

# DeOccNet: Learning to See Through Foreground Occlusions in Light Fields

Yingqian Wang<sup>1</sup>, Tianhao Wu<sup>1</sup>, Jungang Yang<sup>1\*</sup>,  
Longguang Wang<sup>1</sup>, Wei An<sup>1</sup>, Yulan Guo<sup>1,2</sup>

<sup>1</sup>College of Electronic Science and Technology, National University of Defense Technology, China

<sup>2</sup>School of Electronics and Communication Engineering, Sun Yat-sen University, China

{wangyingqian16, yangjungang, yulan.guo}@nudt.edu.cn

## Abstract

Background objects occluded in some views of a light field (LF) camera can be seen by other views. Consequently, occluded surfaces are possible to be reconstructed from LF images. In this paper, we handle the LF de-occlusion (LF-DeOcc) problem using a deep encoder-decoder network (namely, DeOccNet). In our method, sub-aperture images (SAIs) are first given to the encoder to incorporate both spatial and angular information. The encoded representations are then used by the decoder to render an occlusion-free center-view SAI. To the best of our knowledge, DeOccNet is the first deep learning-based LF-DeOcc method. To handle the insufficiency of training data, we propose an LF synthesis approach to embed selected occlusion masks into existing LF images. Besides, several synthetic and real-world LFs are developed for performance evaluation. Experimental results show that, after training on the generated data, our DeOccNet can effectively remove foreground occlusions and achieves superior performance as compared to other state-of-the-art methods. Source codes are available at: <https://github.com/YingqianWang/DeOccNet>.

## 1. Introduction

Seeing through foreground occlusions is beneficial to many computer vision applications such as detection and tracking in surveillance [5, 13, 33, 34]. However, due to foreground occlusions, some rays cannot hit the sensors of traditional single-view cameras (e.g., Digital Single Lens Reflex). Therefore, objects behind occlusions cannot be fully observed and reliably reconstructed. In recent years, camera arrays [8, 18, 19, 26, 27] have undergone a rapid development since they can record light fields (LFs) and provide a large number of viewpoints with rich angular information. The complementary information among different viewpoints is

\*Corresponding author

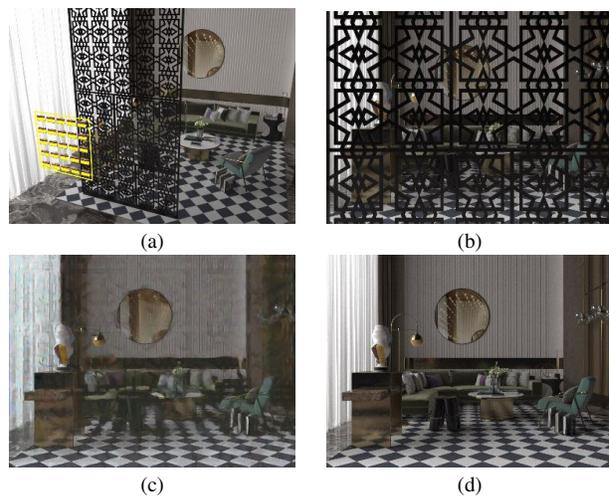


Figure 1: An illustration of LF-DeOcc using our rendered scenes Syn01. (a) Configuration of the scene. Yellow boxes with  $5 \times 5$  blocks represent camera arrays. (b) Occluded center-view SAI. (c) Results of our DeOccNet. (d) Occlusion-free groundtruth.

beneficial for the reconstruction of occluded surfaces since background objects occluded in some views can be seen by other views.

As illustrated in Fig. 1, light field de-occlusion (LF-DeOcc) aims at removing foreground occlusions using sub-aperture images (SAIs) captured by a camera array<sup>1</sup>. The pioneering work on LF-DeOcc is proposed by Vaish *et al.* [18] using a refocusing method. However, this method cannot recover a clean surface of occluded objects since rays from occlusions and background are mixed. In fact, it is important but challenging to correctly select pixels only belonging to occluded objects. To this end, existing methods

<sup>1</sup>In the area of LF-DeOcc, images captured by camera arrays are widely used due to their wide baselines. Therefore, we follow the existing work and use camera arrays for LF-DeOcc.

[11,12,31,35] generally built different models to handle LF-DeOcc problem. Due to the highly complex structures of scenes in real world, these methods with handcrafted feature extraction and stereo matching techniques cannot achieve satisfactory performance. In recent years, deep learning has been successfully used in different LF tasks such as depth estimation [14, 15], image super-resolution [22, 23, 36, 39], view synthesis [21, 28, 29], and LF intrinsics [1, 2]. These networks have achieved state-of-the-art performance in numerous areas. However, to the best of our knowledge, deep learning has not been used for LF-DeOcc due to several issues. In this paper, we design a novel and effective paradigm, and propose the first deep learning network (*i.e.*, *DeOccNet*) to handle LF-DeOcc problem. Specifically, we summarize three major challenges in deep learning-based LF-DeOcc, and provide solutions to these challenges using our proposed paradigm.

The first challenge is that, as compared to LF depth estimation networks [14, 15] and LF super-resolution networks [22, 23, 36, 39], LF-DeOcc networks should use as much information from occluded surfaces as possible, while maintaining a larger receptive field to cover occlusions of different types and scales. We address this challenge by employing an encoder-decoder network to encode LF structures. We concatenate all SAIs along the channel dimension to fully use the information of occluded surfaces. Besides, we use a residual atrous spatial pyramid pooling (ASPP) module to extract multi-scale features and enlarge receptive fields.

