



A Coarse-to-Fine Convolutional Neural Network for Light Field Angular Super-Resolution

Gaosheng Liu[✉], Huanjing Yue[✉], and Jingyu Yang^(✉)[✉]

School of Electrical and Information Engineering, Tianjin University, Tianjin, China
{gaoshengliu,huanjing.yue,yjy}@tju.edu.cn

Abstract. Densely-sampled light field (LF) images are drawing increasing attention for their wide applications, such as 3D reconstruction, virtual reality, and depth estimation. However, due to the hardware restriction, it is usually challenging and costly to capture them. In this paper, we propose a coarse-to-fine convolutional neural network (CNN) for LF angular super-resolution (SR), which aims at generating densely-sampled LF images from sparse observations. Our method contains two stages, *i.e.*, coarse-grained novel views synthesis and fine-grained view refinement. Specifically, our method first extracts the multi-scale correspondence in the sparse views and generates coarse novel views. Then we propose a structural consistency enhancement module to regularize them for LF parallax structure preservation. Experimental results on both real-world and synthetic datasets demonstrate that our method achieves state-of-the-art performance. Furthermore, we show the promising application of the reconstructed LF images by our method on the depth estimation task.

Keywords: Light field · Angular super-resolution · View synthesis

1 Introduction

Different from 2D imaging, light field (LF) imaging can record light rays on not only spatial but also angular dimensions, where the geometric information of the real-world scenes is encoded. Benefited from the rich angular clues, the densely-sampled LF images are highly desirable for various applications, such as 3D reconstruction [10, 24], post-refocusing [4], and virtual reality [28]. However, it is challenging and costly to acquire densely-sampled LF images. For example, the conventional LF imaging devices, including camera array [19] and computer-control gantry [16], are usually bulky and expensive. The recent hand-held commercial LF cameras [1, 2] suffer from the intrinsic tradeoff between spatial and angular resolutions since the sensor is shared for spatial-angular recording. To tackle the tradeoff, developing computational methods to enhance the angular resolutions of LF images is an attractive topic in recent years.

As shown in Fig. 1, the LF image can be represented by an array of sub-aperture images (SAIs). The horizontally or vertically stacked SAIs build up an epipolar-plane image (EPI) volume, and the EPIs are the 2D slices of the EPI volume. In the literature, various approaches have been proposed from different perspectives, *e.g.* reconstructing novel SAIs, or high-resolution EPI volume (EPIs). And they can be roughly classified into two categories, *i.e.*, non-learning-based methods [12–15, 18, 29, 31] and learning-based methods [8, 9, 17, 20–23, 25, 27].

The non-learning-based methods usually either predict the scene depth from the sparse views as auxiliary information [12, 18] or require handcrafted image priors, such as sparsity in continuous Fourier domain [14] or shearlet transform domain [15]. Recently, following the board application of deep learning technology, the learning-based methods have shown remarkable performance in LF angular SR. Among them, both depth-dependent methods [8, 9, 20] and depth-independent methods [11, 17, 21–23, 25, 27] are widely studied. Specifically, Yoon *et al.* [27] proposed to jointly reconstruct the high-spatial and -angular resolution LF images. However, their performance is limited by the under-used spatial-angular correlations in surrounding SAIs. Kalantari *et al.* [9] proposed an end-to-end network, which estimates the scene depth that is utilized to synthesize novel SAIs via a physically-based warping operation. They also proposed a color network to refine the warped views. Yeung *et al.* [25] proposed to utilize spatial-angular alternating convolutions [26] to explore the spatial-angular clues in the input SAIs, and the final novel SAIs are generated via 4D convolutions. Wang *et al.* [17] proposed a pseudo-4D CNN framework to reconstruct high-resolution EPI volumes. Wu *et al.* [21, 23] introduced a *blur-restoration-deblur* scheme that works on 2D EPIs. Later, they proposed to fuse the pre-upsampled sheared EPIs [20] to generate the final high-resolution EPIs. By taking advantage of the geometry information, Jin *et al.* [8] proposed a depth-based network for LF images with large baselines. Recently, Wu *et al.* [22] incorporated the non-local attention mechanism to explore the spatial-angular correlations in the EPI volumes.

