DCM-CNN: Densely Connected Multiloss Convolutional Neural Networks for Light Field View Synthesis

LU LIU, ZICHENG NIAN, (Student Member, IEEE), AND CHEOLKON JUNG, (Member, IEEE)
School of Electronic Engineering, Xi’an University, Xi’an 710071, China
Corresponding author: Cheolkon Jung (zhengzk@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872280, and in part by the International S&T Cooperation Program of China under Grant 2014DFG12780.

ABSTRACT Different from traditional cameras, the light field cameras record spatial information with its angular information. Since the angular and spatial resolutions are limited due to the hardware shortage, view synthesis is required to provide arbitrary views. In this paper, we propose densely connected multiloss convolutional neural networks for light field view synthesis, called DCM-CNN. We build DCM-CNN for view synthesis based on the feature reuse, thereby leading to low computational complexity. Moreover, we present a multiloss function that consists of pixel loss, feature loss and edge loss to produce accurate edges and textures in the synthesized views. Experimental results show that the proposed method generates high quality views from light field data with low computational complexity and outperforms state-of-the-art ones in terms of structural similarity (SSIM), peak-signal-to-noise ratio (PSNR), and runtime.

INDEX TERMS Convolutional neural networks, densely connected, light field, multiloss, view synthesis.

I. INTRODUCTION Light field camera captures the 4D light field by capturing the intensity and direction of ray, thus recording both spatial and angular information in light field images. Thus, users can use light field data for computational photography such as image refocus [1], image matting [2] and depth reconstruction [3]–[5]. To generate virtual views from light field data, Levoy and Hanrahan [6] rendered a 2D array of images for a same scene. Ren et al. [7] developed a plenoptic camera by placing a micro-lens array in front of sensors, and sampled light field data from only one exposure.

A. BACKGROUND With the development of commercial light field cameras such as Lytro and Raytrix, light field photography has attracted more and more attention in recent years. Owing to the limitation of the hardware configuration of the cameras, it needs a balance between spatial and angular resolutions [8]. Thus, it is difficult for light field data to have high resolution in both spatial and angular domains. Similar to single image super-resolution, light field super-resolution is used to address this problem. According to the trade-off of the camera, light field super-resolution can be achieved in two ways: spatial domain and angular domain. The spatial super-resolution is similar to single image super-resolution and aims to improve the resolution of each sub-aperture image in light field data. On the other hand, the angular super-resolution aims to increase the sub-aperture numbers, i.e. virtual view synthesis. The virtual view synthesis increases the angular resolution of the light field image, thus increasing the spatial resolution of the light field camera. Chaurasia et al. [9] and Wanner and Goldluecke [10] provided a depth-based method to synthesize virtual views for a light field image by estimating the depth of light field image. Based on the estimated depth, they warped the input images to synthesize the virtual views. However, because of the depth estimation errors, their synthesized virtual views included blurring, tearing and ghosting. They did not consider that the light field image was often not captured under ideal conditions. Some previous studies [11]–[13] used epipolar plane images (EPI) to achieve light field image super-resolution. Since 4D light field images with angular and spatial dimensions were converted to 2D slices through EPI, the EPI
dimension was greatly reduced compared with the light field image [14]. However, the performance of these approaches was limited due to the low quality light field image captured by a commercial light field camera.