The second challenge is that, as compared to single image inpainting networks [9, 10, 32, 38], LF-DeOcc networks have to learn the scene structure to automatically recognize, label and remove foreground occlusions. We address this challenge by setting the occlusion-free center-view SAI as groundtruth, and train our *DeOccNet* in an end-to-end manner. In this way, our network can recognize occlusions from background through disparity discrepancy, and automatically remove foreground occlusions.

The third challenge is that, LF-DeOcc networks face an insufficiency of training data since large-scale LF datasets with removable foreground occlusions are unavailable. Moreover, test scenes are also insufficient for performance evaluation. We address this challenge by proposing a data synthesis approach to embed different occlusion masks into existing LF images. Using this approach, more than 1000 LFs are generated to train our network. Moreover, we develop several synthetic and real-world LFs for performance evaluation.

Experimental results have demonstrated the effectiveness of our paradigm. Our *DeOccNet* achieves superior performance on both synthetic and real-world scenes as compared to other state-of-the-art methods.

## 2. Related Works

### 2.1. Single image inpainting

Single image inpainting methods aim at filling holes in an image using both neighborhood information and global priors. The major challenge of single image inpainting lies in synthesizing visually realistic and semantically plausible pixels for missing regions. Recent deep learning based methods [9, 10, 32, 38] have achieved promising results for inpainting large missing regions in an image. Specifically, Yu *et al.* [38] proposed a deep generative model-based inpainting method to synthesize novel image structures and textures. Liu *et al.* [9] used partial convolutions for inpainting with irregular holes and achieved the state-of-the-art performance.

Compared to single image inpainting, LF-DeOcc can use complementary information provided by SAIs to generate improved results. The difference between single image inpainting and LF-DeOcc is significant. In single image inpainting, holes or masks are always pre-defined. However, LF-DeOcc requires automatical extraction of foreground occlusions by analyzing scene structures. That is, occlusions are closer to cameras than background objects, and thus have larger disparities. Due to the complex structures of real-world scenes, it is highly challenging for algorithms to correctly select rays originating from occluded objects.

### 2.2. Light field de-occlusion

LF-DeOcc is an active research topic and has been investigated for decades [11, 12, 17, 18, 31, 35]. Vaish *et al.* [18] proposed a refocusing method by warping each SAI by a specific value, and then averaging the warped SAIs along angular dimension. Due to the large equivalent aperture of camera arrays, when background is refocused on, occlusions are extremely blurred and the *see through* effect can be achieved. However, the resulting images are always blurred due to the indiscriminate use of rays from both occlusions and background. Vaish *et al.* further proposed an improved version using both median cost and entropy cost [17]. Since these methods [17, 18] do not exactly exploit scene structures, their performance is limited for scenes with heavy occlusions.

To solve this problem, Pei *et al.* proposed a pixel-labeling method [12] to remove occlusions. Specifically, pixels corresponding to occlusions are labeled by stereo matching and masked out during refocusing process, resulting in a clean image. However, this method can only generate images refocused on a specific depth, leaving objects in other depth ranges suffering from various degrees of blurs. Subsequently, they used an image-matting approach to perform all-in-focus synthetic aperture imaging [11]. Besides, Yang *et al.* used visible layers [35] to address the all-in-focus imaging issue. Xiao *et al.* [31] used k-means clus-

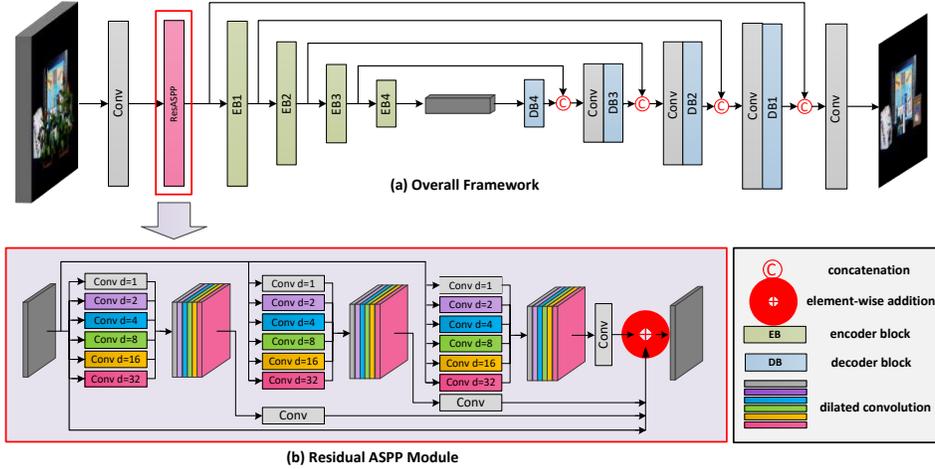


Figure 2: An overview of our *DeOccNet*. (a) The overall architecture. (b) The structure of the residual ASPP module.

tering to classify pixels of occlusions and background. All these methods [11, 12, 31, 35] use handcrafted feature extraction and stereo matching techniques, and cannot achieve a satisfactory performance in scenes with complex structures and heavy occlusions.