Though the previous methods have achieved remarkable performance, the depth-dependent methods rely heavily on the photo-consistency assumption and fail to handle the non-Lambertian effect. The multi-scale correlations in the input SAIs are under-exploited in the depth-independent methods. In addition, there is still room to improve the reconstruction quality of parallax structure in reconstructed LF images. Based on the above observation, in this paper, we develop a depth-independent coarse-to-fine method for LF angular SR. Specifically, our method first models the multi-scale correlations in the sparse views and synthesizes coarse novel views. Then, we propose a structural consistency enhancement module to regularize the intermediate results. We conduct experiments on both real-world and synthetic LF datasets, and the results demonstrate that our method achieves state-of-the-art performance in quantitative results and visual quality.

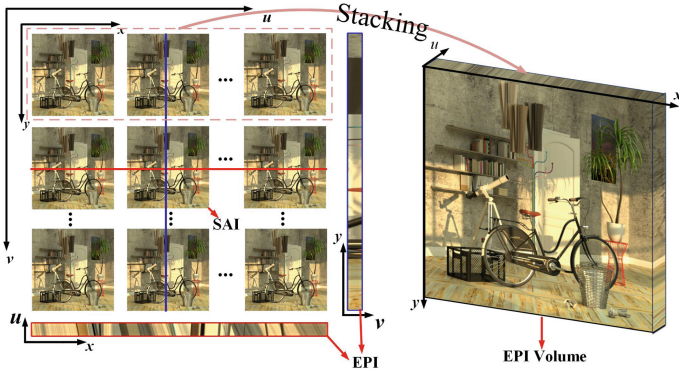


Fig. 1. Illustrations of LF images. The x, y are spatial coordinates and u, v are the angular coordinates. The LF image can be represented as an array of sub-aperture images (SAIs). The epipolar-plane image (EPI) volume is constructed by stacking one-direction SAIs. The EPIs are the 2D slices of the EPI volume.

2 Proposed Method

The two-plane model [5] is widely utilized to parameterize the 4D LF as $L(x, y, u, v) \in \mathbb{R}^{U \times V \times H \times W}$. Given a sparsely-sampled 4D LF image $L_{LR} \in \mathbb{R}^{U \times V \times H \times W}$ with angular resolution of $U \times V$ and spatial resolution of $H \times W$. This paper aims to reconstruct the corresponding densely-sampled LF image $L_{HR} \in \mathbb{R}^{\alpha U \times \alpha V \times H \times W}$. We follow previous work [11, 17, 21–23, 25] to perform the proposed method on the Y channel of the input LF images, and the Cb and Cr channels are up-sampled using bicubic algorithm on the angular domain. The overall architecture of our method is depicted in Fig. 2 (a), which consists of a coarse-grained view synthesis sub-network (CVSNet) and a fine-grained view refinement sub-network (FVRNet). Specifically, the CVSNet takes the sparsely-sampled LF image as inputs and generates novel views. Then the input views and the synthesized views are concatenated to generate the intermediate results. Finally, the intermediate results are fed into FVRNet to generate the final densely-sampled LF image. In the following, we give the details of CVSNet and FVRNet.

2.1 Coarse-grained View Synthesis Sub-network

Previous methods [11, 25] utilize spatial-angular alternating convolution or 3D convolution to extract the correlations in the sparsely-sampled LF image. However, the multi-scale correlations are under-exploited. To this end, we introduce CVSNet which is a modified UNet architecture to model the multi-scale correlations in the sparse views. The architecture of CVSNet is depicted in Fig. 2 (b).

As shown in Fig. 2 (b), the backbone of the proposed CVSNet follows the encoder-decoder structure with skip connections of the UNet. The filter num-

