Removing Foreground Occlusions in Light Field Using Micro-lens Dynamic Filter

Shuo Zhang[1,2,3] , **Zeqi Shen**[1,2] and **Youfang Lin**[1,2,3*]

[1]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing, China
[3]CAAC Key Laboratory of Intelligent Passenger Service of Civil Aviation, Beijing, China
{zhangshuo,shenzeqi,yflin}@bjtu.edu.cn

Abstract

Foreground occlusion removal task aims to automatically detect and remove foreground occlusions and recover background objects. Since for Light Fields (LFs), background objects occluded in some views may be seen in other views, the foreground occlusion removal task for LFs is easy to achieve. In this paper, we propose a learning-based method combining 'seeking' and 'generating' to recover occluded background. Specifically, the micro-lens dynamic filters are proposed to 'seek' occluded background points in shifted micro-lens images and remove occlusions using angular information. The shifted images are then combined to further 'generate' background regions to supplement more background details using spatial information. By fully exploring the angular and spatial information in LFs, the dense and complex occlusions can be easily removed. Quantitative and qualitative experimental results show that our method outperforms other state-of-the-art methods by a large margin.

1 Introduction

Removing the dense and complex occlusions is beneficial for many high-level computer vision applications, such as object detection, recognition and tracking, since the accuracy may be highly improved if the background objects are recovered.However, it is difficult to automatically detect and remove dense and irregular occlusions in one single image. On one hand, the depth information is unknown and foreground occlusions are hard to identify. On the other hand, no reliable information about the occluded background is provided. By contrast, the recently developed Light Fields (LFs) contain Sub-Aperture Images (SAIs) that capture scenes in different directions, in which background points that are occluded in the central SAI may be visible in other SAIs. Therefore, LFs show great potential in the foreground occlusion removal task.

In recent years, with the rapid development of deep learning, single image inpainting methods [Liu *et al.*, 2018; Xie *et al.*, 2019; Li *et al.*, 2020] have made significant breakthroughs. By learning plausible structures from a large amount of data, these inpainting methods try to combine surrounding semantical information to 'generate' labeled missing regions. However, for the occlusion removal task in LFs, it is more important to find and further 'seek' occluded regions in different SAIs. Since LFs contain structure information among different SAIs, the foreground occlusions can be identified accordingly and the occluded background is also possible to find in other SAIs.

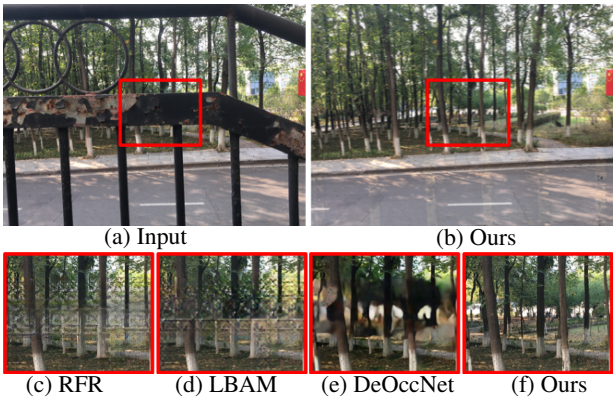

Figure 1: A real-world occlusion removal example. (a) The input central view with occlusions. (b) Our occlusion removal result. The enlarged results of (c) RFR (d) LBAM (e) DeOccNet (f) Ours.

Recently, [Wang *et al.*, 2020] developed an end-to-end learning-based framework to calculate occlusion-free central SAIs using LFs as input. However, the related results show blurry background and aliasing effects. The main problems for LF occlusion removal are: 1) Although lots of LF depth estimation methods [Tsai *et al.*, 2020] have been proposed, it is still difficult and costly to find accurate depth information, especially for occluded points. 2) How to combine all SAIs in grid-like angular sampling to completely remove occlusions. 3) It is difficult to design one network that has large receptive fields to find all occluded points in other SAIs.

In this paper, we propose a novel end-to-end learning based framework to solve the above problems. The angular and spatial information in LFs is fully considered to simultaneously 'seek' and 'generate' occluded background regions. The LFs are first shifted to extract effective features so that differ-

*corresponding author: yflin@bjtu.edu.cn

ent disparities are possible to estimate. We then develop the micro-lens dynamic filters for the shifted LFs to find correct background points using angular information. The features of shifted LFs are further integrated to learn surrounding spatial information to supplement more details. The occlusion-free central SAIs are finally obtained once the abundant information in LFs is fully explored.