### 2.3. Deep learning in light field

Deep neural networks have been widely used in various LF tasks such as image super-resolution [22, 23, 36, 37, 39], view synthesis [21, 28, 29], and depth estimation [14, 15]. Compared to these tasks, networks for LF-DeOcc should have a larger receptive field and use more information of occluded surfaces. Currently, no existing work on deep learning based LF-DeOcc is available in literature. It is worth noting that works in [1, 2] are similar to ours. Specifically, a fully convolutional auto-encoder is proposed in [2] to separate diffuse and specular components of an LF. Both [2] and our work require high-level features and global priors of scene structures. Consequently, we built our network upon encoder-decoder architecture to encode LF structures. Note that, there are two significant differences between [2] and our network. First, only horizontal and vertical SAIs (*e.g.*, 9 cross-views in a  $5 \times 5$  LF) are used in [2]. In contrast, all SAIs are used in our network to fully exploit the information of occluded objects. Second, our *DeOccNet* uses residual ASPP module to enlarge the receptive field, and uses multiple skip layers to have a holistic understanding of the scene while preserving fine details.

## 3. The Proposed Method

### 3.1. Network architecture

The task of our *DeOccNet* is to replace pixels of occlusions with pixels from the background. To achieve this

task, our network is required to find correspondence and incorporate complementary information from SAIs. Note that, foreground occlusions generally have shallow depths and large disparities. That is, pixels of occlusions always have very large position variations among SAIs. Therefore, multi-scale features with large receptive fields are required for our network. In this paper, we first use a residual ASPP module for hierarchical feature extraction, and then use an auto-encoder to incorporate both spatial and angular information. The architecture of our *DeOccNet* is shown in Fig. 2. Different from existing LF networks [2, 15, 39] where only part of SAIs are stacked as inputs, we stacked all SAIs along the channel dimension (*e.g.*, 75 channels for  $5 \times 5$  RGB SAIs) to use as much information as possible since LF-DeOcc highly depends on the information of occluded objects. Consequently, our *DeOccNet* takes the stacked SAIs as its input, and finally generates an occlusion-free center-view SAI.

**Residual ASPP module.** In our network, the input volume is first processed by a  $1 \times 1$  convolution layer to generate features with a fixed depth (*i.e.*, 64 in this paper). Then, a residual ASPP module is used to generate hierarchical features. As shown in Fig. 2(b), similar to the module used in [20], we first combine six dilated convolutional layers (with dilation rates of 1, 2, 4, 8, 16, and 32) to form an ASPP group, then cascade three ASPP groups in residual manner to achieve high learning efficiency. The residual ASPP module can enlarge the receptive field and extract multi-scale information around occlusions. It is demonstrated in the ablation study (see Table 1) that our residual ASPP module is beneficial to the overall LF-DeOcc performance.

**Encoder pathway.** Features generated by the residual ASPP module are then transferred to the encoder pathway, where 4 encoder blocks are cascaded to incorporate both

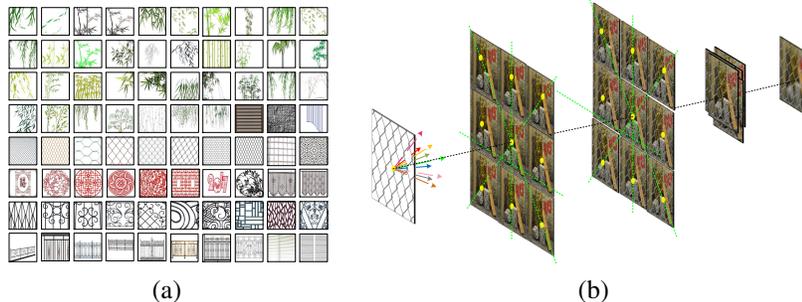


Figure 4: An illustration of our *Mask Embedding* approach. (a) Masks used in our approach. Note that, cropping and scaling are performed for better visualization. (b) The pipeline of our *Mask Embedding* approach. Here, a  $3 \times 3$  LF is used as an example.

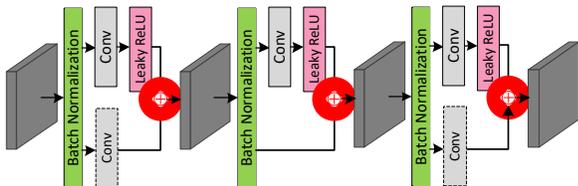


Figure 3: The structure of the encoder and decoder blocks. Note that, the encoder and decoder blocks share mirrored structures. That is, strided convolution is used in the third unit in each encoder block, while de-convolution is used in the first unit in each decoder block.

spatial and angular information. Specifically, as shown in Fig. 3, each encoder block contains three cascaded units. In each unit, the batch-normalized features are given to two separated paths to achieve local residual learning. The first path includes a  $3 \times 3$  convolution and a Leaky ReLU (with a leaky factor of 0.1), while the second path either keeps the input unchanged or passes it through a strided convolution or de-convolution. Therefore, features produced by the two paths have the same resolution. Features from both paths are added to produce the output of this unit. For each encoder block, the first two units keep depth and resolution unchanged, while the third unit applies a strided convolution (with a stride of 2) to halve the resolution and double the feature depth. Consequently, the final feature generated by the encoder (bottleneck of our *DeOccNet*) is  $1/16$  time of the input feature in resolution and has 1024 channels in depths.

**Decoder pathway.** After passing through the bottleneck, features are decoded through the decoder pathway. Note that, decoder blocks have mirrored structures as encoder blocks. That is, the decoder block is also chained by three residual units, and the first unit uses a de-convolution to exactly revert the encoder block on the corresponding level. Moreover, to preserve fine details in the final output image, features of different resolutions on the encoder

pathway are concatenated with their counterparts on the decoder pathway by skip connections. Consequently, the decoder can be guided to gradually add details onto restored images. Since feature depth is doubled by concatenation, an additional  $1 \times 1$  convolution layer is employed before the last three decoder blocks to halve the feature depth. Similarly, a  $1 \times 1$  convolution layer is also applied to the output feature to reduce its channel to 3.