Recently, deep learning has shown its superiority in image processing tasks such as image fusion [15], [16], deblurring [17] and super-resolution [18], [19]. Moreover, deep learning has been used for virtual view synthesis [20]–[23]. Flynn et al. [24] used a sequence of baseline images to generate virtual views based on deep learning. Wang et al. [25] proposed a CNN architecture for EPI slices to be built for material recognition instead of the EPI restoration task. Zhou et al. [23] estimated appearance flow using CNN, and generated the virtual view by the warped input images. Kalantar et al. [20] introduced learning-based view synthesis for light field cameras. Their networks consisted of two CNNs: disparity estimation CNN and color estimation CNN, and produced more reasonable and accurate results than the previous work. Although they obtained more clear and accurate results than the previous work, it was difficult for them to train the network end-to-end and they synthesized blurry views. Yoon et al. [22] proposed light field image super-resolution CNNs. In this method, the first part was a CNN with three layers that the spatial super-resolution for each sub-image was similar to the single image super-resolution CNN, while the second one was an angular super-resolution to synthesize the virtual views. Because they learned the relationship between the virtual view and input images in the pixel space, the networks could not learn the complex relationship between the input images and virtual view, thus resulting in blurring artifacts. The loss in the pixel space such as MSE did not perform well when restoring high frequency information. In other words, it was limited to compute the perceptual differences by the pixel level loss because it calculated pixel level differences [26]–[29]. Minimizing mean squared error (MSE) found the average difference in the pixel space, thus producing typically over-smoothed results especially in image super-resolution [29]. To overcome this shortcoming, Johnson et al. [30] provided the perceptual loss instead of the MSE loss to reconstruct fine details in super-resolution. Ledig et al. [29] applied a generative adversarial network to the single image super-resolution and used the adversarial and content losses to train the networks. However, sub-aperture images in light field data are very similar because the disparity is small for the same scene, and thus the PSNR between different sub-images is very high. The better PSNR value does not mean high performance, which means that only pixel-wise loss is not able to measure the difference between virtual view and ground truth images.

**B. CONTRIBUTIONS**

In this paper, we propose DCM-CNN for light field view synthesis. We adopt DenseNet [31] with a narrow width instead of the network structures with hundreds of thousands of filters. Due to the feature reuse, DenseNet does not require any additional parameters for each layer, and thus it greatly reduces the amount of parameters. We build DCM-CNN based on DenseNet to synthesize light field views. Moreover, we use a multiloss function for network optimization that consists of pixel loss (MSE), feature loss and edge loss instead of the traditional pixel-wise loss. The multiloss function and some preliminary results of this paper were presented in [32]. In this extended paper, we combine densely connected CNN with the multiloss function to produce high quality views with low computational complexity. Experimental results show that the proposed method successfully generates light field views with low computational complexity and outperforms state-of-the-art ones in terms of quantitative measurements and runtime. Fig. 1 illustrates the whole framework of the proposed method.

Compared with existing work, the main contributions of this paper are as follows:

- We use a feature loss for light field view synthesis instead of the pixel loss. The feature loss is more effective than the pixel loss in measuring the similarity between the ground truth and virtual view.
- We utilize edge loss to generate clear structures in virtual views. Edge loss is very effective in measuring the edge similarity.
- We combine the multiloss function with DenseNet, i.e. DCM-CNN, to produce high quality light field views with low computational complexity.

The remainder of this paper is as follows. Section II briefly reviews the related work, while Section III addresses the proposed method in detail. Experimental results and their corresponding analysis are provided in Section IV. Finally, the conclusion of this paper is drawn in Section V.

**II. RELATED WORK**

**A. IMAGE SUPER-RESOLUTION**

Image super-resolution (SR) is a classic computer vision problem that aims at reconstructing a high resolution image from a given low resolution one. Early interpolation based approaches, such as bicubic, bilinear, and Lanczos [33], are fast but achieve a poor performance. There are also some prior knowledge based approaches such as statistical methods [34] and edges based methods [35]. A category of state-of-the-art SR methods learns a mapping between low and high resolution patches, including example learning [36]–[38], neighbor embedding [39]–[42], and sparse dictionary [43], [44]. These studies aim at learning a compact dictionary to match low and high resolution patches. Deep learning has achieved an excellent performance in single image SR in recent years. Dong et al. [18] first proposed SR convolutional neural network (SRCNN) with three layers to model a mapping function, which also further showed that a traditional sparse-coding (SC) approach for SR had a strong correlation with SRCNN. Kim et al. [45] proposed an image SR method using a deeply recursive convolutional network (DRCN). They increased the recursion depth of the network to improve performance without adding new parameters for
additional convolutions layers. With recursive supervision and skip connection, they reduced the difficulty in training. Then, they proposed residual networks for training very deep convolutional networks (VDSR) and achieved good SR performance [19]. In [19], they found that increasing the network depth led to a significant improvement in accuracy so that finally used a model of 20 layers and residual learning for training. Lim et al. [46] developed an enhanced deep super-resolution network (EDSR) which achieved significant performance by removing redundant layers in conventional ResNet. They also developed a multiscale deep SR network (MDSR) which simplified the model and reduced training time showing outstanding performance than other SR ones. They have won the NTIRE2017 SR Challenge.