One typical example is shown in Fig 1, in which our method is able to recover the complex scenes without any prior occlusion information. Our main contributions are summarized as follows:

- We design a novel end-to-end learning-based framework based on the shift operation for occlusion removal task, which fully explores the spatial and angular information of LFs and implicitly learns the disparity information.
- We develop the micro-lens dynamic filter, which is effective to find occluded pixels from other SAIs.
- Experiments show that our method can automatically identify and remove occlusions in specified disparities. Comparisons on both synthetic dataset and real-world dataset show that our method achieves superior performances over the state-of-the-art methods.

2 Related Work

In this section, we briefly overview the related single image inpainting methods and LF occlusion removal methods.

Single Image Inpainting. If foreground occlusions are accurately masked, the widely developed single image inpainting methods can be applied to recover the background regions. [Liu *et al.*, 2018] proposed Partial Convolution by assigning the missing and non-missing pixels in the convolution operation with different weights so that the model focused more on restoring the missing areas. [Xie *et al.*, 2019] further extended the Partial Convolution from the encoding layers to the decoding layers, in order to further improve the visual effect. [Li *et al.*, 2020] creatively transformed the process of inpainting method into an iterative structure. However, the training of these inpainting methods is quite difficult since the models need a large amount of dataset to learn the plausible semantics. Moreover, when occluded background regions are complex, only using the learned semantic information cannot achieve reliable results.

Light Field Occlusions Removal. In order to remove occlusions in LFs, some traditional methods have been proposed. [Vaish *et al.*, 2004] proposed a refocusing method to warp LFs in spatial domain by a special value and then average all SAIs to remove occlusions. Then, the median cost and entropy cost are further proposed in [Vaish *et al.*, 2006] to improve their method. Because these methods can't distinguish the light from occlusion and background very well, their results are quite blurred. [Pei *et al.*, 2013] used pixel labeling via energy minimization to solve this image blur problem. However, their method cannot get the image clear in all depth levels. Then they came up with an image matting approach in [Pei *et al.*, 2018] to get all-in-focus images. [Yang *et al.*, 2014] presented a depth free all-in-focus method to remove occlusions. Specifically, they divided the scene into different visible layers to directly deal with layer-wise occlusions. [Xiao *et al.*, 2017] used the clustering method to distinguish occlusion layers and background layers. [Pendu *et al.*, 2018] noted that there is redundancy in different views so that they removed occlusions via constructing the low-rank matrix. However, all these methods used handcrafted features and cannot achieve satisfactory results in complex background areas. [Wang *et al.*, 2020] recently first proposed to remove foreground occlusions using LFs throughout an end-to-end network. They constructed their network with the U-shape structure and used Atrous Spatial Pyramid Pooling (ASPP) [Wang *et al.*, 2019] to extract features from input SAIs. They also proposed to use LFs and several occlusion images to construct the training dataset and prove that the trained model can handle both synthetic and real-world occluded scenes. However, since all SAIs are simply concatenated as the input, the angular information cannot be fully extracted and the related regions cannot be well recovered.

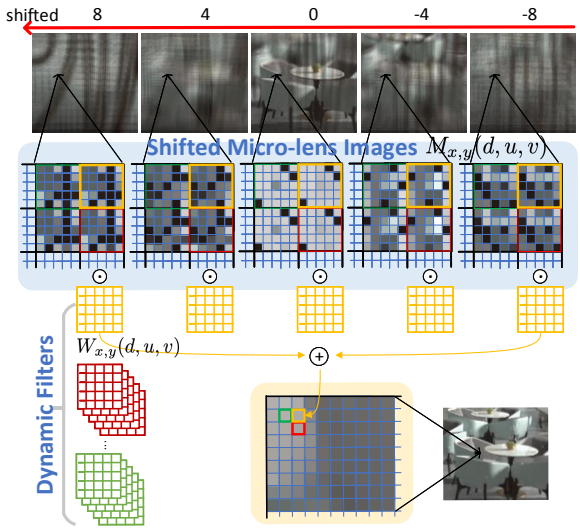

Figure 2: The shifted micro-lens images and the proposed Micro-lens Dynamic Filter.

3 Motivation

In this paper, we use $L(u,v,x,y) \in \mathbb{R}^{U\times V\times X\times Y}$ to represent 4D LFs, where (u,v) and (x,y) are the angular and spatial coordinates, respectively. By fixing angular coordinates (u,v), SAI $I_{u,v}(x,y)$ is obtained. Similarly, by fixing spatial coordinates (x,y), the micro-lens image is represented as $M_{x,y}(u,v)$. LFs occlusion removal task aims to get a clean center SAI $I^{clean}_{u^c,v^c}(x,y)$ from $L(u,v,x,y)$, where u^c, v^c represent the center angular coordinates.