### 3.2. Mask embedding for training data synthesis

It is important to provide sufficient data to train our *DeOccNet*. Although LFs with removable occlusions can be acquired by capturing real-world scenes with/without foreground occlusions, or by rendering synthetic scenes using softwares such as *3dsMax*<sup>2</sup> and *Blender*<sup>3</sup>, these approaches are significantly labour-intensive and even infeasible. Consequently, it is important to design an efficient approach to generate a large amount of data for network training. In this paper, we propose *Mask Embedding*, a training data synthesis approach to synthesize LFs with removable foreground occlusions. An illustration of our *Mask Embedding* approach is shown in Fig. 4.

As shown in Fig. 4(a), we manually collected 80 mask images from Internet using tags such as salix leaves, grids, fences, and paper cuts. All these masks are common foreground occlusions in daily life. Meanwhile, we collected 60 LFs from the Stanford LF dataset [16], the Old HCI 4D LF dataset [25], the New HCI 4D LF benchmark [4], and the MIT Synthetic LF Archive [7]. Note that, to improve the generalization capability of our network, the RGB channels of both LFs and mask images are randomly shuffled, and the masks are randomly selected to be embedded into shuffled LFs.

The pipeline of our *Mask Embedding* approach is illustrated in Fig. 4(b). We randomly pick a mask from the mask set and then embed it into each SAI according to LF con-

<sup>2</sup><https://www.autodesk.com/products/3ds-max/overview>

<sup>3</sup><https://www.blender.org/>

figurations. Specifically, a disparity with a shallow depth is randomly set and allocated to the selected mask. Then, the mask is warped according to the angular coordinate of the target view and the allocated disparity. Note that, bilinear interpolation is used when masks do not fall into integer coordinates. Next, warped masks are added to each SAI, resulting in LFs with foreground occlusions. Finally, refocusing is performed on each generated LF to check the LF configuration.

Although the *Mask Embedding* approach can easily generate a large number of LFs for training, the generated LFs only have occlusions in a single depth, which is significantly different from real-world scenarios. To address this issue, we use generated LFs as original LFs and repeat this process twice to synthesize LFs with occlusions at two and three depth layers. Using the proposed approach, we totally synthesize 1500 LFs to train our models. Although LFs synthesized by our approach only have fence-like and front-parallel occlusions, experimental results show that, our network trained on the synthetic data can generalize well to real-world cases (*e.g.*, CD scene in Fig. 6). That is, our network can successfully learn the scene structure through disparity discrepancy using LFs synthesized by our *Mask Embedding* approach.

### 3.3. Training details

We trained two models with  $5 \times 15$  and  $5 \times 5$  input SAIs. The  $5 \times 15$  model was used for the Stanford CD scene [16], and the  $5 \times 5$  model was used for our self-developed scenes. During the training phase, we used LFs synthesized by the *Mask Embedding* approach as the training data. All the 60 LFs were used for the  $5 \times 5$  model. In contrast, due to the angular resolution limitation of existing LF datasets, only LFs from the Stanford LF dataset [16] were used to generate training data for the  $5 \times 15$  model.

It is worth noting that foreground and background in a scene are relative concepts. That is, some occlusions can also be considered as background objects in multi-occlusion situations as shown in Fig. 5. Consequently, both training and test scenes should be rectified to specific depths for LF-DeOcc. In this paper, we perform rectification by cropping each SAI accordingly to make occlusions have positive disparity values while backgrounds have negative disparity values. In this way, our *DeOccNet* can effectively achieve LF-DeOcc by simply removing objects with positive disparity values. Finally, we cropped occluded SAIs into  $224 \times 224$  pixel patches by a stride of 112, and performed  $2 \times$  upsampling for data augmentation. Meanwhile, occlusion-free center-view SAI was cropped and upsampled accordingly to generate groundtruths.

Our *DeOccNet* was implemented in Pytorch on a PC with an Nvidia RTX 2080Ti GPU. All models were trained using an MSE loss and optimized using the Adam method



Figure 5: Multi-occlusion situation in LF-DeOcc. (a) Occluded center-view SAI. (b) Result of our *DeOccNet* with inputs being rectified at a shallow depth (the blue dotted line in (a)). The front-most tree is considered as a foreground occlusion. (c) Result of our *DeOccNet* with inputs being rectified at a deep depth (the red dotted line in (a)). Three front trees are considered as foreground occlusions. Consequently, different results can be generated by our *DeOccNet* with the same inputs being rectified at different depth values.

[6] with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and a batch size of 8. The initial learning rate was set to  $1 \times 10^{-3}$  and reduced to  $1 \times 10^{-4}$  after 100 epochs. The training took about 2 days and was stopped at 200 epochs.

## 4. Experiments

In this section, we first introduce the test scenes used in our experiments, and then compare our method to several state-of-the-art methods. Finally, we present ablation study and analyses.

### 4.1. Test scenes

**Real-world scenes.** We followed [11, 17, 35] and tested our method on the publicly available CD scene [16]. The original CD scene consists of 105 views distributed on a  $5 \times 21$  grid. We selected the central  $5 \times 15$  views for performance evaluation. Groundtruth image is provided by a second capture with occlusions being removed. Besides, we captured several real-world scenes using a moving *Leica Q* camera (with an  $F = 10$ ,  $f = 28$  mm lens) mounted on a gantry. It is argued in [24, 30] that the scanning scheme is equivalent to a single shot by a camera array in static occasions. We shifted the camera to 25 positions on a  $5 \times 5$  grid with 3 cm baselines. The captured images were calibrated using the method in [40].