In addition to them, there are also some researches on the loss functions. Most deep learning methods used MSE as the loss function for SR, while MSE and PSNR were widely used as a performance measure for SR. However, MSE was a pixel-wise loss function that did not capture the perceptual difference between the network output and the ground truth. Zhao et al. [47] showed the importance of perceptually motivated losses when the resulting image was to be evaluated by human observers. They compared the performance of different losses and showed that the results improved significantly even when the network architecture was left unchanged. Johnson et al. [30] provided a perceptual loss to reconstruct fine details for image SR. Ledig et al. [29] proposed a method of using a generative adversarial network based on content and adversarial losses. Their performance was closer to the high resolution ground truth images than other state-of-the-art methods in the perceptual similarity. However, we cannot directly apply these single image SR methods to the light field SR whose target includes not only improving the resolution for each sub-aperture of light field images but also increasing spatial resolution.

B. LIGHT FIELD IMAGE SUPER-RESOLUTION

Due to the hardware shortage, the angular and spatial resolutions of light field images are limited when captured by light field cameras. Thus, many methods have been proposed to enhance the resolution of light field images at the software level. To achieve light field image super-resolution in both spatial and angular domains, some researchers have used epipolar plane images (EPIs) [10], [11], [48], [49] that are 2-D slices of constant angular and spatial directions. Mitra and Veeraraghavan [11] proposed a patch based approach based on a Gaussian mixture model (GMM). Wu et al. [49] presented a CNN-based framework (blur-restoration-deblur) from a sparse set of views to reconstruct a high quality light field image to resolve ghost effects on EPI due to an asymmetric information between the spatial and angular dimensions. They also trained CNN for light field image super-resolution based on the assumption that a patch in a shared EPI exhibits a clear structure when the shared values are equal to the depth [48]. Some light field image super-resolution methods have obvious defects, and are required to estimate an explicit disparity for one or more light field views. Rossi and Frossard [50] used a super-resolution method based on multi-frame similarity to improve the spatial resolution of the whole light field image. They used different light field
views to augment the spatial resolution of the whole light field image. Farrugia et al. [51] enhanced the spatial resolution of different views in a consistent manner over all sub-aperture images of the light field data. This method could be extended to the improvement of the angular resolution which used multivariate ridge regression to approximate the middle sub-aperture image of 2-D patches to the adjacent sub-aperture one. Their another work [52] proposed a learning based spatial light field image super-resolution that restored the whole light field image based on the consistency of sub-aperture images.

C. LIGHT FIELD VIEW SYNTHESIS

There exists a balance between angular and spatial resolutions due to the limit of the light field camera sensor. Many SR methods have been proposed to improve spatial and angular resolutions. In this work, we pay attention to the angular resolution and investigate synthesizing high quality virtual views. The view synthesis approaches are divided into two main categories: with depth estimation and without depth estimation.

1) WITH DEPTH ESTIMATION

Wanner and Goldluecke [10] developed a variational method for light field disparity estimation and spatial/angular SR. In their study, disparity was locally estimated using EPI. Then, they used convex optimization to consolidate the local disparity into global disparity maps. They utilized the estimated disparity and then warped the input images to synthesize virtual views. Jeon et al. [3] proposed depth estimation for narrow baseline sub-aperture images. They used a phase shift theorem to estimate the sub-pixel shifts in the Fourier domain, and estimated stereo correspondence accurately. Zhang et al. [53] proposed disparity assisted phase based synthesis (DAPS) to estimate disparity maps. They provided an analysis by synthesis based on DAPS to generate binocular images from the other warped one, and minimized phase differences between the warped image and the ground truth by iterative optimization of disparity. Penner and Zhang [54] proposed soft 3D reconstruction for view synthesis based on depth of input views. They reconstructed scene geometry of view using discretized projective volume representation. Kalantari et al. [20] introduced a deep learning method to synthesize virtual views. Their networks consisted of two CNNs: disparity estimation CNN and color estimation CNN. They extracted features such as mean and standard deviation from the input views and used the disparity estimation network for disparity map estimation at the virtual view. Finally, they used the warped images and the estimated disparity map as inputs to generate the virtual view. However, this method caused ghosting artifacts in the occluded regions and achieved poor performance where the warped images did not contain enough valid information. In general, the view synthesis methods with depth estimation used the estimated depth to synthesize virtual views. It was not easy to get accurate depth when the light field data could not contain enough valid information. Thus, these approaches often result in poor performance in occluded regions, non-Lambertian surfaces and transparent regions.

