused by the angular displacement and sub-view decomposition distortions. Therefore, we propose a convolutional neural network to characterize the non-linearity between these sub-views. In order to make full use of the non-linear transformation of the neural network, a skip connection adds the average of the surrounding sub-views in the input layer to the output layer. Moreover, the skip connection also can boost the convergence speed of the neural network training and achieve better performance in some other applications [9, 10]. The average of the neighbouring sub-views is added to the non-linear residuals derived from the proposed network to generate the un-coded sub-views in the set $S_{generate}$.

2.2.2. *Parameter Updating*

The convolutional kernels are initialized with the method described in [18] and the biases in the proposed network are set to zero at the beginning of the training. Since the network takes the surrounding coded views as the inputs to generate the un-coded sub-views, the parameter updating is formulated by minimizing the distance between the original dropped sub-views and the output of the neural network. Given a collection of N training instances, the mean squared error is used as the loss function to measure this distance. The training procedure can be formulated as:

$$\min_{W_{(i)}, b_{(i)}} \frac{1}{N} \sum_{j=1}^{N} \| S(I_j^{uncode}) - I_j^{orig} \|_2^2 \tag{2}$$

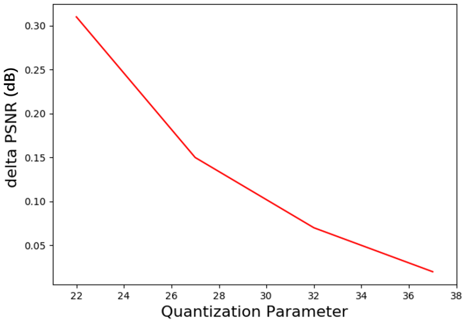


Fig. 3. The difference between the CNN based synthesis performance and the linear approximation with the quantization parameters.

where the I_j^{orig} is the corresponding dropped sub-views and the I_j^{uncode} is the decoded sub-views around I_j^{orig}.

The derivation of the trainable parameters in the neural network is calculated by the back-propagation. The learning rate is fixed to 0.001. We use the Adam algorithm [19] with the default hyper-parameter settings to update the convolution kernels and biases.

2.3. Multi-view joint enhancement

2.3.1. *Performance analysis for Compressed Videos*

In general, the trained network is sensitive to the change of the data characteristics. However, the compression artifacts introduced by video coding usually can change the statistical properties of the sub-views, and further degrade the performance of the neural network especially in low bit-rate scenario. In Fig. 3, we show the performance variation of the CNN based un-coded sub-view synthesis along with quantization parameters, where the X-axis represents the quantization parameters and the Y-axis represents the quality improvement of the CNN based sub-view synthesis compared with the average of the surrounding sub-views. The performance degeneration indicates that the coding distortions play an obvious negative influence on the trained convolution network.

2.3.2. *CNN based Multi-view Joint Enhancement*

To depress the influence of the compression artifacts, we design a quality enhancement convolution network to reduce the compression artifacts before the sub-view generation. Considering the variant distortion characteristic in different reconstructed sub-views, we can take multiple sub-views to enhance the sub-view quality by reducing the compression artifacts. To deal with the complicated non-linear correlation between these lossy sub-view images, a convolutional neural network $E(\cdot)$ is utilized to jointly suppress these artifacts.

The multi-view joint enhancement network $E(\cdot)$ has similar structure with the sub-view synthesis network $S(\cdot)$. The $E(\cdot)$ is composed of six convolution layers and each layer contains 64 convolution kernels with the shape 3×3. The number of the output channels is the same as that of the input layers. The tensor in the input layer is added to that in the output layer to improve the performance [10].

2.3.3. *Training of the Enhancement Network*

Since the training data plays an important role in the training process, we collect a large scale and diverse training samples by downsampling more than a hundred lenslet images from the JPEG Pleno [4]. The decomposed 5-D LF images are sparsely sampled as the S_{code} and these sub-views are rearranged into a pseudo sequence, which is further compressed using HEVC (reference software HM16.5). The un-coded sub-views and its surrounding sub-views in set S_{code} are used as the training labels and the corresponding inputs. These images are cropped into the size 32×32 to enrich the diversity of each training batch. All of the image values are normalized to the unit interval $[0, 1]$ for stable training process.