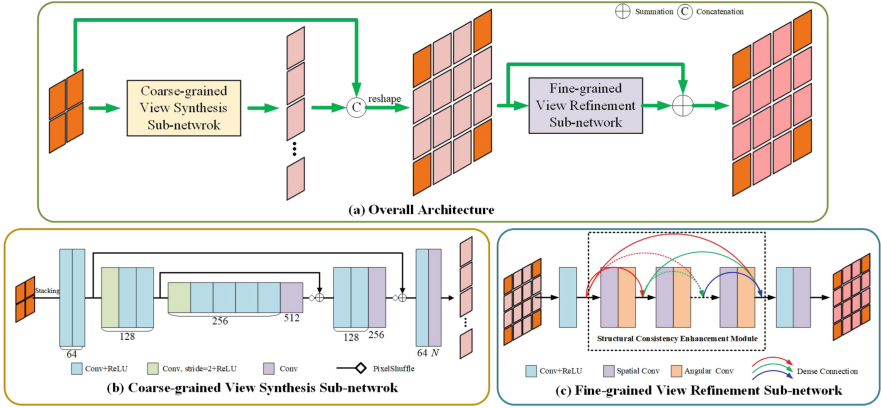


Fig. 2. The Framework of the proposed method. It contains two sub-networks, *i.e.*, coarse-grained views synthesis network (CSVNet), and fine-grained view refinement network (FVRNet). The CSVNet is utilized to synthesize novel views and FVRNet is utilized to refine them.

ber of each convolutional layer is shown at the bottom of each block. The filter number of the last convolutional layer N is the number of novel views to be synthesized (*e.g.*, $N = 60$ for $2 \times 2 \rightarrow 8 \times 8$ angular SR). The CVSNet consists of two down-sampling and up-sampling operations. The down-sampling operation is achieved by applying a convolutional layer with stride 2, and the up-sampling operation is achieved by the PixelShuffle layer. Specifically, the input SAIs are stacked along the channel dimension before being fed into CVSNet. Then CVSNet generates a set of novel SAIs by incorporating the multi-scale correlations in the SAIs of the input sparsely-sampled LF image, which can be represented as

$$L_{\text{Coarse}} = H_{\text{CVSNet}}(L_{\text{LR}}), \tag{1}$$

where the H_{CVSNet} denotes the CVSNet and L_{Coarse} denotes the synthesized views. The generated novel views and the input views are concatenated, which are fed into the FVRNet.

2.2 Fine-grained View Refinement Sub-network

An important property of the LF image is the valuable 4D structure, which is also known as the parallax structure. To produce high-quality high-angular-resolution LF images, the parallax structure should be well preserved. In our coarse views synthesis, the novel views are synthesized without considering the structural consistency among the intermediate views. Therefore, further refinement on the intermediate LF image to enhance the structural consistency is required.

To preserve the LF parallax structure, an intuitive way is to apply high-dimensional (*e.g.*, 4D and 3D) CNNs. However, high-dimensional CNNs will

bring a huge number of parameters and computational complexity. As an alternative, the spatial-angular alternating (SAA) convolution [26] is proposed to utilize interleaved 2D convolutions on spatial and angular dimensions. However, as analyzed in [7], the spatial angular alternating convolutions still suffer from inefficient feature flow. To this end, we introduce a structural consistency enhancement (SCE) module by combining the spatial angular alternating convolutions and dense connections to regularize the intermediate LF image.

The structure of the FVRNet is depicted in Fig. 2 (c). In FVRNet, a shared-weight convolution is first applied to extract the initial feature from the intermediate SAIs, generating $\mathcal{F}_{\text{init}} \in \mathbb{R}^{\alpha^2 UV \times C \times H \times W}$, where C is the number of feature maps. Then $\mathcal{F}_{\text{init}}$ is fed into the structural consistency enhancement module to explore the spatial-angular correlations. In the SCE module, the spatial convolution is performed on each SAI-wise feature, $\mathcal{F}_s^i \in \mathbb{R}^{C \times H \times W}$, $i \in \{1, 2, \dots, \alpha^2 UV\}$. Then the output features are reshaped to stacks of angular patches, *i.e.*, $\mathcal{F}_a \in \mathbb{R}^{HW \times C \times \alpha U \times \alpha V}$. The angular convolution is performed on each angular feature, $\mathcal{F}_a^j \in \mathbb{R}^{C \times \alpha U \times \alpha V}$, $j \in \{1, 2, \dots, HW\}$.