2) WITHOUT DEPTH ESTIMATION

Vagharshakyan et al. [12] proposed light field SR from a sparse set of perspective views without depth estimation. The desired intermediate views were synthesized by processing the corresponding EPI in the shearlet domain. However, they needed a large number of input views due to the assumption that the densely sampled EPI was a square image. Cao et al. [55] designed learning-based interpolation to reconstruct high quality light field images. They used sparse coding to train a dictionary for encoding natural image struc-
tures and use it to reconstruct light field images. Yoon et al.’s work [22] also does not perform depth estimation, and uses the MSE loss to train their network, thus thus resulting in blurring, ghost and inaccurate edges.

III. PROPOSED METHOD

A. DENSELY CONNECTED MULTILOSS CNN

We formulate the virtual view synthesis as follows:

\[ V_{v} = f(V_{1}, V_{2}, \ldots, V_{n}) \]  

(1)

where \( V_{1}, V_{2}, \ldots, V_{n} \) means a set of input sub-aperture views and \( V_{v} \) means the synthesized virtual view, and \( f(\cdot) \) is our target function to synthesize the virtual view from the input views. Eq. (1) shows the relationship between the sub-aperture views and the virtual view. Thus, we aim at modeling this relationship by formulating \( f(\cdot) \). However, different views usually have occlusion and relative shift, thus making the relationship very complex. We propose a densely connected multiloss CNN (DCM-CNN) to build the relationship as illustrated in Fig. 1. Inspired by DenseNet [31], we adopt CNN with densely connected layers to model \( f(\cdot) \). The input of the proposed network is the sub-aperture views, while the output is the virtual view. The input of each convolution layer is a concatenation of the output of the preceding several convolutions, while the output features are the input of each convolution layers. This structure is able to reuse the feature while reducing the parameters. Unlike the original DenseNet, we remove the pooling and BN layers for light field view synthesis. To preserve the output size the same as the input, we apply zero padding to each layer. Every layer has 16 filters with size of 3 × 3, while the output is 3 channels (RGB). Yoon et al. [14] used MSE between the output view and the corresponding ground truth view as loss function to train their networks. However, the MSE loss often causes blurring, incorrect warping and ghosting artifacts as shown in Fig. 3. That is because the MSE loss in the pixel space is not effective in restoring high frequency information. In other words, the MSE loss fails to compute the perceptual difference.

To address the disadvantage of the MSE loss, we propose a multiloss function that consists of pixel loss (MSE), feature loss and edge loss. The multi-loss function \( l_{mul} \) is formulated as follows:

\[ l_{mul} = \alpha \times l_{mse} + \beta \times l_{fea} + \gamma \times l_{edg} \]  

(2)

where \( l_{mse} \) is the loss function in the pixel space as follows:

\[ l_{mse} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (I_{x,y} - f(I_{x,y}^{outputs}))^{2} \]  

(3)

where \( l_{mse} \) is the most widely used loss function in deep learning, \( f(I_{x,y}^{outputs}) \) is the synthesized view, and \( I_{x,y} \) is the corresponding ground truth view. Minimizing MSE finds the average difference in the pixel space, thus producing typically over-smoothed results and leading to poor performance in image SR. Thus, we combine MSE loss, feature loss and edge loss in the multiloss function instead of the MSE loss. \( l_{fea} \) is the feature loss defined as follows:

\[ l_{fea} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (\phi(I_{x,y}^{view}) - \phi(f(I_{x,y}^{outputs})))^{2} \]  

(4)

where \( l_{fea} \) is mean squared error in the feature space between the synthesized view and corresponding ground truth view, \( \phi(f(I_{x,y}^{outputs})) \) is the feature image of the virtual view generated by the proposed networks while \( \phi(I_{x,y}^{view}) \) is the feature image of the ground truth view. We acquire the feature image through the pre-trained VGG-16 networks [56]. \( l_{edg} \) is an edge loss function defined as follows:

\[ l_{edg} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} (\varphi(I_{x,y}^{view}) - \varphi(f(I_{x,y}^{outputs})))^{2} \]  

(5)

where \( l_{edg} \) is the edge difference between the ground truth view and corresponding synthesized view, \( \varphi(f(I_{x,y}^{outputs})) \) is the edge map of the synthesized views, and \( \varphi(I_{x,y}^{view}) \) means the edge map of the ground truth.

