I See-Through You: A Framework for Removing Foreground Occlusion in Both Sparse and Dense Light Field Images

Jiwan Hur*, Jae Young Lee*, Jaehyun Choi, and Junmo Kim

School of Electrical Engineering, KAIST, South Korea
{jiwan.hur, mcneato, chlwogus, junmo.kim}@kaist.ac.kr

Abstract

Light field (LF) camera captures rich information from a scene. Using the information, the LF de-occlusion (LF-DeOcc) task aims to reconstruct the occlusion-free center view image. Existing LF-DeOcc studies mainly focus on the sparsely sampled (sparse) LF images where most of the occluded regions are visible in other views due to the large disparity. In this paper, we expand LF-DeOcc in more challenging datasets, densely sampled (dense) LF images, which are taken by a micro-lens-based portable LF camera. Due to the small disparity ranges of dense LF images, most of the background regions are invisible in any view. To apply LF-DeOcc in both LF datasets, we propose a framework, ISTY, which is defined and divided into three roles: (1) extract LF features, (2) define the occlusion, and (3) inpaint occluded regions. By dividing the framework into three specialized components according to the roles, the development and analysis can be easier. Furthermore, an explainable intermediate representation, an occlusion mask, can be obtained in the proposed framework. The occlusion mask is useful for comprehensive analysis of the model and other applications by manipulating the mask. In experiments, qualitative and quantitative results show that the proposed framework outperforms state-of-the-art LF-DeOcc methods in both sparse and dense LF datasets.

1. Introduction

In various computer vision tasks such as image classification [7, 26], semantic segmentation [25, 1, 8], and object detection [21, 22, 23], the performance becomes unstable and degrades by the foreground objects which occlude the region of interest. To diminish such performance drop, the de-occlusion task aims to capture the foreground occlusion object in the image and fill the region with the backgrounds.

*Equal contribution.

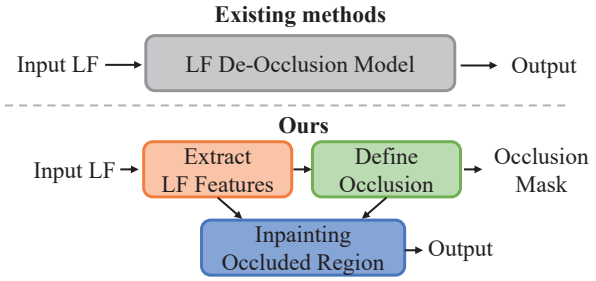

Figure 1. The framework of the existing LF-DeOcc methods (top) and the proposed method (bottom). While the existing methods consist of a single black-box model, the proposed method is composed of three separate components, generating the occlusion mask as an intermediate representation of the framework which is useful for the analysis and other applications. Note that the proposed LF-DeOcc framework also works in an end-to-end manner.

Recently, light fields (LFs) are utilized in the de-occlusion task (LF-DeOcc) [30, 37]. As LFs can capture the scene in various views with different angular information, they can offer guidance of the background object, the key goal of the de-occlusion task. While the existing deep learning-based LF-DeOcc methods show reasonable performance, they mainly focus on the sparsely sampled (sparse) LFs, which are collected through a camera array and have large disparity ranges. In other words, most of the occluded region can be seen in the other views of sparse LFs. Contrarily, the densely sampled (dense) LFs are collected through a micro-lens-based *portable* LF camera [2] which has narrow disparity ranges consisting of scarce information about the backgrounds. In summary, while it is easy for the sparse LFs to obtain visible background information beyond the occlusion, it is hard for the dense LFs to obtain visible background information beyond the occlusion. However, from a practical perspective, whereas it is easy for the dense LFs to obtain various scenes by being portable and affordable, it is

relatively hard for the sparse LFs to collect various scenes. The detailed statistics of the dataset and different characteristics of both LFs are provided in supplementary materials.

In this paper, different from the existing LF-DeOcc methods, we considered two scenarios to make a model work on both sparse and dense LFs. First, if the occluded region is visible in the other views, the visible information should be used to fill the region. Second, if not, the context information around the occluded region should be used to fill the region. Although both scenarios are presented in both LFs, the first scenario dominates in the sparse LFs whereas the second scenario dominates in the dense LFs.

Since each scenario requires quite different solutions, we divide the framework and define the separate roles and functions in the proposed framework (Fig. 1). For the first scenario, likewise DeOccNet [30], a component for extracting the LF features from the sub-aperture images (SAIs) is presented in the proposed framework. For the second scenario, a component for inpainting a single image is modified with an additional component to define the occlusion mask in the proposed framework. Contrary to the existing methods which implicitly have those roles in their models, the proposed framework explicitly divides the roles and connects them. By explicitly dividing the roles, the proposed method not only shows better performance but also makes the development and analysis easier than the existing methods.