We apply dense connection in the SCE module to enhance the information flow. Specifically, The output of the k -th ($2 \leq k \leq 4$) SAA convolution can be formulated by

$$\mathcal{F}_{\text{SAA}}^{(k)} = H_{\text{SAA}}^k([\mathcal{F}_{\text{SAA}}^{(1)}, \dots, \mathcal{F}_{\text{SAA}}^{(k-1)}]), \quad (2)$$

where $[\cdot]$ denote the concatenation operation, $\mathcal{F}_{\text{SAA}}^{(k)}$ denotes the output feature of the k -th SAA convolution, H_{SAA}^k denotes the k -th SAA convolution.

Specifically, we utilize four SAA convolutions in the SCE module. The output feature of the SCE module is processed by two convolutional layers to output the final results. Finally, a global residual connection in FVRNet is also performed.

2.3 Training Details

We utilize the L_1 loss to minimize the distance between the ground truth and the output of our method

$$\mathcal{L} = \|L_{\text{SR}} - L_{\text{GT}}\|_1, \quad (3)$$

where L_{SR} is the angularly super-resolved LF image. The number of filters of the convolutional layers in FVRNet is set to $C = 32$ (the last convolution has one filter). We cropped the SAIs into patches of 64×64 for training. The batch size was set to one, and the learning rate was initially set to $2e-4$ which is reduced by half after every 15 epochs. The training was stopped after 70 epochs. During training, the data is augmented via random flipping and 90-degree rotation. We implemented the network in PyTorch and utilized an NVIDIA RTX 2080 TI GPU to train it. The ADAM algorithm is applied to optimize the network.

Table 1. Quantitative comparisons (PSNR/SSIM) with the state-of-the-arts on real-world datasets for $2 \times 2 \rightarrow 8 \times 8$ angular SR. The best results are highlighted in red.

Test sets	Kalantari <i>et al.</i> [9]	Wu <i>et al.</i> [21]	Wu <i>et al.</i> [20]	Yeung <i>et al.</i> [25]	SAA-Net [22]	Ours
<i>30scenes</i>	39.88/0.979	35.25/0.928	36.74/0.950	40.67/0.979	39.90/0.977	41.83 / 0.993
<i>Occlusions</i>	35.34/0.962	33.02/0.922	32.98/0.943	36.24/0.977	35.04/0.962	37.19 / 0.986
<i>Reflective</i>	34.99/0.940	33.64/0.927	34.38/0.941	35.72/0.945	35.36/0.945	36.27 / 0.970
Average	36.74/0.960	33.97/0.926	34.70/0.945	37.54/0.967	36.77 / 0.961	38.43 / 0.983

3 Experiments

The experiments are conducted on both real-world and synthetic LF scenes. Specifically, we select 100 real-world scenes from Stanford Lytro Archive [3] and Kalantari *et al.* [9] to train the model. We extract the central 8×8 SAIs from the original 14×14 SAIs for training and testing. For synthetic scenes, we select 20 scenes from *HCInew* dataset [6] to train the model. We compare our methods with five state-of-the-art methods, *i.e.*, Kalantari *et al.* [9], Yeung *et al.* [25], Wu *et al.* [21], Wu *et al.* [20], and SAA-Net [22]. The Kalantari *et al.* [9], Yeung *et al.* [25], and Wu *et al.* [20] are trained using the same training datasets as ours. Since the training codes of Wu *et al.* [21] and SAA-Net [22] are not available, we test their methods using their released models. The comparisons are conducted on $2 \times 2 \rightarrow 8 \times 8$ task. Specifically, we sample the input sparse 2×2 views from the four corners of the ground-truth 8×8 views. To compute the PSNR and SSIM scores of the angular SR results, only the Y channel of synthesised views (*e.g.*, 60 novel views for $2 \times 2 \rightarrow 8 \times 8$ angular SR) are utilized for quantitative evaluation.

3.1 Comparison on Real-world Scenes

We used three test sets which contains 70 real-world LF scenes are utilized for performance comparison, namely *30scenes* [9], *Occlusions* [3], and *Reflective* [3].