B. NETWORK TRAINING

To train the proposed multi-loss networks, we use the public light field dataset provided by Yoon et al. [14]. We use a PC with E5-2640 CPU, GTX1080Ti GPU, and RAM 32GB. For training, a set of sub-aperture images as the input views and the ground truth images label views are required. We decode the light field images to \( N \times N \times W \times H \times C \) sub-image array, where \( N \) is the angular resolution of the light field data and \( W, H, C \) are the size of each sub-image. Thus, we regard the \( i_{th} \) sub-image as the ground truth between the horizontal \( (i - 1)_{th} \) and \( (i + 1)_{th} \) views. Analogously, we synthesize a virtual view between the vertical views and the central virtual view from four views. In the original data set, the image angular resolution is \( 14 \times 14 \), however, there are blind areas in some angular views. Thus, we filter out the data set with the angular resolution of \( 8 \times 8 \) to ensure that each pixel has a value. Finally, we select \( 2 \times 2 \) images of four corners in the \( 8 \times 8 \) resolution data set so that the input becomes four views. Our task is to use four different angular views to generate the central view, thus there is no angular coordinate in the central view. The light field image has angular resolution of \( 8 \times 8 \), and we only use four sub-images as the input. For training, we randomly crop the patches with the size of \( 120 \times 120 \) from the input sub-images to generate a large amount of the input patches. Meanwhile, we also crop the same size patches from the same region in the ground truth sub-images as the ground truth. During the training process, we use 64 batch size, randomly initialize the weights and ADAM solver with a learning rate of 0.01 dropping every 1000 iteration.

IV. EXPERIMENTAL RESULTS

We conduct experiments using the light field dataset provided by Raj et al. [57]. It contains 30 LF images of \( 8 \times 8 \times 541 \times 376 \times 3 \) size (angular resolution: \( 8 \times 8 \), spatial resolution: \( 541 \times 376 \times 3 \)). Since we consider Lytro Illum V2 camera with
FIGURE 3. Edge map comparison. (a) Ground truth. (b) Edge map of the ground truth. (c) Edge map of the synthesized virtual view by the Yoon et al.’s method [14]. (d) Edge map difference between the ground truth and the synthesized view. There are obvious differences between two edge maps.

FIGURE 4. Edge map comparison. (a) Ground truth. (b) Edge map of the ground truth. (c) Edge map of the synthesized virtual view by the Kalantari et al.’s method [20]. (d) Edge map difference between the ground truth and the synthesized view. There are obvious differences between two edge maps.

A small baseline, we use four sub-aperture images as the input to synthesize the central view. Synthesizing the central view can be easily and visually compared with the ground truth, and shows the effectiveness of the light filed view synthesis. Since most pixels of different views are very similar, it is difficult to find their differences. Thus, we perform edge map comparison on the synthesized views by the proposed method (multiloss) and Yoon et al.’s method [14] (single MSE loss). We use PSNR and structural similarity (SSIM) as evaluation metrics for performance comparison. To see the computational efficiency, we compare the runtime between the proposed method and Kalantari et al.’s [20]. All experiments are performed in MATLAB2014a.

A. ABLATION STUDY
As shown in Fig. 7, we perform an ablation study to prove the effectiveness of dense connection and multiloss function. We provide PSNR and SSIM scores on different test images and the proposed method achieves the highest score. Apparently, the results in Fig. 7 indicate that dense connection and multiloss function can effectively improve the quality of synthesized view. We also perform a subjective comparison on the results. Regardless of scenes, the proposed method achieves good similarity to the ground truth. Notice that there are still some lost details and artifacts in the synthesized view such as the edge of the cloth (see the first row of Fig. 7) when using only one or a few loss functions. Although it seems that the proposed method with densely connected multiloss CNN (DCM-CNN) performs only a little worse than the proposed method without DCM-CNN, DCM-CNN greatly reduces the size of the model and ensures the stability of model training. The proposed method still maintains better edges when the objects have relatively simple structures and details (see the bottom row of Fig. 7). In conclusion, using multiloss function keeps many specific features of the synthesized view so as to guarantee that the synthesized view is closer to the ground truth.
truth with better performance. Simultaneously, using dense connection is able to let the model be lightweight.