In addition, because the occlusion is explicitly represented in the proposed framework, it is flexible to define the occlusion, which helps not only prevent artifacts in the non-occluded regions but also remove occlusion in arbitrary depth planes while preserving the foreground objects of interest by manipulating the occlusion mask.

The contributions of the proposed LF-DeOcc framework are summarized as follows.

- We propose a framework, ISTY, which works on both sparse and dense LFs, achieving *state-of-the-art* performance in the majority of the settings.
- We modularized black-box framework into three separate components, making further development and analysis easier.
- Occlusion mask generator offers flexibility in defining the occlusion by explicitly giving the mask representation and enables additional applications.

2. Related work

2.1. Light Field De-Occlusion (LF-DeOcc)

By the digital refocusing algorithm [18], in the refocused image, the occluded regions are blurry but partially visible. Thus, the digital refocusing algorithm has also been utilized to see through the occlusion [28]. Vaish et al. [28] proposed a refocus method that re-parameterizes LF image by a specific value and average along with the angular dimension. In their method, even though foreground occlusion could be seen through, the images are highly blurred.

Recently, Wang et al. [30] proposed a deep learning-based end-to-end LF-DeOcc model (DeOccNet). DeOccNet reconstructs the occlusion-free center-view (CV) image with a deep encoder-decoder model and residual atrous spatial pyramid pooling (ResASPP) module from sparse LF images. They also propose a mask embedding approach to generate a training dataset which synthesizes the occlusion LF image using the mask image and occlusion-free LF image allowing the fully supervised end-to-end LF-DeOcc learning. However, DeOccNet generates blurry outputs and does not appropriately deal with occlusion with the large invisible region, making artifacts from occlusions. Zhang et al. [37] proposed a filter to extract features from the shifted lenslet images to seek background information to reconstruct the occluded regions. Although their method works well on the sparse LFs, the performance on the dense LF is not as good as that on the sparse LF because they strongly assume that background object is visible. Furthermore, using a set of shifted-lenslet images requires large memory and long pre-processing time.

The recently proposed deep learning-based LF-DeOcc methods focused on the sparse LFs, filling the occluded regions with the visible background information from the other views. Since it is more difficult to collect the sparse LF dataset than the dense LF dataset, it is reasonable to use the dense LFs, which can be easily collected, to train and apply a model with the advantage of a large number of data from practical perspectives. Thus, different from existing LF-DeOcc methods, we propose a framework that works on both sparse and dense LF images.

2.2. Single Image Inpainting

Single image inpainting in an RGB image. The goal of the single image inpainting is to recover the missing (masked) regions of a single image with realistic content. The rapid development of deep learning algorithms and the vast amount of single RGB image datasets [6, 38] makes it possible to fill the masked regions with a plausible structure without information beyond the mask [17, 16, 32].

Partial convolution (PConv) [17] helps to encode the context features while avoiding the artifacts from invalid pixels of the masked region through the masked and renormalized convolution. Based on PConv, Li et al. [16] introduced recurrent feature reasoning (RFR) to reconstruct the large continuous hole through recurrent inpainting the part of the image and average the generated feature group if they have no invalid pixels. Xie et al. [32] use learnable bidirectional attention maps (LBAM) to replace the PConv. They used attention not only in the encoder but also in the decoder so that the decoder can focus on filling the masked regions only. In addition, unlike PConv, LBAM allows soft