In order to find occluded pixels in other views, we follow the shift strategy in LF depth estimation [Tsai *et al.*, 2020] to construct shifted LFs. Specifically, the shifted LFs $L^d(u,v,x,y)$ with different disparities d are calculated:

$$L^d(u,v,x,y) = L(u,v,x+(u-u^c)\cdot d, y+(v-v^c)\cdot d). \quad (1)$$

The example shifted LFs with different d, displaying with micro-lens images, are shown in Fig 2. Following the LF dig-

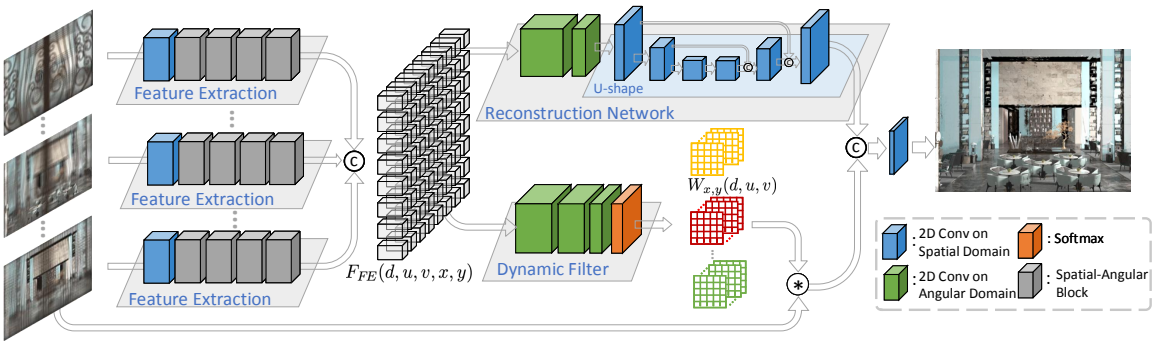

Figure 3: The network architecture of our model.

ital refocusing in [Ng *et al.*, 2005], the refocused images can be obtained by directly summing up each micro-lens image as one pixel. If we shift LFs with one disparity d^*, the objects in disparity d^* are in focus and the pixels in the related micro-lens image $M^{d^*}_{x,y}(u,v)$ show consistent color. However, in Fig 2, the background regions are occluded by dense and fine objects. The micro-lens image $M^{d_{back}}_{x,y}(u,v)$, which is shifted using background disparity d_{back}, contains not only background colors but also occlusion colors. If we directly average the pixels in the micro-lens image, the result refocused image shows the background with blurry occlusions. In our method, we propose to delete occlusion points and keep background points in the shifted micro-lens images so that images without occlusion can be recovered.

Specifically, we first construct D possible shifted LFs $L^d(u,v,x,y)$. As Fig 2, for one specific point (x,y) in central SAIs, D shifted micro-lens images $M^d_{x,y}(u,v)$ are available. In one specific shifted micro-lens image, the occluded background pixels can be found and further extracted. Inspired by the dynamic upsampling filters in video super-resolution [Jo *et al.*, 2018], we design novel micro-lens dynamic filters for the LF occlusion removal task. The micro-lens dynamic filter measures the possibility of each pixel in shifted micro-lens images belonging to background points and is calculated for each point in the central SAI. Then, by summing up the weighted micro-lens images, the pixel is recovered with the correct background color.

4 The Proposed Approach

The overview architecture of our proposed method is shown in Fig 3. The shifted LFs are first fed into the Feature Extraction (FE) module H_{FE} , which explores the latent relationship between angular domain and spatial domain. The extracted features are then separately fed into a novel Micro-lens Dynamic Filter (MDF) module H_{MDF} and U-shape Reconstructed (UR) module H_{UR}. H_{MDF} is designed to calculate the micro-lens filters so that occluded pixels are effectively found using angular information. H_{UR} aims to reconstruct images with more details based on spatial information.

4.1 Feature Extraction Module

For each shifted LF, we construct one branch to extract both spatial and angular features. Specifically, we design a spatial-angular block, which contains 4 parallel branches, 2 for 2D convolution with dilation =1, 2 on spatial domain and other 2 for angular domain. The spatial and angular features are then concatenated in the end of each block. In our method, 2D spatial convolution and 4 spatial-angular blocks are built in order in each branch H_{FE} to extract effectively features:

$$F^d_{FE}(u,v,x,y) = H_{FE}(L^d(u,v,x,y)). \tag{2}$$

In order to reduce the complexity of the network, the weights are shared in different feature extraction branches.