Table 1 lists the PSNR and SSIM scores for each test set and the average results for the three test sets. From Table 1, it can be observed that our method consistently outperforms other state-of-the-art methods. Compared with Kalantari *et al.* [9], our method achieves an average gain of 1.69 dB. This is because their method incorporates estimated disparities to warp novel views from input views. However, the warping operation is difficult to handle challenging cases, such as occluded regions and non-Lambertian surfaces. The results of Wu *et al.* [21] and Wu *et al.* [20] are inferior to others. The reason is that their methods work on one-direction EPis, which can not fully exploit the spatial correlations in the SAIs. Our method outperforms Yeung *et al.* [25] by 1.16 dB on *30scenes* test set. Compared with SAA-Net [22], our method achieves 1.66 dB and 0.022 gain in terms of average PSNR and SSIM.

Figure 3 presents the visual comparison results of two scenes. It can be observed that our method produces fine-grained details in the synthesized views.

Table 2. Quantitative comparisons (PSNR/SSIM) with the state-of-the-arts on synthetic datasets for $2 \times 2 \rightarrow 8 \times 8$ angular SR. The best results are highlighted in red.

Test sets	Kalantari <i>et al.</i> [9]	Wu <i>et al.</i> [21]	Wu <i>et al.</i> [20]	Yeung <i>et al.</i> [25]	SAA-Net [22]	Ours
<i>HCInew</i>	32.37/0.905	28.82/0.773	27.85/0.793	32.07/0.895	30.54/0.862	32.87/0.951
<i>HCIold</i>	38.22/0.944	34.79/0.874	36.03/0.902	37.44/0.927	38.40/0.918	39.89/0.964
Average	35.30 / 0.925	31.81 / 0.824	31.94 / 0.848	34.76/0.911	34.47/0.890	36.38/0.957

The EPI-based method, Wu *et al.* [21], Wu *et al.* [20], SAA-Net [22] are prone to producing ghosting artifacts. Kalantari *et al.* [9] struggles to recover the boundary of the rock in scene *Rock*. The results of Yeung *et al.* [25] suffer from ringing artifacts. We also provide the EPIs recovered by each method for comparison. We can observe that Wu *et al.* [21] and Wu *et al.* [20] can not recover the linear patterns in EPIs. The EPIs reconstructed by Yeung *et al.* [25] and SAA-Net [22] also have artifacts. By contrast, our method recovers fewer artifacts and more linear structures. This demonstrates that our method is a strong baseline for high-quality LF angular SR.

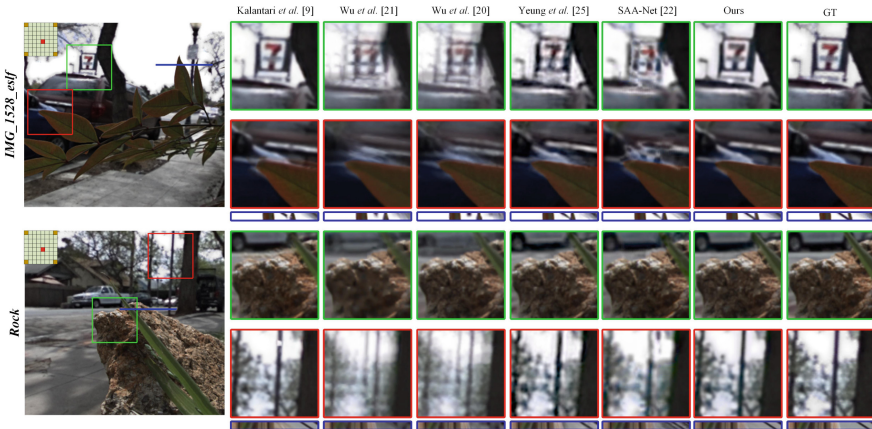


Fig. 3. Visual comparison on real-world scenes for $2 \times 2 \rightarrow 8 \times 8$ angular SR. We selected patches (highlighted using green and red boxes) in SAI that locates at the angular position of (5, 5). The EPIs are cut along the blue line. (Color figure online)

3.2 Comparison on Synthetic Scenes

For synthetic scenes, we select four scenes from *HCInew* and five scenes from *HCIold* dataset for comparison. Table 2 lists the quantitative comparison results with the state-of-the-art methods in terms of PSNR and SSIM. From Table 2, we can observe that our method achieves the best performance. Specifically, our

method outperforms Yeung *et al.* [25] by 2.45 dB on *HCIold* dataset. Compared with Kalantari *et al.* [9], our method achieves an average gain of 1.08 dB. Compared with Wu *et al.* [21] and Wu *et al.* [20], Ours achieves more than 4 dB on average. Compared with SAA-Net [22], Ours also achieves an average gain of 1.91 dB.