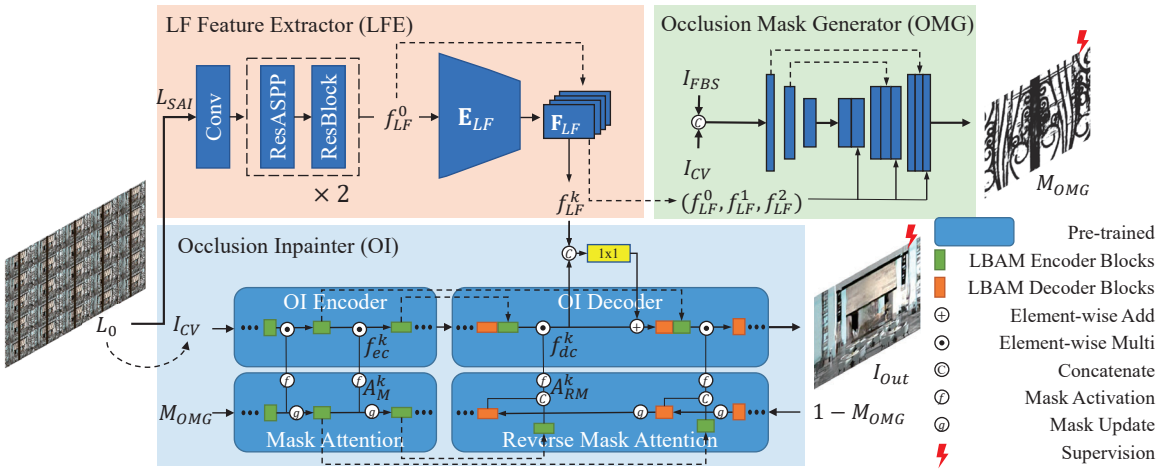

Figure 2. Illustration of the proposed model architecture. ⓕ and ⓖ denotes the mask activation and update function respectively, defined in LBAM [32]. The I_{CV} and M_{OMG} are supervised with the loss function defined in section 3.5 using I_{gt} and M_{gt}, respectively.

attention map and differentiable mask updates and activation function which gives more leeway for further improvements and stability to model in training. However, directly applying the single image inpainting method on LF-DeOcc is not feasible as the occlusion mask can not be defined in the single RGB image. In our framework, to utilize the inpainting method, we explicitly define the occlusion mask from the LFs and pass it as input of the inpainting method.

Single image inpainting in LFs. In the LF, the single image inpainting has been used to solve the LF completion which aims to fill the entire LF views with consistent information. Rather than directly filling the 4-D manifold [33, 3, 4], Zhang et al.[35] and Pendu et al.[10] used single image inpainting to the CV image and propagated it to the remaining views. However, since their main goal is to propagate the information of the inpainted CV image to the remaining views, they naively used the single image inpainting method for the CV image with a given inpainting mask.

2.3. Foreground Background Seperation (FBS)

The foreground-background separation (FBS) [11, 13, 15], which represents whether each pixel is classified into the foreground or background, could be regarded as a sub-representation of the depth map [14]. Using the optical phenomenon "flipping" [12], Lee and Park [11] estimated depth map from LF by accumulating the binary maps, which separated the foreground and background based on the focused (or zero disparity) plane. A single slice of FBS in form of the binary map is obtained by thresholding a score map, whose pixel values are bounded from -1 to 1, with zero. In this paper, we define the occlusion mask by refining the FBS in form of the score map, to avoid losing information.

3. Method

The proposed framework is composed of three components, each of which performs a different role, named LF feature extractor (LFE), occlusion mask generator (OMG), and occlusion inpainter (OI), respectively. In the following subsections, we first describe our overall architecture. Subsequently, we introduce each part of the model and the composition of our loss function. Lastly, we introduce an additional application possible in the proposed framework. The architecture of the proposed framework is shown in Fig. 2.

3.1. Overall Architecture

From the input LF $L_0 \in \mathbf{R}^{U\times V\times X\times Y\times 3}$, we extract various LF information in LFE, define the occlusion mask in OMG, and reconstruct the occlusion-free CV image in OI, where (U, V) and (X, Y) denote the angular and spatial dimensions of the LF image respectively, and 3 is color channel dimension. In the LFE, using the L_0 in form of the SAIs concatenated along with the channel dimension, named $L_{SAI} \in \mathbf{R}^{(3\times U\times V)\times X\times Y}$, a set of LF features $\mathbf{F}_{LF} = \{f^0_{LF}, \cdots, f^K_{LF}\}$ are extracted, where K indicates the number of layers in LF encoder ($\mathbf{E}_{LF}$). In the OMG, an FBS score map $I_{FBS} \in \mathbf{R}^{X\times Y\times 1}$ is directly obtained from L_0 without a deep learning-based method. The occlusion mask $M_{OMG} \in \mathbf{R}^{X\times Y\times 1}$ is obtained by refining the I_{FBS} with a CV image I_{CV} and a set of LF features $\{f^0_{LF}, f^1_{LF}, f^2_{LF}\}$. The OI reconstructs the occlusion-free CV image I_{out} from I_{CV} and M_{OMG} in a single image inpainting manner. We combine the LFE and OI by infusing the $\mathbf{F}_{LF}$ with OI decoder features $\mathbf{F}_{dc}$, to utilize the background information gathered from LFs during the inpainting step. We train our model in a fully supervised manner using ground truth occlusion free CV image I_{gt} and mask M_{gt}.

3.2. LF Feature Extractor

The main role of the LF feature extractor (LFE) is to find and extract rich information from L_{SAI}, including depth information, unoccluded background object information, and background context information. To effectively handle large disparity objects scattered around the SAIs and extract context features of background information while avoiding large occlusion, LFE requires a large receptive field.

LF feature initialization. DeOccNet [30] shows that the ResASPP module [29] is beneficial for large receptive field and helps to extract useful features required for LF-DeOcc task. We use multi-layers of the ResASPP module and residual block (ResBlock) together for the large receptive field with a dense sampling rate, which is used in Wang et al.'s method [31]. We use 1×1 convolution followed by two ResASPP Block and ResBlock layers to initialize the LF feature (f^0_{LF}). In our model, ResASPP has four parallel dilated convolutions with dilation rates of 1, 2, 4, and 8, respectively, and our ResBlock consists of 3 convolution layers and two leakyReLU layers, alternately.