**Synthetic scenes.** Since the number of real-world test scenes is very small, we rendered 4 synthetic scenes with removable foreground occlusions for further evaluation. All elements in our synthetic scenes were collected from Internet, and parameters (*e.g.*, lighting, depth range) were tuned to better reflect real scenes. The angular resolution of each scene was set to  $5 \times 5$ , while baselines and occlusion ranges were varied in different scenes. Occlusion-free center-view SAIs were also rendered for quantitative evaluation.

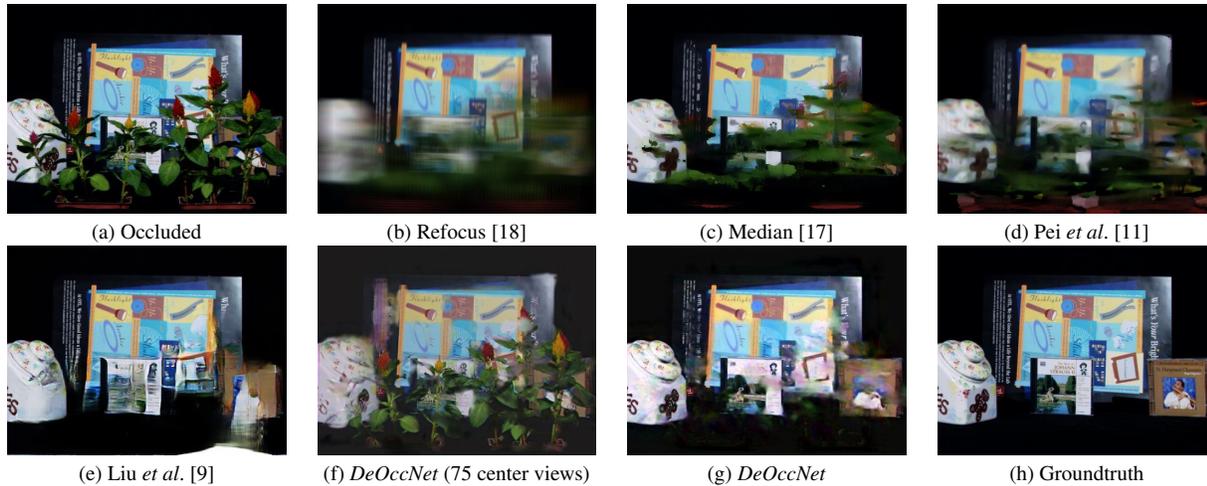


Figure 6: Qualitative results achieved on the CD scene [16] (with an occlusion rate of 40.2%). (a) Occluded center-view SAI. (b)-(e) Comparative results achieved by different methods. (f) Result achieved by our *DeOccNet* using 75 identical center-view SAIs as its inputs (discussed in Section 4.4). (g) Our result. (h) Occlusion-free center-view SAI.

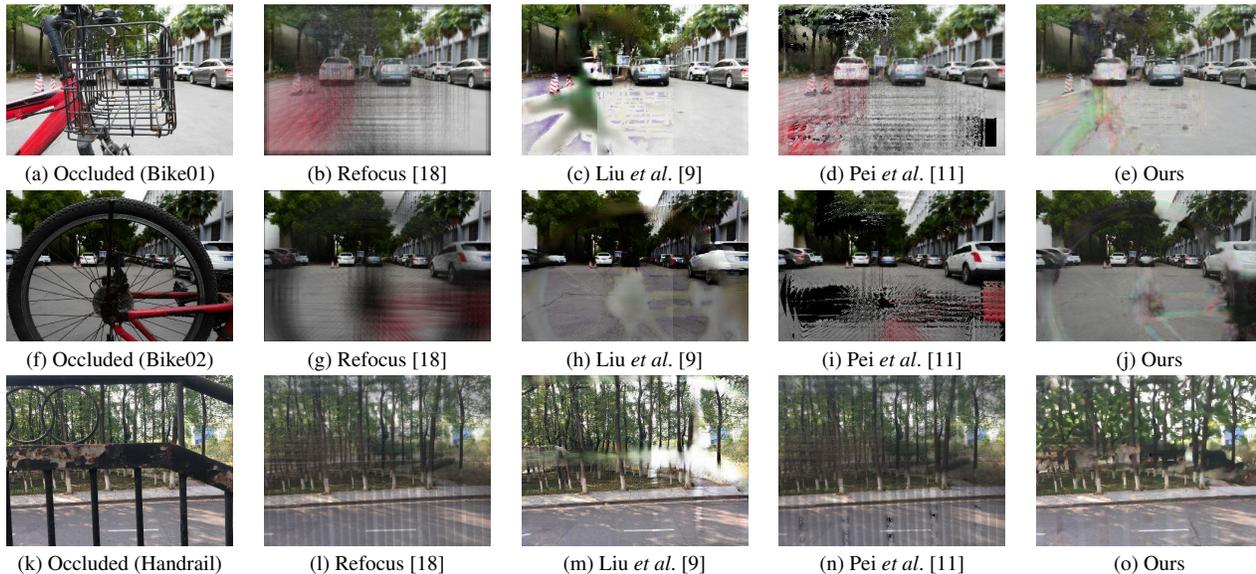


Figure 7: Qualitative results achieved on our self-developed real-world scenes. Note that, the occlusion rates of the three scenes are 61.8%, 57.7%, and 39.1%, respectively.