The loss function is another important factor to ensure the CNN converge to an optimal result. In this paper, we utilize the L2 norm distances between the enhanced sub-views and the dropped viewpoints as the loss function to generate high quality dropped sub-views, which can be formulated as:

$$distance(I_i, I^{orig}) = \frac{1}{N}\sum_{j=1}^{N} \|E(S(I_j^{code})) - I_j^{orig}\|_2^2 \quad (3)$$

where I_j^{orig} is the dropped views in the set $S_{generate}$ and I_j^{code} is the tensor where each channel denotes one of the neighbouring accessible coded sub-views. $S(\cdot)$ and the $E(\cdot)$ are the sub-view approximation network and multi-view joint enhancement network respectively. The sub-view synthesis network $S(\cdot)$ works as a part of the differentiable loss function.

When training the enhancement network $E(\cdot)$, the multiple convolution layer in the synthesis network $S(\cdot)$ are initialized with the training parameters which are described in the Section 2.2. The trainable parameters in the enhancement network $E(\cdot)$ are initialized using the random method [18] and are updated with the same gradient descent method Adam [19]. The parameters in the synthesis network $S(\cdot)$ do not change during the training process of the enhancement network. The synthesis network $S(\cdot)$ and the enhancement network $E(\cdot)$ are combined into an end-to-end network framework, which enhances the coded sub-views in the set S_{code} and make full use of the self-similarity between these sub-views to reconstruct the dropped sub-views in the set $S_{generate}$.

Table 1. Quality improvement of the proposed scheme compared with the algorithm with the linear prior [8] (dB).

Test Images	QP=22	QP=27	QP=32	QP=37
Vespa	0.44	0.34	0.26	0.23
Bikes	0.56	0.38	0.28	0.23
Color_Chart_1	0.79	0.61	0.52	0.59
Danger_de_Mort	0.49	0.35	0.25	0.22
Desktop	0.45	0.35	0.24	0.18
Flowers	0.46	0.33	0.24	0.16
Friends_1	0.15	0.12	0.09	0.08
ISO_Chart_12	0.91	0.85	0.73	0.62
Magnets_1	0.25	0.22	0.22	0.24
Stone_Pillars _Outside	0.30	0.19	0.11	0.08
Ankylosaurus_& _Diplodocus_1	0.27	0.20	0.15	0.17
Fountain_& _Vincent_1	1.02	0.73	0.47	0.28

3. VALIDATIONS

3.1. Testing Condition

The test images are from the JPEG Pleno Call for Proposal [20] and all test images are downloaded from the EPFL LF dataset. It is worth noting that there is no overlap between training data and testing data. The lenslet images are decomposed into multiple sub-views with the spatial shape $15 \times 15 \times 434 \times 542 \times 3$ using the MATLAB Light Field (LF) toolbox 0.4 [17]. Only the internal 11×11 sub-views are adopted because of the significant decomposed distortion of the other sub-views. The bit depth of these sub-views is 10 bits and the color space is RGB 422.

To demonstrate the performance of the proposed scheme, we compare the proposed scheme with two state-of-the-art methods. The first one (anchor-1) is rearranging all sub-views into a pseudo sequence following the serpentine order [5], such that the LF image can be compressed in terms of video sequences. Another compared algorithm (anchor-2) is the light field compression with the linear approximation [8], which achieves the state-of-the-art compression performance of the Light Field images. The pseudo video sequences are encoded by the reference software (HM-16.15) of the HEVC standard with the $encoder\_lowdelay\_main10$ (LDB) configuration.

Instead of compressing all sub-views, our proposed method reorders only 61 sub-views in the set S_{code} into a pseudo video sequence, such that only half of the sub-views are compressed. The pseudo sequences are encoded with the same testing condition as the above anchor algorithms. The sub-view synthesis network and the multi-view joint enhancement network are trained with the Caffe package and the networks are called with the trained convolution kernels and biases for our proposed scheme. The proposed neural network is only responsible for generating the luminance component and the chroma components are obtained by averaging the surrounding corresponding components.