LF feature encoding. With the initialized feature f^0_{LF}, a set of LF features $f^k_{LF}(k > 0)$ are extracted using a LF encoder ($\mathbf{E}_{LF}$). $\mathbf{E}_{LF}$ consists of a K encoder blocks, each of which consists of a convolution block followed by a self-attention module to give a more long-range dependency. The convolution block uses 2D convolution with kernel size of 4, stride of 2, and padding size of 1, LeakyReLU and batch normalization. Following the self-attention module which is defined in Zhang et al. [36], we use a 1×1 convolution to generate the key, query, and value matrix from the output of the convolution block. The output of the self-attention module is element-wisely added to the output of the convolution block with a learnable weight γ which is initially set to 0.25. For memory efficiency, only encoder layers that have direct skip connections to the OI use the self-attention module ($k > 1$).

3.3. Occlusion Mask Generator

The OMG generates occlusion mask M_{OMG}, from the I_{FBS} and U-shaped refinement module. I_{FBS} divides the foreground and background with respect to the zero disparity plane, and our 3-layer U-shaped network refines I_{FBS} using the I_{CV} as a guidance. In the decoder part, we reuse the LF features extracted from LFE ($f^k_{LF}, k = (0, 1, 2)$) to efficiently take advantage of depth information encoded from LFE, by concatenating the f^k_{LF} to the OMG decoder feature. Finally, we generate an occlusion mask with a softmax layer, occlusion regions as 0, and non-occlusion regions as 1, ideally. Note that contrary to the single image inpainting mask which is hard digit mask $\subset \{0, 1\}$, our generated mask is soft continuous mask $\subset [0, 1]$. Rather than thresholding the mask, which might cause a loss of information, we use the soft mask and use an appropriate inpainting method which can utilize the soft mask.

3.4. Occlusion Inpainter

Inpainting Method. We use a U-shaped single image inpainting architecture for OI. Contrary to PConv[17] which only adopts hard 0-1 mask, mask attention used in LBAM [32] can adopt soft mask due to the learnable attention map. Thus, following the LBAM architecture [32], the encoder feature of OI, $\tilde{f}^k_{ec}$, is re-normalized with mask attention A^k_M. That is, $f^k_{ec} = \tilde{f}^k_{ec} \odot A^k_M$ where $\odot$ represents element-wise multiplication. LBAM uses a mask attention map A^k_M for encoder as well as reverse mask $(1 - M_{OMG})$ attention map A^k_{RM} for decoder. The decoder feature of OI $\tilde{f}^k_{dc}$ is re-normalized with reverse mask attention, $f^k_{dc} = \tilde{f}^k_{dc} \odot A^k_{RM}$, which helps the decoder only focus on the masked region.

Feature Fusion Method. One of the important functions of the OI is to fuse the decoder features of OI (f^k_{dc}) and f^k_{LF} from LFE to reconstruct the occlusion-free CV image utilizing the visible background information from LFs. We found that 1x1 convolution shows competitive or superior performance compared to other more complicated fusion methods. Thus, for simplicity, we concatenate two features, f^k_{LF} and f^k_{dc}, and infuse them with 1x1 convolution. The fused features are element-wisely added to the f^k_{dc} with a learnable parameter γ, which is initially set to 0.25.

3.5. Loss Function

We follow the objective function used in LBAM [32] to guide the I_{out} using the I_{gt}. Our image reconstruction loss $\mathcal{L}_I$ consists of ℓ_1 loss, perceptual loss and style loss, which are described in detail in LBAM [32],

$$\mathcal{L}_I = \mathcal{L}_{\ell_1} + \lambda_1 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{style}. \tag{1}$$

Furthermore, with our modularized framework, the intermediate representation of the network, the occlusion mask M_{OMG}, can be directly guided using the M_{gt}. Thus we add a mask generation loss $\mathcal{L}_M$ and use an MSE loss for it. Finally, our entire objective function is defined as