## 4.2. Comparison to the state-of-the-arts

We compared our *DeOccNet* to the state-of-the-art LF-DeOcc method [11]. We also used traditional refocusing method [18] and its improved version [17] as baselines. Moreover, to investigate the benefits of complementary information introduced by additional perspectives, we compared our method to the state-of-the-art image inpainting method [9]. Note that, the inpainting method [9] cannot automatically recognize occlusions in an image. Therefore, we manually labeled occlusions in each center-view

image. Since the codes of [11, 17, 18] are unavailable, we used our own implementations with their default parameter settings. Following the state-of-the-art image inpainting methods [9, 38], mean  $l_1$  error, peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) are used as quantitative evaluation metrics in this paper. Readers are referred to [9, 38] for more details about these metrics. Qualitative results on real-world and synthetic datasets are shown in Figs. 6, 7, and 8, and quantitative results are listed in Table 1.

It can be seen from Fig. 6(a) that occlusions in the CD

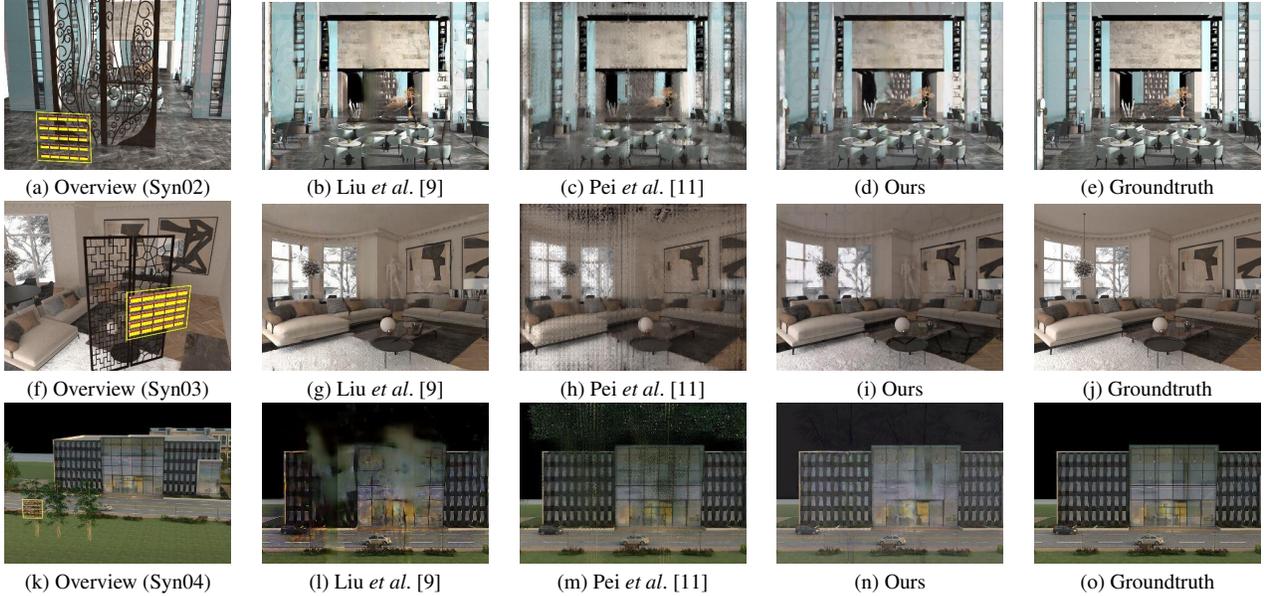


Figure 8: Qualitative results achieved on our synthetic scenes. Sub-figures on the leftmost column represent the configurations of different scenes, yellow boxes with  $5 \times 5$  blocks represent camera arrays. Note that, the occlusion rates of these scenes are 30.2%, 19.6% and 14.8%, respectively.

scene occupy a large portion. Consequently, even though 75 views are provided, it is still highly challenging to reconstruct occluded objects. Method [18] removes occlusions by warping and averaging SAIs. As shown in Fig. 6(b), although occlusions are extremely blurred at the focused depth, occluded object is also unclear and its contrast is low. Besides, since results produced by [18] only focus on limited depth ranges, objects in other depth ranges are highly blurred and cannot be recognized. In contrast, Vaish *et al.* [17] use median cost to achieve all-in-focus synthetic aperture imaging. However, the performance achieved by [17] is very limited.

Pei *et al.* [11] achieve a better performance than method [17]. That is because, only rays belonging to occluded objects are used in [11]. However, due to various textures and shapes of occlusions and background, it is difficult to exactly select rays only from background, and the resulting images can be deteriorated by incorrect classification of occlusion and background. More specifically, if pixels from occlusions are misused, the resulting images will be blurred (*e.g.*, Fig. 6(d)). In contrast, if background are incorrectly considered as occlusions, the resulting images will have blank areas (*e.g.*, Figs. 7(d) and 7(i)) since all information in these areas is removed.

The single image inpainting method [9] tries to hallucinate missing part and render reasonable results using knowledge learned from a large number of daily scenes (*e.g.*, the *ImageNet* database [3]). It can generate promising results on scenes with simple structure and small occlusions (*e.g.*,

Fig. 8(g)). However, since no additional information is used in the inpainting process, this method cannot handle scenes with rich textures/details or heavy occlusions (*e.g.*, Figs. 7(c) and 8(l)).