**B. EDGE DETECTION**

We compare the edge maps of the synthesized view and its ground truth as shown in Figs. 2, 3 and 4. The edge map of the proposed method is almost the same as that of the ground truth because their difference map is nearly zero (see Fig. 2(d)). The results indicate that the proposed method produces accurate edge information for the synthesized views. However, the edge maps of Figs. 3(d) and 4(d) by the Yoon *et al.*‘s [14] and the Kalantari *et al.*’s [20] are much different from the ground truth, i.e. not accurate.

The Yoon *et al.*’s method [14] uses an MSE loss for network optimization. Minimizing MSE loss computes the average difference in the pixel space and produces over-smoothed results without considering high frequency components. For view synthesis, the MSE loss has a limit of computing the perceptual difference causing inaccurate edges with ghost and blur artifacts. Although the Kalantari *et al.*’s method [20] produces more accurate results with higher PSNR and SSIM scores than the Yoon *et al.*’s [14], it synthesizes blurry views. Our multiloss function consists of the pixel loss, edge loss and feature loss, and thus it effectively considers the high frequency components with accurate edges.
FIGURE 6. Performance comparison between different methods. Left to right: Ground truth, zoomed results by Wanner et al. [10], Yoon et al. [14], Tao et al. [5], Jeon et al. [3], Kalantari et al. [20], the proposed method, and the ground truth.
C. DETAIL PRESERVATION

As shown in Fig. 2, the details of the proposed method are nearly the same as the ground truth. Although the synthesized views contain a little blur in some textures, the proposed method generally achieves outstanding view synthesis performance. We further compare the results of the proposed method with those of state-of-the-arts ones, including Jeon et al.’s method [3], Wang et al.’s method [4], Kalantari et al.’s method [20] and Yoon et al.’s method [14].

We evaluate PSNR and SSIM for performance comparison. PSNR is based on the mean square error (MSE) between the synthesized view and the ground truth, which indicates the image quality and similarity to the original image. The higher PSNR represents higher image quality. SSIM ranges [0,1] and the higher the better. We provide PSNR and SSIM scores on four test images in Table 1. It can be observed that the proposed method obviously performs better than the others. This is because our synthesized views are closest to the

| Method          | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
|-----------------|------|------|------|------|------|------|------|------|------|------|
| Wanner et al. [10] | 28.79 | 0.898 | 36.11 | 0.937 | 36.57 | 0.951 | 36.34 | 0.969 | 38.45 | 0.983 |
| Yoon et al. [14] | 31.63 | 0.869 | 40.34 | 0.968 | 38.12 | 0.950 | 38.21 | 0.952 | 44.64 | 0.985 |
| Tao et al. [5] | 26.51 | 0.748 | 35.48 | 0.942 | 35.21 | 0.951 | 35.04 | 0.949 | 39.85 | 0.977 |
| Jeon et al. [3] | 22.86 | 0.669 | 27.36 | 0.906 | 27.73 | 0.930 | 27.53 | 0.928 | 31.81 | 0.951 |
| Kalantari et al. [20] | 19.74 | 0.792 | 24.31 | 0.871 | 27.27 | 0.946 | 27.51 | 0.949 | 31.65 | 0.966 |
| Proposed method | 45.0024 | 0.9864 | 40.0226 | 0.9750 | 40.6159 | 0.9729 | 37.6470 | 0.9756 | 40.5512 | 0.9778 | 41.1904 | 0.9771 | 44.0023 | 0.9819 |

FIGURE 7. Performance comparison between different loss function. Left to right: Ground truth, zoomed results by ground truth, the proposed method, only MSE loss, only feature loss, only edge loss, MSE+feature loss, MSE+Edge loss, feature + Edge loss, without Densely Connected Multiclass CNN. There are obvious differences between these results.
the runtime comparison verifies that the proposed method is very cost-effective by using a simple network structure.