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_M. \tag{2}$$

3.6. Arbitrary Depth Occlusion Removal

By explicitly defining the occlusion mask, more flexible applications are available. In LF-DeOcc, the occlusion is defined as *all objects placed in the foreground* based on the disparity plane d. However, in various real-world situations, the occlusion can be placed in some *arbitrary* depth plane, and the object of interest may be placed in the foreground, which is undesired to be removed. Different from previous researches, the proposed framework can only remove objects in an arbitrary depth plane while preserving the foreground object by simply manipulating

the occlusion mask M_{OMG}, without additional finetuning. Let the occlusion is placed between two disparity planes $d_1, d_2 (d_1 < d_2)$. The foreground occlusion mask for two disparity plane is denoted as $M^{d_1}_{OMG}, M^{d_2}_{OMG}$, respectively. Note that the occlusion mask M^{d}_{OMG} defines all foreground occlusions between the disparity ranges $[d, \infty]$ as 0 and background as 1, ideally. Then, the occlusion mask of objects placed between two disparity planes $[d_1, d_2]$ is defined as $M^{d_1,d_2}_{OMG} = 1-(M^{d_1}_{OMG}-M^{d_2}_{OMG})$. With the M^{d_1,d_2}_{OMG}, the occlusion in an arbitrary depth plane can be removed, without affecting the foreground object. In addition, unwanted occlusion removal can be prevented by the user definition.

4. Experiments

4.1. Experimental Setup

Training dataset For the LF-DeOcc training, the requirement of the ground truth occlusion-free CV image is strong prior. It has been researched that the mask embedding approach proposed by Wang et al. [30] which synthesizes the occlusion LF image with occlusion-free LF image and occlusion image could solve this problem. Even though it uses synthetic occlusion training data, the trained model could be well applied to various real-world occlusion scenes [30, 37]. The mask embedding approach randomly embeds 1-3 occlusion images in an LF image for a multi-disparity occlusion scenario. To allow the model to effectively learn the inpainting scenario, that is, a low disparity occlusion scenario, we place more occlusions in the low disparity planes. For the occlusion image, 21 thick and large real occlusion images are added to the original 80 mask images used in Wang et al.'s method [30] to train the inpainting scenario caused by large occlusion. A detailed description of the mask embedding approach that we applied and the disparity plane are provided in supplementary materials.

We embed the occlusion in the positive disparity planes. Thus, our ground truth occlusion-free LFs should contain only negative disparity objects. We choose 1418 LFs out of 2957 LFs from DUTLF-V2 [19] training dataset, which is a dense LF dataset captured by Lytro Illum camera [2].

Test dataset To evaluate the performance quantitatively in sparse LFs, we use 4 synthetic sparse LF scenes (4-Syn) and 9 synthetic sparse LF scenes (9-Syn) for the quantitative comparison which is synthesized by Wang et al. [30] and Zhang et al. [37], respectively. A real sparse LF image, Stanford CD scene [27] is also used for quantitative comparison as it has a ground truth. To evaluate the performance quantitatively in dense LFs, we choose 615 LFs out of 1247 LFs from DUTLF-V2 [19] test dataset and collect another 33 real occlusion images. Using the mask embedding approach with a disparity range of $[1, 4]$, single or double occlusions are embedded to evaluate multi-disparity occlusion scenario, which is denoted as *Single Occ* and *Double Occ* respectively. For the qualitative comparison, various publicly available real-world sparse and dense occlusion LF scenes are used. The sparse LF dataset is composed of LF scenes captured by Wang et al. [30] and Stanford CD scene [27]. The dense LF dataset is composed of EPFL-10 [24] and Stanford Lytro dataset [20], both captured by the Lytro Illum camera.

Training Detail The angular and spatial resolution of LF images in DUTLF-V2[19] is $(U \times V \times X \times Y) = (9 \times 9 \times 600 \times 400)$. For training and testing, we use central 5×5 images and the spatial resolution is resized to 300×200. Randomly center-cropped and horizontally flipped images with a resolution of $(X \times Y) = (256 \times 192)$ are used for our training with the mask embedding approach. We randomly choose 1-3 masks with a random RGB shuffle and horizontally and vertically flipping for embedded occlusion masks in training time. Our model is optimized by ADAM optimizer with (β_1, β_2)=(0.5,0.9), and a batch size, λ_1, and λ_1 are set to 16, 0.01, and 120. The learning rate is initially set to 0.0005 and multiplied by 0.5 every 200 epochs. We train our model on 4 Nvidia TITAN X Pascal GPUs in Pytorch framework. The epoch is set to 500 and the training step is ended within 1 day.

4.2. Experimental Results

We compare our model with the state-of-the-art LF-DeOcc methods, DeOccNet [30] and Zhang et al.'s method [37]. We train the DeOccNet with the same learning strategy and the dataset used by the original paper. For Zhang et al.'s method, we used the pre-trained model provided by the author. DeOccNet* and Zhang et al.* denotes each model trained on the same dataset and mask embedding as ours for a fair comparison in Dense LF dataset. We trained DeOccNet* from the scratch whereas Zhang et al.* is finetuned from pre-trained model provided by the author as the model trained from scratch does not converge. We also compare our model with the single image inpainting methods, RFR [16] and LBAM [32], to investigate a information gathered from various views in LFs. Since single image inpainting models can not define the foreground occlusions by itself, we additionally attach the OMG module to the single image inpainting model, where additional $(f^0_{LF}, f^1_{LF}, f^2_{LF})$ are removed to prevent the information from LFs and I_{FBS} are provided to define the occlusion. Both RFR and LBAM pretrained on the Paris Street View dataset [6] are finetuned on the same dataset we used.