As compared to existing methods, our method achieves superior performance, especially on scenes with heavy occlusions (*e.g.*, scenes CD and Syn04). That is, *DeOccNet* successfully learns to discriminate occlusions from backgrounds, and incorporates information of occluded objects from different viewpoints. It can be also seen from Table 1 that our method achieves the best quantitative results on scenes CD, Syn01 and Syn02. Note that, our method is slightly inferior to [9] on scene Syn03. That is because, method [9] uses manually labeled groundtruth occlusions while our method do not rely on any groundtruth information. Since scene Syn03 has relatively small occlusions and simple textures, method [9] works well using learned spatial priors and labeled groundtruth occlusions.

### 4.3. Ablation study

We conducted ablation study to investigate the improvement introduced by the residual ASPP module and skip connections. Quantitative results are presented in Table 1.

First, we removed the residual ASPP module and re-trained *DeOccNet* from scratch using the same training data. We can observe from Table 1 that the residual ASPP module significantly contributes to the overall LF-DeOcc performance. Specifically, it introduces more than 3 dB improvement in average PSNR, and nearly 0.1 improvement

Table 1: Quantitative results achieved by different methods and different design choices of *DeOccNet*. Note that, for  $l_1$ -error, smaller scores indicate better performance, and for *PSNR* and *SSIM*, higher scores indicate better performance.

	CD	Syn01	Syn02	Syn03	Syn04	Average
$l_1$ -error (Pei <i>et al.</i> [11])	0.233	0.188	0.240	0.204	<b>0.156</b>	0.204
$l_1$ -error (Liu <i>et al.</i> [9])	0.196	0.264	0.180	<b>0.078</b>	0.187	0.181
$l_1$ -error (ours_noASPP)	0.296	0.200	0.271	0.220	0.339	0.265
$l_1$ -error (ours_3skips)	0.222	0.202	0.346	0.319	0.207	0.259
$l_1$ -error (ours)	<b>0.185</b>	<b>0.138</b>	<b>0.165</b>	0.163	0.242	<b>0.178</b>
<i>PSNR</i> (Pei <i>et al.</i> [11])	16.75	19.78	18.10	19.60	<b>21.57</b>	19.16
<i>PSNR</i> (Liu <i>et al.</i> [9])	15.95	18.19	18.99	<b>25.58</b>	20.44	19.83
<i>PSNR</i> (ours_noASPP)	18.80	20.37	17.69	20.78	17.79	19.09
<i>PSNR</i> (ours_3skips)	19.95	20.41	16.25	17.74	21.20	19.11
<i>PSNR</i> (ours)	<b>21.27</b>	<b>24.68</b>	<b>21.74</b>	23.98	20.56	<b>22.45</b>
<i>SSIM</i> (Pei <i>et al.</i> [11])	0.508	0.636	0.568	0.656	0.569	0.587
<i>SSIM</i> (Liu <i>et al.</i> [9])	0.647	0.595	0.682	0.848	0.485	0.651
<i>SSIM</i> (ours_noASPP)	0.625	0.586	0.617	0.809	0.523	0.632
<i>SSIM</i> (ours_3skips)	0.621	0.613	0.560	0.742	0.530	0.613
<i>SSIM</i> (ours)	<b>0.694</b>	<b>0.699</b>	<b>0.734</b>	<b>0.858</b>	<b>0.650</b>	<b>0.727</b>

in average *SSIM*. That is because, the residual ASPP module can help the network to have a large receptive field to cover foreground occlusions of various shapes and scales.

Then, we investigated the benefits introduced by skip connections. Note that, during the training process, we discovered that the network cannot achieve a reasonable convergence without any skip connection. Therefore, we only removed the outermost skip connection in the ablation study, and infer the contribution of skip layers from experimental results. As shown in Table 1, the outermost skip layer introduces 3.34 dB improvement in average *PSNR* and 1.14 improvement in average *SSIM*. That is because, the high-frequency details are tend to be lost by strided convolutions, and it is difficult to recover these details from low-resolution features. Therefore, skip connections are necessary to provide a short path for low-level and high-frequency information.

#### 4.4. Further discussion

It is worth noting that LFs generated by our *Mask Embedding* approach were used as the only training data of our *DeOccNet*, and there is no intersection between the training and test scenes. In fact, the styles of occlusion and background between the training and test scenes differ significantly. That is, all occlusions in our training data are fence-like and front-parallel masks. Although we ran the *Mask Embedding* approach three times to simulate occlusions at multiple depths, we cannot generate occlusions within a continuous range of depths as in the CD and Bike scenes. However, as shown in Figs. 6(g) and 7(e), our network can generalize well to these cases to handle slant occlusions. That is because, both training and test scenes share the same LF structures. Specifically, occlusions have positive disparity values while backgrounds have negative disparity values.

To investigate the intrinsic mechanism of our network in dealing with foreground occlusions, we stacked 75 identi-

cal center-view SAIs of the CD scene, and fed them into our *DeOccNet*. It can be seen from Fig. 6(f) that our *DeOccNet* network cannot work with replicated inputs with identical information. That is, the mechanism of our *DeOccNet* is significantly different from that of the single image inpainting network [9]. Specifically, rather than using spatial information within one perspective in [9], our network uses complementary information from different viewpoints. It introduces significant performance improvements in scenes with heavy but pierced occlusions (*e.g.*, basket in scene Bike01) because useful information can be introduced by different perspectives. However, only focusing on complementary information from different views makes our network perform unsatisfactorily on scenes with solid block regions (*e.g.*, the right bottom corner in scene Bike02) since background is occluded in all perspectives and spatial information is need for LF-DeOcc. In the future, we will incorporate perspective information with neighborhood prior for LF-DeOcc, which is likely to introduce further performance improvement.