4.2 Micro-lens Dynamic Filter Module

After feature extraction, we combine all features of different shifted LFs as $F_{FE}(d,u,v,x,y)$. We then design the MDF module H_{MDF} to calculate the micro-lens filters $W_{x,y}(d,u,v) \in \mathbb{R}^{D\times U\times V}$. H_{MDF} contains several angular convolution in order to fully explore the potential relationship between pixels in shifted micro-lens images. A softmax layer is then implemented on the (d,u,v) dimension to normalize $W_{x,y}$:

$$W_{x,y}(d,u,v) = H_{MDF}(F_{FE}(d,u,v,x,y)). \tag{3}$$

$W_{x,y}(d,u,v)$ represents the possibilities of pixels in shifted micro-lens images belonging to the occluded background points. The higher the value, the larger possibility it has. We calculate $W_{x,y}(d,u,v)$ for each point (x,y) in the central SAI.

As analyzed before, for each point (x,y) in the central SAI, D shifted micro-lens images are available, which can be expressed as $M_{x,y}(d,u,v)$. By multiplying the shifted micro-lens images $M_{x,y}(d,u,v)$ with the corresponding micro-lens filters $W_{x,y}(d,u,v)$, the pixels belonging to occluded background are extracted and foreground occlusions are deleted:

$$F_{MDF}(x,y) = \sum_d \sum_u \sum_v M_{x,y}(d,u,v) \odot W_{x,y}(d,u,v). \tag{4}$$

F_{MDF} indicates the calculated occlusion removal images using local angular information. $\odot$ denotes the dot product.

4.3 U-shape Reconstructed Module

In MDF module, angular information from the same point of other SAIs is extracted in order to restore occluded pixels. We further introduce the UR module to integrate spatial information from adjacent regions around the occluded points

Method	resize	disparity	Evaluation using Occlusion Masks				Evaluation w/o Masks	
			RFR	LBAM	DeOccNet*	Ours*	DeOccNet	Ours
4-synLFs	1	(-12, 9)	25.73/0.8150	25.00/0.8099	24.56/0.8103	**27.09/0.8664**	21.42/0.6788	24.98/0.8019
4-synLFs	2/3	(-8, 6)	24.03/0.7560	23.57/0.7499	25.26/0.8219	**28.50/0.8691**	22.32/0.7256	26.43/0.8104
4-synLFs	1/2	(-6, 4.5)	24.19/0.7720	23.39/0.7663	25.06/0.8124	**26.96/0.8233**	22.31/0.7270	24.96/0.7561
9-synLFs	1	(-5, 5)	27.94/0.8817	27.85/0.8764	20.92/0.7382	**28.39/0.9137**	17.73/0.6469	25.70/0.8443

Table 1: Quantitative results (PSNR/SSIM) of different methods on synthetic LFs.

for more details. Specifically, the 2D convolution is first implemented on angular domain without padding in order to reduce the angular dimension. The widely used U-Shape structures [Wang *et al.*, 2020; Ronneberger *et al.*, 2015] is then introduced to enlarge the receptive field and extract multi-scale spatial features. The U-shape structure includes encoding layers, decoding layers and skip-connections is constructed as shown in Fig 3. In order to effectively extract spatial features in SAIs, the 2D convolution is performed on the spatial domain. The output $F_{UR}(x, y)$ is then computed as:

$$F_{UR}(x, y) = H_{UR}(F_{FE}(d, u, v, x, y)). \quad (5)$$

Finally, we fuse $F_{MDF}(x, y)$ and $F_{UR}(x, y)$ using 2D spatial convolution to get the final output $I^{clean}_{u^c,v^c}(x, y)$.

4.4 Loss Functions

Besides the widely used L1 loss $loss_{L1}$, we introduce one mask loss $loss_{mask}$ to make the model focus more on occlusion areas, which is calculated as:

$$loss_{mask} = \left\| I^{clean}_{u^c,v^c} \odot mask - I^{gt}_{u^c,v^c} \odot mask \right\|_1, \quad (6)$$

where $I^{gt}_{u^c,v^c}$ denotes the ground truth occlusion-free image. Since the occlusion mask is available in the training dataset, the $mask$ can be accordingly set, *i.e* 1 for occluded areas and 0 for other regions. We also introduce the perceptual loss $loss_{vgg}$ in [Wei *et al.*, 2019] based on the VGG-19 Network [Simonyan and Zisserman, 2014] to achieve better visual effects. The final loss function is calculated as:

$$loss_{total} = loss_{L1} + \lambda_1 loss_{mask} + \lambda_2 loss_{vgg}, \quad (7)$$

where λ_1 and λ_2 denote loss weights and are set as 0.5 and 0.1 in the following experiments, respectively.