4.2.1 Qualitative Results

The qualitative comparisons of ours and other methods on real-world sparse and dense LFs are shown in Figs. 3 and 4. The RFR [16] does not reconstruct the scene in both LF datasets because PConv can not accept the soft mask generated by OMG. With a learnable attention map, LBAM

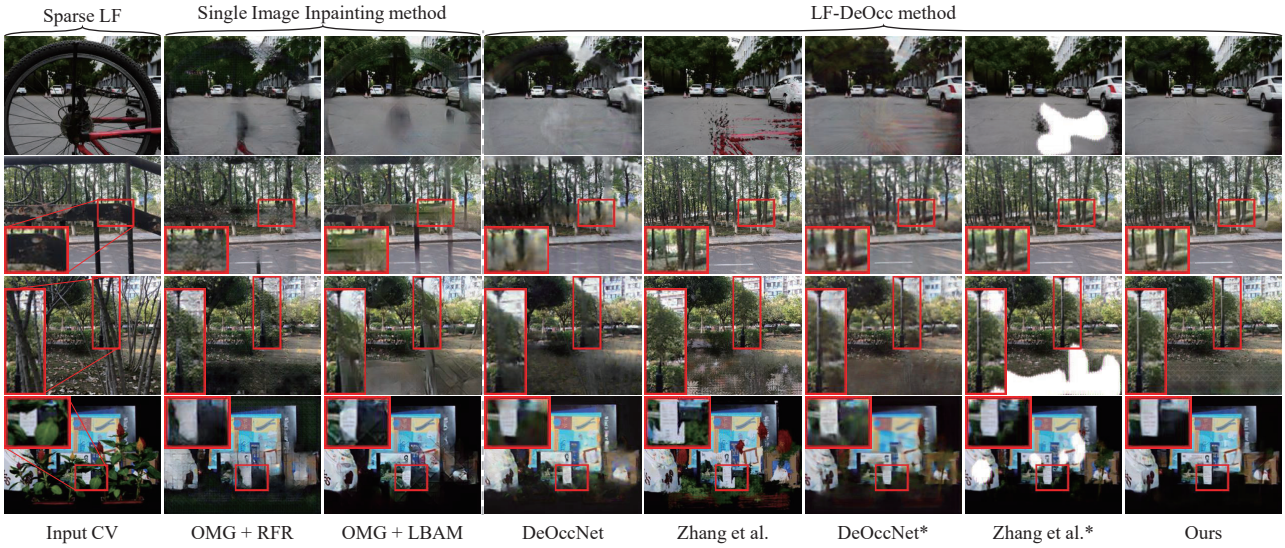

Figure 3. Qualitative comparisons on the *sparse* LF dataset. Some parts of the outputs are magnified with red boxes for a detailed comparison. Ours could reconstruct sharper occlusion-free CV images from the scene utilizing occluded background information visible in other views. The scenes in the first last which is denoted as CD [27] are used for quantitative comparison.

[32] outperforms in the dense LFs compared to RFR [16] and existing LF-DeOcc methods. However, LBAM [32] can not accurately reconstruct the output in the sparse LFs since they entirely depend on the context information and can not accurately define the complicated occlusion without $\mathbf{F}_{LF}$. We emphasize once again that RFR and LBAM do not work without OMG since the occlusion mask can not be defined in a single RGB image. The DeOccNet [30] shows blurry outputs around the occlusion in the sparse LFs and remains occlusion artifacts in both LF datasets. DeOccNet* shows better reconstruction performance in both LF datasets, but still shows blurry outputs and remains occlusion artifacts. Zhang et al.'s method [37] shows clear output if the single disparity occlusion has a large disparity (second row in Fig. 3). However, the artifacts from occlusions remain in the multi-disparity occlusions, especially in the dense LFs. Even though they are trained on the same dataset that we used, Zhang et al.* [37] shows artifacts in the dense LFs. Contrary to the single image inpainting models and other LF-DeOcc methods which only performs well in the dense and sparse LF datasets, respectively, our proposed framework generally shows better de-occlusion performance in both LF dataset. Our model generates output with few occlusion artifacts compared to other LF-DeOcc models, reconstructing clear occlusion-free CV images.

4.2.2 Quantitative Results

We use peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to quantitatively evaluate how precisely the model reconstructs the occlusion-free CV image, which is widely used metrics in LF-DeOcc and single image inpainting [30, 37, 32, 16].