We compare different algorithms with the average PSNR measure, which is calculated following JPEG Call for Proposal [20],

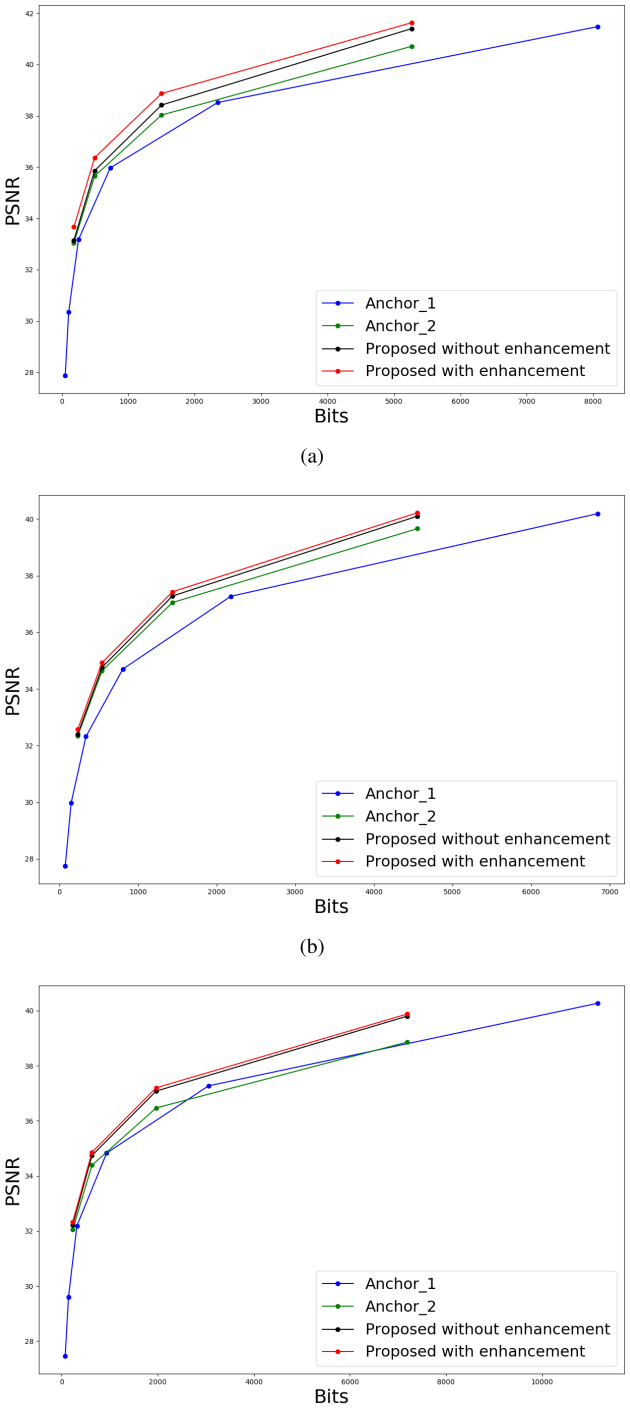
$$PSNR_{YUV}(I, I_{ref}) = \frac{6}{8} \times PSNR_Y(I, I_{ref}) \qquad (4)$$
$$+\frac{1}{8} \times PSNR_U(I, I_{ref}) + \frac{1}{8} \times PSNR_V(I, I_{ref})$$

3.2. Experimental Results

The rate-distortion performance of our proposed scheme is shown in Fig. 4. From Fig. 4, we can observe that the deep learning based sub-view synthesis scheme performs significantly better than the pseudo sequence compression algorithm. The multi-view joint enhancement network can well reconstruct the sub-views, such that the influence of the compression artifacts can be alleviated and the performance can be boosted beyond the sub-view synthesis network, which illustrates the efficiency of our proposed scheme. More detailed experimental results are elaborated in the Table 1. The comparisons are carried out between our proposed scheme and the state-of-the-art compression algorithm via the linear prior [8]. The results provide useful evidence on the high efficiency of our deep learning based scheme, and it is shown that our scheme can achieve up to 0.78 dB and 0.36 dB objective quality improvement upon the linear approximation methods at the same bit-rate. Benefiting from the powerful non-linear representation of the convolutional neural network, our proposed non-linear sub-view synthesis can well characterize the correlation between the sub-views than the linear approximation, such that significant performance improvement of the LF image compression can be achieved.