5 Experiment

We first introduce the datasets and training details. The proposed method is then compared with other methods. Finally, the ablation study is taken for further evaluation.

Datasets. We follow [Wang *et al.*, 2020] to use 61 LF images from [Honauer *et al.*, 2016; Lanman *et al.*, 2011; Wanner *et al.*, 2013; Vaish and Adams, 2008] and 80 occlusion images to synthesize the training dataset. Since only 4 synthetic LFs with ground truth (4-synLFs) are used in [Wang *et al.*, 2020] for testing, we further synthesize another LF dataset (9-synLFs) using 3D MAX and the public 3D models in [Company, 2013]. The new synthetic dataset contains 9 LFs with complex backgrounds and different foreground occlusions. We further resize the 4-synLFs with different scale factors so that images with different disparities can be further evaluated. In our experiment, the original 4-synLFs are used as the validation set and the model is finally tested on the resized 4-synLFs [Wang *et al.*, 2020], new 9-synLFs and one real-world dataset [Wang *et al.*, 2020].

Disparity Setting. Following the training strategy in [Wang *et al.*, 2020], our model is also designed to remove occlusions that have positive disparities and keep background regions that have negative disparities. In this way, multiple occlusions in different disparities can be removed step by step by shifting LFs. In order to find micro-lens images in background disparities, we choose totally D shifted values d, which is evenly spaced in $d \in \{d_{min}, \cdots, 0\}$. In our experiments, we choose d_{min}=-9 and D = 10, which is able to handle the above dataset with different disparities. Experiments with different parameters are further analyzed in Sec 5.2.

Training Details. We crop the training images into $5 \times 5 \times 64 \times 64$ pixel patches. The RGB color is put into the channel dimension in the input and output. Our model has 2.7M parameters, which is 14 times fewer than DeOccNet, 24 times fewer than LBAM, 11 times fewer than RFR. We perform upsampling, flipping, rotation, color conversion for data augmentation. The proposed model is optimized using the Adam [Kingma and Ba, 2014] algorithm with a batch size of 5. The initial learning rate is set as $1e-3$ in the first 150 epochs and is then divided by 10 every 50 epochs. We implement the model with pyTorch framework and the training process roughly takes 3 days with an Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz with a Titan XP GPU. Occlusions are removed in one LF with around 10s using our model.

5.1 Comparison with State-of-the-Arts

The only learning-based LF occlusion removal network DeOccNet [Wang *et al.*, 2020] is compared. The state-of-the-art inpainting methods LBAM [Xie *et al.*, 2019] and RFR [Li *et al.*, 2020] are also compared by providing the central SAI and the ground truth occlusion mask. In order to learn plausible semantics information, the inpainting methods need a large amount of images (around 15000 images) for training. LBAM [Xie *et al.*, 2019] and RFR [Li *et al.*, 2020] cannot achieve comparable results if only our 61 training dataset is used for training. Therefore, we choose to carefully fine-tune their pre-trained models on our training dataset.

Synthetic Scenes

We calculate the PSNR and SSIM metrics of the recovered RGB central view image for quantitative comparison. For the image inpainting methods, only occlusion regions are recovered and the other regions are directly copied from the original images. Therefore, we also replace the non-occluded ar-