## 5. Conclusion

In this paper, we propose *DeOccNet*, the first deep learning based method for LF-DeOcc. We embed masks into existing LFs to generate a large training dataset. Experiments on both synthetic and real-world scenes show that, our *DeOccNet* can automatically remove foreground occlusions through disparity discrepancy, and achieve superior performance as compared to existing methods.

## 6. Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (Nos. 61401474, 61602499, 61972435), and Fundamental Research Funds for the Central Universities (No. 18lgzd06).

## References

- [1] A. Alperovich, O. Johannsen, and B. Goldluecke. Intrinsic light field decomposition and disparity estimation with deep encoder-decoder network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2165–2169. IEEE, 2018.
- [2] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke. Light field intrinsics with a deep encoder-decoder network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9145–9154, 2018.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.
- [4] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision (ACCV)*, pages 19–34. Springer, 2016.
- [5] N. Joshi, S. Avidan, W. Matusik, and D. J. Kriegman. Synthetic aperture tracking: tracking through occlusions. In *International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.
- [6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2014.
- [7] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich, and R. Raskar. Polarization fields: dynamic light field display using multi-layer lcds. In *ACM Transactions on Graphics*, volume 30, page 186. ACM, 2011.
- [8] X. Lin, J. Wu, G. Zheng, and Q. Dai. Camera array based light field microscopy. *Biomedical optics express*, 6(9):3179–3189, 2015.
- [9] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.
- [10] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016.
- [11] Z. Pei, X. Chen, and Y.-H. Yang. All-in-focus synthetic aperture imaging using image matting. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(2):288–301, 2018.
- [12] Z. Pei, Y. Zhang, X. Chen, and Y.-H. Yang. Synthetic aperture imaging using pixel labeling via energy minimization. *Pattern Recognition*, 46(1):174–187, 2013.
- [13] Z. Pei, Y. Zhang, T. Yang, X. Zhang, and Y.-H. Yang. A novel multi-object detection method in complex scene using synthetic aperture imaging. *Pattern Recognition*, 45(4):1637–1658, 2012.
- [14] J. Peng, Z. Xiong, D. Liu, and X. Chen. Unsupervised depth estimation from light field using a convolutional neural network. In *2018 International Conference on 3D Vision (3DV)*, pages 295–303. IEEE, 2018.
- [15] C. Shin, H.-G. Jeon, Y. Yoon, I. So Kweon, and S. Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4748–4757, 2018.
- [16] V. Vaish and A. Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7), 2008.
- [17] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2331–2338. IEEE, 2006.
- [18] V. Vaish, B. Wilburn, N. Joshi, and M. Levoy. Using plane+ parallax for calibrating dense camera arrays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I. IEEE, 2004.
- [19] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar. Picam: An ultra-thin high performance monolithic camera array. *ACM Transactions on Graphics*, 32(6):166, 2013.
- [20] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An, and Y. Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [21] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan. End-to-end view synthesis for light field imaging with pseudo 4d-cnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–348, 2018.
- [22] 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*, 27(9):4274–4286, 2018.
- [23] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo. Spatial-angular interaction for light field image super-resolution. *arXiv preprint arXiv:1912.07849*, 2019.
- [24] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An. Selective light field refocusing for camera arrays using bokeh rendering and superresolution. *IEEE Signal Processing Letters*, 26(1):204–208, 2018.
- [25] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Vision, Modelling and Visualization (VMV)*, volume 13, pages 225–226. Citeseer, 2013.
- [26] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz. High-speed videography using a dense camera array. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–II. IEEE, 2004.
- [27] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. In *ACM Transactions on Graphics*, volume 24, pages 765–776. ACM, 2005.
- [28] G. Wu, Y. Liu, Q. Dai, and T. Chai. Learning sheared epipolar structure for light field reconstruction. *IEEE Transactions on Image Processing*, 2019.

- [29] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai. Light field reconstruction using convolutional network on epi and extended applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
- [30] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017.
- [31] Z. Xiao, L. Si, and G. Zhou. Seeing beyond foreground occlusion: a joint framework for sap-based scene depth and appearance reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):979–991, 2017.
- [32] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6721–6729, 2017.
- [33] T. Yang, Y. Zhang, X. Tong, X. Zhang, and R. Yu. Continuously tracking and see-through occlusion based on a new hybrid synthetic aperture imaging model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3409–3416. IEEE, 2011.
- [34] T. Yang, Y. Zhang, X. Tong, X. Zhang, and R. Yu. A new hybrid synthetic aperture imaging model for tracking and seeing people through occlusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(9):1461–1475, 2013.
- [35] T. Yang, Y. Zhang, J. Yu, J. Li, W. Ma, X. Tong, R. Yu, and L. Ran. All-in-focus synthetic aperture imaging. In *European Conference on Computer Vision (ECCV)*, pages 1–15. Springer, 2014.
- [36] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2019.
- [37] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 24–32, 2015.
- [38] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5505–5514, 2018.
- [39] S. Zhang, Y. Lin, and H. Sheng. Residual networks for light field image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11046–11055, 2019.
- [40] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 2000.