Perceptual loss for light field reconstruction in high-dimensional convolutional neural networks

Nan Meng, Tianjiao Zeng, Edmund Y. Lam

Department of Electrical and Electronic Engineering,
The University of Hong Kong, Pokfulam, Hong Kong
elam@eee.hku.hk

Abstract: We explore the benefits of perceptual loss for light field (LF) spatial reconstruction in a high-dimensional convolutional neural network. The results outperform some state-of-the-art methods for LF or image super-resolution. © 2019 The Author(s)
OCIS codes: 100.0100, 100.4996, 100.6640.

1. Introduction

Light field (LF) imaging is a powerful imaging tool for capturing 3D environment, which has been widely utilized in various applications [1–3]. Despite appealing benefits and additional information light field has brought, it often suffers from great limitation in spatial resolution [1]. To mitigate such problem, various computational methods have been proposed and presented promising results, especially deep learning based approaches using convolutional neural networks (CNNs). However, in most existing literature [2, 3], the networks were learnt by minimizing the mean square error (MSE), which usually results in the lack of high-frequency spatial details for the super-resolved (SR) LFs. Therefore, in this paper, we proposed a novel 4D CNN network with perceptual loss acquired by enhancing MSE with spatial content loss for enhancing spatial details in light field.

2. Method

We establish the network as stacked convolution blocks as is illustrated in Fig. 1. We increase the resolution of the input LF with trained sub-pixel convolution layers inspired by [4]. The output SR LF is then send to a pre-trained VGG network [5] together with ground truth to calculate the content loss for updating the model. The proposed loss function ℓ is described as the weighted sum of a spatial content loss (ℓ_{S}) and a MSE loss (ℓ_{MSE}), i.e. $\ell = \ell_{\mathrm{S}} + 10^{-1} \cdot \ell_{\mathrm{MSE}}$. The content loss is defined based on the LeakyReLU activation layer of the pre-trained VGG network. Like most existing work, the MSE loss calculates the residual between ground truth I_{GT} and predicted HR output I_{SR}, which is formulated as $\ell_{MSE} = \frac{\|I_{\mathrm{GT}} - I_{\mathrm{SR}}\|_2^2}{U \cdot V}$, where $\|\cdot\|_2^2$ denotes the L_2 norm. U and V represent the number of horizontal and vertical viewpoints respectively. We use $\phi_{i,j}^k(\cdot)$ $(i \in \{1, \cdots, U\}, j \in \{1, \cdots, V\}, k \in \mathbb{N})$ to indicate the feature map obtained by the kth convolution when feeding the $(i, j)^{\mathrm{th}}$ sub-aperture image of a LF. Therefore, the spatial content loss can be described as, $\ell_S = \frac{1}{U \cdot V} \sum_{i=1}^{U} \sum_{j=1}^{V} \sum_{k=1}^{L} \left(\phi_{i,j}^k(I_{\mathrm{GT}}) - \phi_{i,j}^k(I_{\mathrm{SR}}) \right)$.

3. Results and Analysis

We compare the performance of our method with the state-of-the-art LF spatial reconstruction methods, including Yoon et al. [6], LFNet [2], and BM PCA+RR [3]. We evaluate the performance of each algorithm on a HCI synthetic scene – *Buddha* and a real-world scene – *Bikes 16*. As reported in Table 1, our method outperforms other methods in the the peak signal to noise ratio (PSNR) and structural similarity (SSIM) index. As can be seen in Fig. 2, better "fine texture" details have been reconstructed with our proposed perceptual loss, which encourages the network to reconstruct spatial details with high fidelity while preserving the angular correlations.

Table 1. Quantative comparison on synthetic and real-world light field.