Figure 4 presents the visual comparisons. It can also be observed that our method produces the most fine-grained details. For scene *Herbs*, the bowls (highlighted in green box) reconstructed by other methods are over smooth. For scene *StillLife*, the tablecloth recovered by Kalantari *et al.* [9], Yeung *et al.* [25], and SAA-Net [22] have severe artifacts. The results of Wu *et al.* [21] and Wu *et al.* [20] are also blurry. By contrast, our method recovers more fine-granular textures.

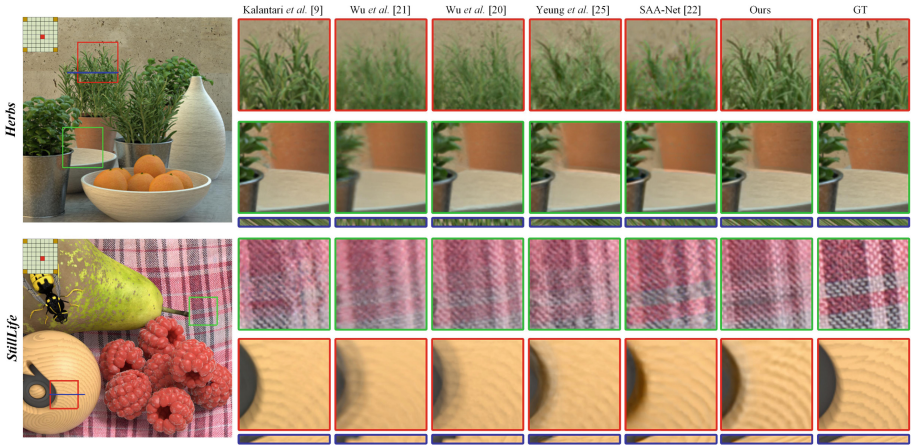


Fig. 4. Visual comparison on synthetic scenes for $2 \times 2 \rightarrow 8 \times 8$ angular SR. We selected patches (highlighted using green and red boxes) in SAI that locates at the angular position of (5, 5). The EPIs are cut along the blue line. (Color figure online)

3.3 Ablation Study

In this subsection, We perform ablation studies to demonstrate the effectiveness of the proposed methods. Specifically, we take advantage of the multi-scale correlations in the sparsely-sampled input, and we propose the SCE module to refine the coarse-grained synthesized views. To this end, we design several variants to verify the effectiveness of the introduced strategy or module. The experiments are conducted on $2 \times 2 \rightarrow 8 \times 8$ task. The models are trained using the real-world 100 scenes, and tested on the *30scenes*, *Occlusions*, and *Reflective* test sets.

The Effectiveness of Multi-scale Modeling. To verify the effectiveness of multi-scale modeling, we design a variant by replacing the down-sampling

Table 3. Ablation results on real-world datasets. The best results are highlighted in red.

Variants	<i>30Scenes</i>	<i>Occlusions</i>	<i>Reflective</i>
<i>w/o</i> _Multi-scale Modeling	41.54/0.9921	36.72/0.9842	36.20/0.9713
<i>w/o</i> _CSE	41.54/0.9917	36.85/0.9845	36.07/ 0.9714
<i>w/o</i> _Dense Connection	41.67/0.9922	37.04/0.9852	36.16/0.9699
<i>w</i> _2SAACnv	41.68/0.9922	36.92/0.9848	36.14/0.9713
<i>w</i> _3SAACnv	41.72/0.9923	37.02/0.9852	36.18/0.9704
Ours	41.83/0.9925	37.19/0.9857	36.27/0.9704

convolutions with normal convolutions and removing the PixelShuffle layers in CVSNet.

Table 3 lists the quantitative results of the ablation study. It can be observed that *w/o*_Multi-scale Modeling suffers from a decrease of 0.29 dB on *30scenes* test set, and 0.47 dB on *Occlusion* test set. This is because the down-sampling operations in CVSNet can help enlarge the receptive field and explore the multi-scale correlations in the SAIs, which are beneficial for the view synthesis.