**RECOMMENDATIONS**

[1] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, “Dappled photography: Mask enhanced cameras for heteroeyd light fields and coded aperture refocusing,” *ACM Trans. Graph.*, vol. 26, no. 3, p. 69, Jul. 2007.

[2] D. Cho, S. Kim, and Y. W. Tai, “Consistent matting for light field images,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 90–104.

[3] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon, “Accurate depth map estimation from a lenslet light field camera,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1547–1555.

[4] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, “Occlusion-aware depth estimation using light-field cameras,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3487–3495.

[5] M. W. Tao, S. Hadap, J. Malick, and R. Ramamoorthi, “Depth from combining defocus and correspondence using light-field cameras,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 673–680.

[6] M. Levoy and P. Hanrahan, “Light field rendering,” in *Proc. 23rd Annul. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, New York, NY, USA, 1996, pp. 31–42. doi: 10.1145/237170.237199.

[7] N. Ren, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photography with a hand-held plenoptic camera,” Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep. CSTR 2005-02, 2005.

[8] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala, “Spatio-angular resolution tradeoffs in integral photography,” in *Proc. Eurographics Symp. Rendering Techn.*, Nicosia, Cyprus, 2006, pp. 263–272.

[9] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, “Depth synthesis and local warps for plausible image-based navigation,” *ACM Trans. Graph.*, vol. 32, no. 3, p. 30, Jun. 2013.

[10] S. Wanner and B. Goldluecke, “Variational light field analysis for disparity estimation and super-resolution,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[11] K. Mitra and A. Veeraraghavan, “Light field denoising, light field super-resolution and stereo camera based refoocussing using a GMM light field patch prior,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 22–28.

[12] S. Vagharshakyan, R. Bregovic, and A. Gotchev, “Image based rendering technique via sparse representation in shearlet domain,” in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 1379–1383.

[13] T. E. Bishop and P. Favaro, “The light field camera: Extended depth of field, aliasing, and superresolution,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.

[14] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and J. S. Kweon, “Light-field image super-resolution using convolutional neural network,” *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 848–852, Jun. 2017.

[15] Y. Yang, M. Yang, S. Huang, Y. Que, M. Ding, and J. Sun, “Multifocus image fusion based on extreme learning machine and human visual system,” *IEEE Access*, vol. 5, pp. 6989–7000, 2017.

[16] Y. Yang, M. Yang, S. Huang, M. Ding, and J. Sun, “Robust sparse representation combined with adaptive PCNN for multifocus image fusion,” *IEEE Access*, vol. 6, pp. 20138–20151, 2018.

[17] J. Sun, W. Cao, Z. Xu, and J. Ponce, “Learning a convolutional neural network for non-uniform motion blur removal,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 769–777.

[18] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[19] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[20] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, “Learning-based view synthesis for light field camera fields,” *ACM Trans. Graph.*, vol. 35, no. 6, pp. 193, Nov. 2016.

[21] A. Dosovitskiy, J. T. Springenberg, and T. Brox, “Learning to generate chairs with convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1538–1546.

[22] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, “Learning a deep convolutional network for light-field image super-resolution,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 57–65.
[23] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, “View synthesis by appearance flow,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 286–301.

[24] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, “Deep stereo: Learning to predict new views from the World’s imagery,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 5515–5524.

[25] T. C. Wang, Y. J. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, “A 4D light-field dataset and CNN architectures for material recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 121–138.

[26] P. Gupta, P. Srivastava, S. Bhardwaj, and V. Bhateja, “A modified PSNR metric based on HVS for quality assessment of color images,” in Proc. Int. Conf. Comput. Ind. Appl., Dec. 2011, pp. 1–4.

[27] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in Proc. 37th Asilomar Conf. Signals, Syst. Comput., vol. 2, 2003, pp. 1398–1402.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[29] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 4681–4690.

[30] J. Johnson, A. Alahi, and F. F. Li, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 694–711.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 4700–4708.

[32] Z. Nian and C. Jung, “High-quality virtual view synthesis for light field cameras using multi-loss convolutional neural networks,” in Proc. 25th IEEE Int. Conf. Image Process. (ICIP), Oct. 2018, pp. 2605–2609.