Methods	PSNR(dB)						SSIM					
	Buddha			Bikes 16			Buddha			Bikes 16		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Yoon [6]	31.01	31.23	31.38	24.01	27.86	28.53	0.821	0.825	0.828	0.778	0.828	0.839
LFNet [2]	31.08	31.93	32.18	24.81	28.17	28.59	0.830	0.834	0.838	0.788	0.830	0.829
BM PCA+RR [3]	31.16	31.43	31.74	**24.95**	28.05	28.45	0.831	0.837	0.843	0.787	0.833	0.831
Ours ($\ell_{\mathrm{MSE}} + \ell_{\mathrm{S}}$)	**31.27**	**33.06**	**33.88**	24.54	**28.48**	**29.04**	**0.875**	**0.885**	**0.892**	**0.791**	**0.853**	**0.863**

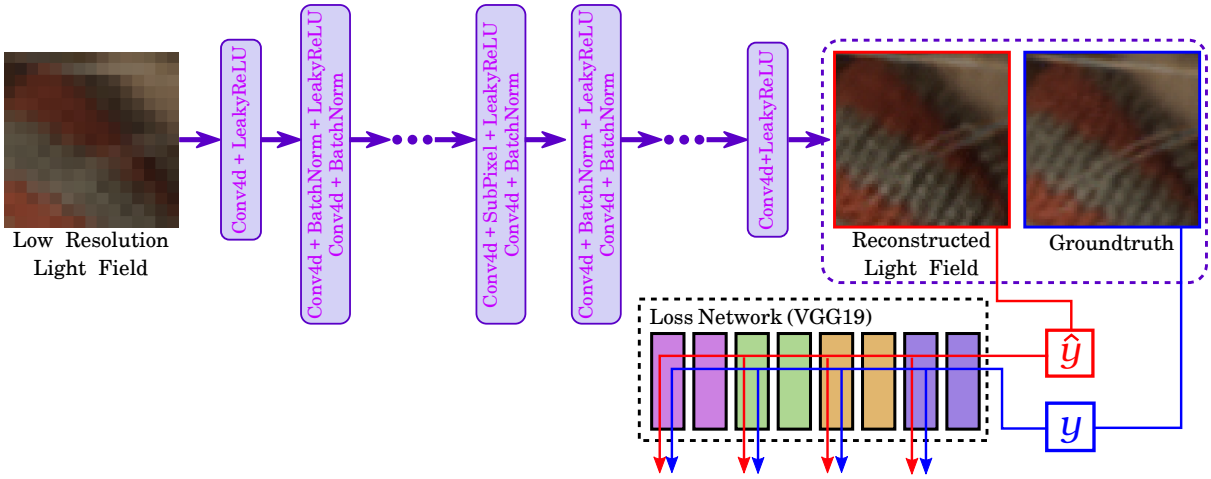

Fig. 1. Proposed network architecture for light field spatial enhancement.

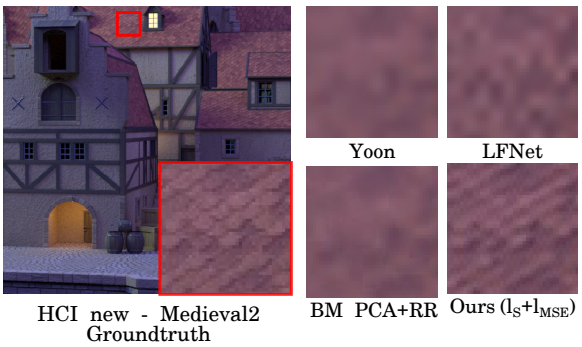

Fig. 2. Comparison of Visual results.

4. Conclusion

In this work, we demonstrate the superior performance of perceptual loss for reconstructing high-frequency spatial details of LFs compared with other state-of-the-art methods. Promising results have been presented in both visual inspection and objective assessments.

References

1. E. Y. Lam, "Computational photography with plenoptic camera and light field capture: tutorial," *JOSA A*, vol. 32, no. 11, pp. 2021–2032, 2015.
2. Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4274–4286, 2018.
3. R. A. Farrugia, C. Galea, and C. Guillemot, "Super resolution of light field images using linear subspace projection of patch-volumes," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1058–1071, August 2017.
4. Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," *Optica*, vol. 4, no. 11, pp. 1437–1443, 2017.
5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *n International Conference on Learning Representations*, 2015.
6. Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Light-field image super-resolution using convolutional neural network," *Signal Processing Letters*, vol. 24, no. 6, pp. 848–852, 2017.