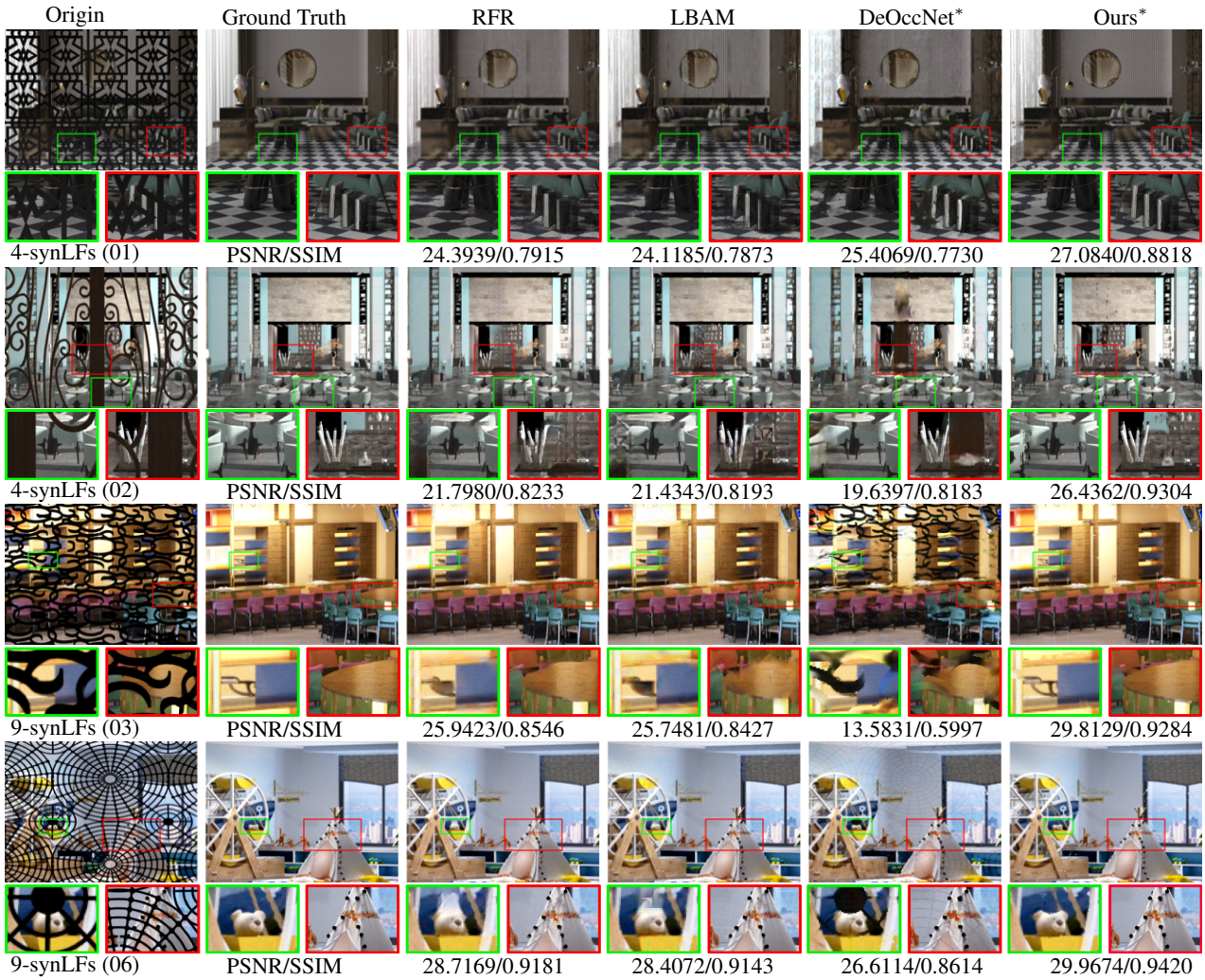

Figure 4: Occlusion removal results on synthetic datasets. The ground truth occlusion masks are provided. Our method integrates information from the angular domain and is able to find the occluded background and recover accurate complex textures even with the dense occlusions.

eas using the original images in DeOccNet and ours, labelled with $*$, to give a further evaluation on how these methods perform on occluded areas.

Table 1 and Fig 4 show the quantitative and qualitative results of 4-synLFs and 9-synLFs. Using the occlusion masks, the related results focus on whether the occluded background regions are accurately recovered. Note that by resizing LFs in spatial domain, the disparities also change. Compared with a small disparity range, removing occlusions in a large disparity range is more difficult since we need to find more possible regions in LFs. From Table 1, we can find that the average PSNR and SSIM of ours* are far better than other methods on different datasets. Qualitative results in Fig 4 also show that our method can recover the complex background textures which are fully occluded in the central SAI. By comparison, other methods produce significant artifacts and wrong textures. The inpainting methods LBAM [Xie *et al.*, 2019] and RFR [Li *et al.*, 2020] generate background regions according to the plausible semantics. It is difficult for them to recover complex backgrounds. DeOccNet [Wang *et al.*, 2020] is also designed to use the abundant information of LFs, but their network cannot find the occluded background since the angular information is not well explored. Moreover, the performance of DeOccNet decreases dramatically in the 9-synLFs, which indicates that their generalization ability is insufficient.

Fig 5 and Table 1 further show the comparison of our method and DeOccNet, in which no occlusion masks are provided. Different with the inpainting methods, foreground occlusions are automatically recognized and removed through the end-to-end network. Numerical results in Table 1 indicate that our results achieve more than 3 dB higher PSNR than DeOccNet. Fig 5 further shows our method produces less artifacts and less color distortion than DeOccNet.