[33] M. Irani and S. Peleg, “Improving resolution by image registration,” CVGIP, Graph. Models Image Process., vol. 53, no. 3, pp. 231–239, May 1991.

[34] K. In Kim and Y. Kwon, “Single-image super-resolution using sparse regression and natural image prior,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[35] J. Sun, Z. Xu, and H.-Y. Shum, “Image super-resolution using gradient profile prior,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2008, pp. 1–8.

[36] W. T. Freeman, T. R. Jones, and E. C. Pasztor, “Example-based super-resolution,” IEEE Comput. Graph. Appl., vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.

[37] W. T. Freeman, T. R. Jones, and E. C. Pasztor, Example-Based Super-Resolution. Washington, DC, USA: IEEE Computer Society Press, 2002.

[38] Z. Xiong, D. Xu, X. Sun, and F. Wu, “Example-based super-resolution with soft information and decision,” IEEE Trans. Multimedia, vol. 15, no. 6, pp. 1458–1465, Oct. 2013.

[39] H. Chang, D.-Y. Yeung, and Y. Xiong, “Super-resolution through neighbor embedding,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 1, Jul. 2004, p. 1.

[40] M. Bevilacqua, A. Roumy, C. Guillermot, and M.-L.-A. Morel, “Low-complexity single-image super-resolution based on nonnegative neighbor embedding,” in Proc. BMVC, 2012, p. 135.

[41] X. Gao, K. Zhang, D. Tao, and X. Li, “Image super-resolution with sparse neighbor embedding,” IEEE Trans. Image Process., vol. 21, no. 7, pp. 3194–3205, Jul. 2012.

[42] J. Jiang, X. Ma, C. Chen, T. Lu, Z. Wang, and J. Ma, “Single image super-resolution via locally regularized anchored neighborhood regression and nonlocal means,” IEEE Trans. Multimedia, vol. 19, no. 1, pp. 15–26, Jan. 2017.

[43] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Trans. Image Process., vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[44] Z. Zhu, F. Guo, H. Yu, and C. Chen, “Fast single image super-resolution via self-example learning and sparse representation,” IEEE Trans. Multimedia, vol. 16, no. 8, pp. 2178–2190, Dec. 2014.

[45] J. Kim, J. K. Lee, and K. M. Lee, “Deeply-recursive convolutional network for image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 1637–1645.

[46] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jul. 2017, pp. 1132–1140.

[47] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss functions for image restoration with neural networks,” IEEE Trans. Comput. Imag., vol. 3, no. 1, pp. 47–57, Mar. 2017.

[48] G. Wu, Y. Liu, Q. Dai, and T. Chai, “Learning sheared EPI structure for light field reconstruction,” IEEE Trans. Image Process., vol. 28, no. 7, pp. 3261–3273, Jul. 2019.

[49] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, “Light field reconstruction using convolutional network on EPI and extended applications,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1681–1694, Jul. 2019.

[50] M. Rossi and P. Frossard, “Graph-based light field super-resolution,” in Proc. IEEE 19th Int. Workshop Multimedia Signal Process., Oct. 2017, pp. 1–6.

[51] R. A. Farrugia, C. Galea, and C. Guillermot, “Super resolution of light field images using linear subspace projection of patch-volumes,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 7, pp. 1058–1071, Oct. 2017.

[52] R. Farrugia and C. Guillermot, “Light field super-resolution using a low-rank prior and deep convolutional neural networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 5, pp. 1162–1175, 2020.

[53] Z. Zhang, Y. Liu, and Q. Dai, “Light field from micro-baseline image pair,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jan. 2015, pp. 3800–3809.

[54] E. Penner and L. Zhang, “Soft 3D reconstruction for view synthesis,” ACM Trans. Graph., vol. 36, no. 6, p. 235, Nov. 2017.

[55] D. Cho, M. Lee, S. Kim, and Y.-W. Tai, “Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 3260–3267.

[56] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent., 2015, pp. 1–14.

[57] A. S. Raj, M. Lowney, and R. Shah. (2016). Stanford Lytro Light Field Archive. [Online]. Available: http://lightfields.stanford.edu/