Table 1 shows the summarized quantitative results. The RFR [16] and LBAM [32], which is the single image inpainting method, shows better results in the dense LF images because the dense LF requires inpainting knowledge to reconstruct the occlusion-free scene. However, single image inpainting models generally shows lower performance in the sparse LFs because they can not utilize the background information from LFs. DeOccNet [30] and DeOccNet* shows reasonable results on both dataset, but generally shows insufficient performance. As shown in Fig. 3, Zhang et al. [37] shows notable performance, especially in the single disparity occlusion of the sparse LF. Zhang et al.* [37] also does not generally shows better results. Our proposed framework generally outperforms other LF-DeOcc and inpainting models in both sparse and dense LFs.

4.3. Various Applications

4.3.1 Prevention of Unwanted Removal

In the third row of the Fig. 3, the unwanted regions may defined as an occlusion and removed, such as the ground, making serious artifacts. Contrary to other LF-DeOcc models, by explicitly defining the occlusion mask, our model could manually prevent the unwanted removal with user guidance. Fig.5 shows the original output I_{out} and edited output I_{out}^{edit} with the edited mask M_{OMG}^{edit}. With explicit user guidance to the occlusion mask, the artifacts from the ground is removed in edited output.

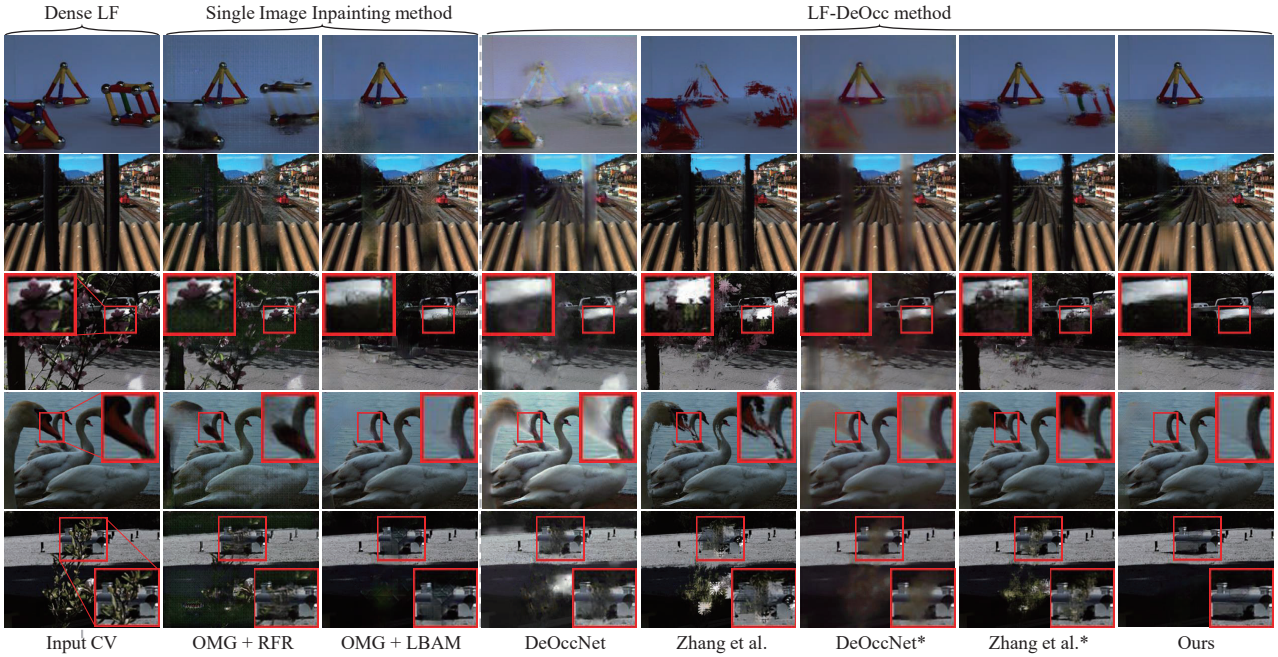

Figure 4. Qualitative comparisons on the *dense* LF dataset. Some parts of the outputs are magnified with red boxes for a detailed comparison. Ours and OMG+LBAM[32] model can generate clear occlusion-free CV image.

Table 1. Quantitative comparison on the sparse and dense LF dataset using PSNR and SSIM (PSNR/SSIM). The higher the both metric, the better the quality of the reconstructed image.