Real-world Scenes

Fig 1 and Fig 6 show the occlusion removal results using the real-world dataset. In order to compare with the inpainting methods, we manually label the occlusion masks for them. By contrast, the results of DeOccNet and ours are directly

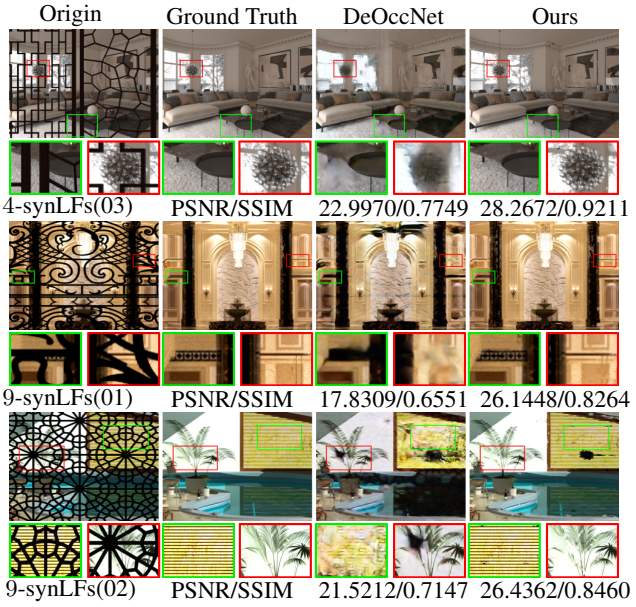

Figure 5: Qualitative results on synthetic datasets without occlusion masks. The occlusion-free images are calculated in an end-to-end manner. Our method is able to automatically detect and remove the occlusions and recover the rich background textures.

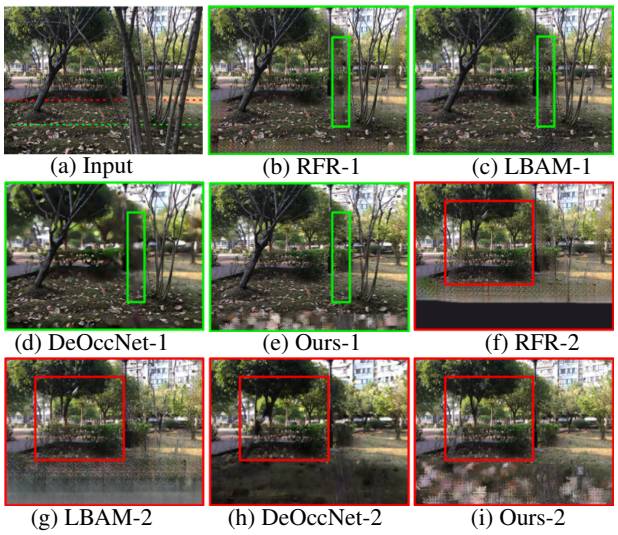

Figure 6: Qualitative results on the real-word image with different occlusion removal tasks. The green and red line indicate the 0 disparity of the input LFs and the objects before the line are needed to remove. Our method outperforms others in different tasks.

produced without occlusion masks. As in Fig 1, our method is the only one that is able to recover the occluded trees with rich textures. In Fig 6, we further compare different tasks in which occlusions in different disparities are removed. As mentioned before, through the shift operation, positions of zero disparity can be adjusted and occlusions in specific disparities are removed. As shown, in this complex example, by shifting the zero disparity to the 'red' or 'green' lines, the objects before the line are successfully removed in our method.

Method		PSNR	SSIM
Parameter	d_{min} = -9, D = 5	23.4753	0.7675
	d_{min} = -4, D = 5	21.5868	0.7032
Structure	w/o H_{UR}	23.8333	0.7949
	w/o H_{MDF}	21.3039	0.5959
	$W \in \mathbb{R}^{1\times U\times V}$ in Equ. 3	24.1922	0.7782
	$W \in \mathbb{R}^{D\times 1\times 1}$ in Equ. 3	20.7103	0.5701
Loss	$loss_{L1}$	23.4878	0.7870
	$loss_{L1} + loss_{vgg}$	23.8270	0.7582
	$loss_{L1} + loss_{mask}$	23.9598	0.7944
Our method		**24.9770**	**0.8019**

Table 2: Ablation studies (PSNR/SSIM) on 4-synLFs.

This means our method successfully learns the disparity information and is able to remove occlusions that we do not need. Although DeOccNet can also recognize disparity information, the results are far less accurate than our method and have blurry background textures. Other inpainting methods cannot recover accurate details in this textured background. Note that in this example, the bottom grass regions are close to the camera and are identified as the foreground that should be removed. However, since no background regions behind the grass can be recovered, all the methods choose to blur the foreground grass.