The Effectiveness of CSE. In the FVRNet, we introduced CSE module to refine the intermediate results. We then conduct several experiments to show the influence of CSE module. We first directly remove the CSE module in FVRNet. In Table 3, we can observe that the results of *w/o*_CSE are decreased by 0.29 dB on *30scenes*, which demonstrates the effectiveness of proposed CSE module. We then remove the dense connections in CSE module and only four SAA convolutions are maintained. We can observe that *w/o*_Dense Connection, the results suffer from a decrease of 0.16 dB on *30scenes*. This is because the dense connections can help enhance the feature flow. We also conduct the experiments by utilizing a different number of SAA convolutions in CSE module. From Table. 3, we can observe that the results of *w*_2SAACnv and *w*_3SAACnv are inferior to Ours (*w*_4SAACnv) by 0.15 dB and 0.11 dB on *30scenes* test set, respectively.

3.4 Depth Estimation

Since one of the most valuable information of the reconstructed LF image is the geometry information of the real-world scene, we further apply our method to depth estimation task to verify the ability to reveal the geometric structures. We utilize SPO [30] to predict the scene depth estimation from the reconstructed densely-sampled LF image. We also compare the visual quality of estimated depth maps with Wu *et al.* [20] and SAA-Net [22]. The ground-truth depth map is estimated from the ground-truth densely-sampled LF image. Figure 5

presents the visual results. We can observe that our method achieves promising depth prediction, such as the boundary of leaves in scene *occlusion_2_eslf* and the rock in scene *Rock*.

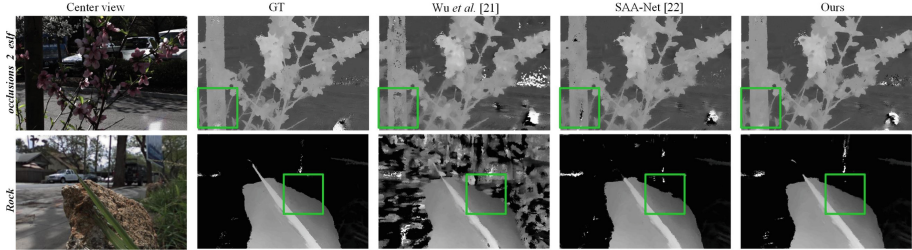


Fig. 5. Visual comparison of depth estimation.

4 Conclusion

In this paper, we propose a coarse-to-fine network for LF angular SR, which aims to reconstruct densely-sampled LF images from sparsely-sampled ones. We introduce two sub-networks, *i.e.*, a CVSNet to synthesize novel views, and an FVRNet to refine the coarse views. Specifically, CVSNet contains a UNet architecture to extract the multi-scale correspondence in the sparse views and generate coarse novel views. In FVRNet, we propose a structural consistency enhancement module to refine the coarse views and help preserve the parallax structure of LF image. The experiments are conducted on both real-world and synthetic datasets, and the experimental results demonstrate that our method achieves state-of-the-art performance. We further apply our method to the depth estimation task, and the visual results show our promising ability to predict the geometric information from scenes. Our method also has limitations, *e.g.*, the visual reconstruction quality still has obvious distance from the ground-truth images in some challenging regions, such as the tablecloth in scene *StillLife* (Fig. 4). In future work, we will explore more effective strategies to improve the visual quality.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 62072331.

References

1. Lytro illum. <https://www.lytro.com/>
2. Raytrix. <https://www.raytrix.de/>
3. The stanford lytro light field archive. <http://lightfields.stanford.edu/LF2016.html>. Accessed 16 Oct 2021