4. CONCLUSION

In this paper, we have proposed a CNN-based light field image compression scheme. Our proposed scheme takes advantage of the intrinsic high redundancy of LF images, and applies a non-linear deep-learning-based view synthesis network to boost the performance of the LF images compression. In particular, we investigate the influence of the coding distortion on the quality of the generated sub-views, and a multi-views joint enhancement network is combined with the sub-view synthesis network as an end-to-end system to generate the remaining sub-views. Experimental results demonstrate that the proposed compression scheme can obviously



(a)

(b)

(c)

Fig. 4. Rate-distortion performance comparisons for the testing sequence.(a) ISO_Chart_12; (b) Bikes; (c) Fountain_&_Vincent_1.

improve performance of the light field images compression over the state-of-the-art methods.

5. REFERENCES

[1] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu, "Saliency detection on light field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.

[2] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy, "High performance imaging using large camera arrays," in *ACM Transactions on Graphics (TOG)*. ACM, 2005, vol. 24, pp. 765–776.

[3] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report CSTR*, vol. 2, no. 11, pp. 1–11, 2005.

[4] Touradj Ebrahimi, Siegfried Foessel, Fernando Pereira, and Peter Schelkens, "Jpeg pleno: Toward an efficient representation of visual reality," *Ieee Multimedia*, vol. 23, no. 4, pp. 14–20, 2016.

[5] Dong Liu, Lizhi Wang, Li Li, Zhiwei Xiong, Feng Wu, and Wenjun Zeng, "Pseudo-sequence-based light field image compression," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–4.

[6] Gary J Sullivan, Jens Ohm, Woo-Jin Han, and Thomas 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.

[7] Jie Chen, Junhui Hou, and Lap-Pui Chau, "Light field compression with disparity-guided sparse coding based on structural key views," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 314–324, 2018.

[8] Shengyang Zhao and Zhibo Chen, "Light field image coding via linear approximation prior," in *Image Processing (ICIP), 2016 IEEE International Conference on*, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[11] Ke Yu, Chao Dong, Chen Change Loy, and Xiaoou Tang, "Deep convolution networks for compression artifacts reduction," *arXiv preprint arXiv:1608.02778*, 2016.

[12] Chuanmin Jia, Shiqi Wang, Xinfeng Zhang, Shanshe Wang, and Siwei Ma, "Spatial-temporal residue network based in-loop filter for video coding," *arXiv preprint arXiv:1709.08462*, 2017.

[13] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 193, 2016.

[14] Ting-Chun Wang, Jun-Yan Zhu, Nima Khademi Kalantari, Alexei A Efros, and Ravi Ramamoorthi, "Light field video capture using a learning-based hybrid imaging system," *arXiv preprint arXiv:1705.02997*, 2017.

[15] Xinfeng Zhang, Ruiqin Xiong, Xiaopeng Fan, Siwei Ma, and Wen Gao, "Compression artifact reduction by overlapped-block transform coefficient estimation with block similarity," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4613–4626, 2013.

[16] Xinfeng Zhang, Weisi Lin, Ruiqin Xiong, Xianming Liu, Siwei Ma, and Wen Gao, "Low-rank decomposition-based restoration of compressed images via adaptive noise estimation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4158–4171, 2016.

[17] Donald G Dansereau, Oscar Pizarro, and Stefan B Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1027–1034.

[18] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[19] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] "Jpeg pleno call for proposals submission details," `https://jpeg.org/downloads/jpegpleno/wg1n75024-REQ-JPEG_Pleno_CfP_-_Submission_Process_Details.pdf`.