5.2 Ablation Study

We conduct several ablation studies on the 4-synLFs with different hyper-parameters, structures and loss functions as in Table 2. For fair comparisons, the number of parameters of all these models is kept almost the same. By comparing different d_{min} and D, we find that for the 4-synLFs with $(-12, 9)$ disparities, larger disparity range and more shifted numbers provide more accurate results. We then compare our model with different structures, in which the proposed H_{MDF} is deleted or modified. As shown, without H_{MDF}, only H_{UR} cannot find correct occluded points. Moreover, using dynamic filters whose size in $\mathbb{R}^{1\times U\times V}$ or $\mathbb{R}^{D\times 1\times 1}$ cannot fully use the abundant information in different shifted LFs or in different SAIs, respectively. By contrast, without H_{UR}, the spatial information cannot be fully explored and the performance also decreases. Finally, combining the proposed mask loss and the commonly used perceptual loss, the results can be further improved by 1.4 dB in PSNR.

6 Conclusion

In this paper, we proposed a novel end-to-end learning-based method for LF occlusion removal, in which depth information is implicitly learned. The micro-lens dynamic filter is designed to remove foreground occlusions and recover background details. The angular and spatial information in LFs is fully extracted and integrated in our network. Experiments on synthetic and real-world datasets prove that our method achieves obviously higher performances than others.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61902018).

References

[Company, 2013] Shenzhen TePeng Network Limited Company. Indoor scene models. https://www.3d66.com, 2013. Accessed: 2021-05-01.

[Honauer *et al.*, 2016] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, volume 10113, pages 19–34, 2016.

[Jo *et al.*, 2018] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3224–3232, 2018.

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

[Lanman *et al.*, 2011] Douglas Lanman, Gordon Wetzstein, Matthew Hirsch, Wolfgang Heidrich, and Ramesh Raskar. Polarization fields: dynamic light field display using multi-layer lcds. *ACM Transactions on Graphics*, 30(6):186, 2011.

[Li *et al.*, 2020] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7757–7765, 2020.

[Liu *et al.*, 2018] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *European Conference on Computer Vision*, volume 11215, pages 89–105, 2018.

[Ng *et al.*, 2005] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenopic camera. *Technical Report CTSR*, 2005.

[Pei *et al.*, 2013] Zhao Pei, Yanning Zhang, Xida Chen, and Yee-Hong Yang. Synthetic aperture imaging using pixel labeling via energy minimization. *Pattern Recognition*, 46(1):174–187, 2013.

[Pei *et al.*, 2018] Zhao Pei, Xida Chen, and Yee-Hong Yang. All-in-focus synthetic aperture imaging using image matting. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(2):288–301, 2018.

[Pendu *et al.*, 2018] Mikael Le Pendu, Xiaoran Jiang, and Christine Guillemot. Light field inpainting propagation via low rank matrix completion. *IEEE Transactions on Image Processing*, 27(4):1981–1993, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *In Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Tsai *et al.*, 2020] Yu-Ju Tsai, Yu-Lun Liu, Ming Ouhyoung, and Yung-Yu Chuang. Attention-based view selection networks for light-field disparity estimation. In *Association for the Advance of Artificial Intelligence*, pages 12095–12103, 2020.

[Vaish and Adams, 2008] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. http://lightfield.stanford.edu/index.html, 2008. Accessed: 2021-05-01.

[Vaish *et al.*, 2004] Vaibhav Vaish, Bennett Wilburn, Neel Joshi, and Marc Levoy. Using plane + parallax for calibrating dense camera arrays. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2–9, 2004.

[Vaish *et al.*, 2006] Vaibhav Vaish, Marc Levoy, Richard Szeliski, C. Lawrence Zitnick, and Sing Bing Kang. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2331–2338, 2006.

[Wang *et al.*, 2019] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jun-Gang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019.

[Wang *et al.*, 2020] Yingqian Wang, Tianhao Wu, Jun-Gang Yang, Longguang Wang, Wei An, and Yulan Guo. Deoccnet: Learning to see through foreground occlusions in light fields. In *IEEE Winter Conference on Applications of Computer Vision*, pages 118–127, 2020.

[Wanner *et al.*, 2013] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *International Symposium on Vision, Modeling and Visualization*, pages 225–226, 2013.

[Wei *et al.*, 2019] Kaixuan Wei, Jiaolong Yang, Ying Fu, David P. Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019.

[Xiao *et al.*, 2017] Zhaolin Xiao, Lipeng Si, and Guoqing 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.

[Xie *et al.*, 2019] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *IEEE International Conference on Computer Vision*, pages 8857–8866, 2019.

[Yang *et al.*, 2014] Tao Yang, Yanning Zhang, Jingyi Yu, Jing Li, Wenguang Ma, Xiaomin Tong, Rui Yu, and Lingyan Ran. All-in-focus synthetic aperture imaging. In *European Conference on Computer Vision*, volume 8694, pages 1–15, 2014.