4. Fiss, J., Curless, B., Szeliski, R.: Refocusing plenoptic images using depth-adaptive splatting. In: 2014 IEEE International Conference on Computational Photography (ICCP), pp. 1–9 IEEE (2014)
5. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 43–54 (1996)
6. Honauer, K., Johannsen, O., Kondermann, D., Goldluecke, B.: A dataset and evaluation methodology for depth estimation on 4D light fields. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10113, pp. 19–34. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54187-7_2
7. Hu, Z., Yeung, H.W.F., Chen, X., Chung, Y.Y., Li, H.: Efficient light field reconstruction via spatio-angular dense network. *IEEE Trans. Instrum. Meas.* **70**, 1–14 (2021)
8. Jin, J., Hou, J., Yuan, H., Kwong, S.: Learning light field angular super-resolution via a geometry-aware network. **34**(07), pp. 11141–11148 (2020)
9. Kalantari, N.K., Wang, T.C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. *ACM Trans. Graphics (TOG)* **35**(6), 1–10 (2016)
10. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.H.: Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.* **32**(4), 73 (2013)
11. Liu, D., Huang, Y., Wu, Q., Ma, R., An, P.: Multi-angular epipolar geometry based light field angular reconstruction network. *IEEE Trans. Comput. Imaging* **6**, 1507–1522 (2020)
12. Mitra, K., Veeraraghavan, A.: Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 22–28 IEEE (2012)
13. Pujades, S., Devernay, F., Goldluecke, B.: Bayesian view synthesis and image-based rendering principles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3906–3913 (2014)
14. Shi, L., Hassanieh, H., Davis, A., Katabi, D., Durand, F.: Light field reconstruction using sparsity in the continuous fourier domain. *ACM Trans. Graphics (TOG)* **34**(1), 1–13 (2014)
15. Vagharshakyan, S., Bregovic, R., Gotchev, A.: Light field reconstruction using shearlet transform. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(1), 133–147 (2017)
16. Vaish, V., Adams, A.: The (new) stanford light field archive (2008). <http://lightfield.stanford.edu/>
17. Wang, Y., Liu, F., Wang, Z., Hou, G., Sun, Z., Tan, T.: End-to-end view synthesis for light field imaging with pseudo 4DCNN. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 340–355. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_21
18. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(3), 606–619 (2013)
19. Wilburn, B., et al.: High performance imaging using large camera arrays. In: ACM SIGGRAPH 2005 Papers, pp. 765–776 (2005)
20. Wu, G., Liu, Y., Dai, Q., Chai, T.: Learning sheared epi structure for light field reconstruction. *IEEE Trans. Image Process.* **28**(7), 3261–3273 (2019)
21. Wu, G., Liu, Y., Fang, L., Dai, Q., Chai, T.: Light field reconstruction using convolutional network on epi and extended applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(7), 1681–1694 (2018)

22. Wu, G., Wang, Y., Liu, Y., Fang, L., Chai, T.: Spatial-angular attention network for light field reconstruction. *IEEE Trans. Image Process.* **30**, 8999–9013 (2021)
23. Wu, G., Zhao, M., Wang, L., Dai, Q., Chai, T., Liu, Y.: Light field reconstruction using deep convolutional network on epi. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6319–6327 (2017)
24. Wu, J., et al.: Iterative tomography with digital adaptive optics permits hour-long intravital observation of 3d subcellular dynamics at millisecond scale. *Cell* **184**, 3318–3332.e17 (2021)
25. Yeung, H.W.F., Hou, J., Chen, J., Chung, Y.Y., Chen, X.: Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018*. LNCS, vol. 11210, pp. 138–154. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_9
26. Yeung, H.W.F., Hou, J., Chen, X., Chen, J., Chen, Z., Chung, Y.Y.: Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Trans. Image Process.* **28**(5), 2319–2330 (2018)
27. Yoon, Y., Jeon, H.G., Yoo, D., Lee, J.Y., So Kweon, I.: Learning a deep convolutional network for light-field image super-resolution. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 24–32 (2015)
28. Yu, J.: A light-field journey to virtual reality. *IEEE MultiMedia* **24**, 104–112 (2017)
29. Zhang, F.L., Wang, J., Shechtman, E., Zhou, Z.Y., Shi, J.X., Hu, S.M.: Plenopatch: patch-based plenoptic image manipulation. *IEEE Trans. Visual. Comput. Graphics* **23**(5), 1561–1573 (2016)
30. Zhang, S., Sheng, H., Li, C., Zhang, J., Xiong, Z.: Robust depth estimation for light field via spinning parallelogram operator. *Comput. Vis. Image Underst.* **145**, 148–159 (2016)
31. Zhang, Z., Liu, Y., Dai, Q.: Light field from micro-baseline image pair. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3800–3809 (2015)