LF Type	Name	RFR + OMG	LBAM + OMG	DeOccNet	Zhang et al. [37]	DeOccNet*	Zhang et al.* [37]	Ours
Sparse(syn)	4-Syn [30]	19.89/0.668	21.11/0.677	25.04/0.807	26.37/**0.871**	23.74/0.701	14.46/0.683	**26.42**/0.836
	9-Syn [37]	20.69/0.672	23.04/0.725	21.07/0.791	**27.97/0.901**	23.70/0.715	22.00/0.758	27.04/0.849
Sparse(real)	CD [27]	21.13/0.646	21.56/0.803	21.22/0.740	18.30/0.662	22.70/0.741	20.19/0.832	**25.17/0.870**
Dense(syn)	*Single Occ*	26.28/0.867	27.92/0.899	24.84/0.863	21.98/0.815	28.67/0.914	23.15/0.900	**32.44/0.947**
	Double Occ	23.25/0.801	24.83/0.827	23.04/0.819	19.71/0.755	25.85/0.867	18.01/0.823	**28.31/0.902**

4.3.2 Arbitrary Depth Occlusion Removal

Using the mask manipulation method described in section 3.6, our model could be applied to the *arbitrary depth occlusion removal*, which selectively removes the occlusion placed in arbitrary depth. Fig. 5 shows the occlusion removal between two disparities d_1 and d_2. The objects in the intermediate depth plane (parallelepiped-shaped magnet and flower bud) are removed while preserving the foreground objects (octahedron-shaped magnet and flower) without explicit guidance by the user.

4.4. Ablation Study

The proposed framework is built upon the fact that LF-DeOcc requires various domain knowledge and by dividing the model into three specified components to deal with the sparse and dense LFs. Since the three components are related closely, our framework does not properly work if one of the components is eliminated. Thus, to verify the effectiveness of the separation, we design a *DeOccNet-large*, which enlarges the DeOccNet [30] in the channel dimension so that the number of the parameter is similar to ours (87.8M), but the model is not explicitly divided into specified components. We further experiment *DeOccNet-large + FBS*, in which I_{FBS} is concatenated to the input L_{SAI}, to verify the effect of explicit guidance of the occlusion information on the performance. Table 2 shows the quantitative results of ablation studies. With large parameters, *DeOccNet-large* and *DeOccNet-large + FBS* shows better performance in dense LFs, but still shows lower performance than ours. Especially, even though they have a large number of parameters and explicit occlusion guidance, the performance improvement on the sparse LFs is insignificant and fails to generally deal with both sparse and dense LF datasets. In addition, since our model combines single image inpainting methods and features of LFs, the feature fusion method affects the performance. We further test several attention-based fusion methods, self-attention based fusion *(SA Fusion)* and mask-feature attention based fusion *(M-F*

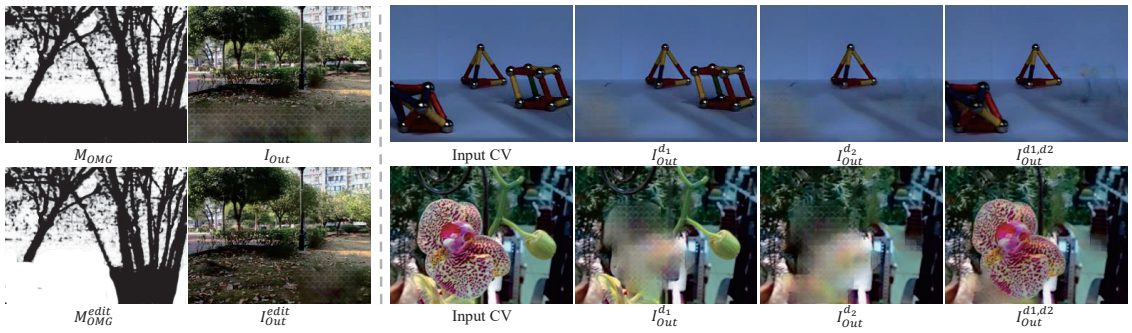

Figure 5. Illustration of the preventing unwanted removal (left) and arbitrary depth occlusion removal (right) in our model. Contrary to existing LF-DeOcc methods, our model could be applied to various de-occlusion tasks with mask manipulation.

Table 2. Ablation studies of ours and its variants on the sparse and dense LF datasets using PSNR and SSIM (PSNR/SSIM). The higher the both metric, the better the quality of the reconstructed image.

LF Type	Name	*DeOccNet-large*	*DeOccNet-large + FBS*	*SA Fusion*	*M-F Fusion*	Ours
Sparse(syn)	4-Syn[30]	24.51/0.741	24.88/0.756	26.16/0.812	26.39/0.833	**26.42/0.836**
	9-Syn[37]	24.24/0.755	25.08/0.774	26.79/0.847	26.89/**0.849**	**27.04/0.849**
Sparse(real)	CD [27]	22.80/0.752	23.72/0.822	24.31/0.862	24.39/0.864	**25.17/0.870**
Dense(syn)	*Single Occ*	29.39/0.927	31.20/0.939	32.18/0.945	32.06/0.944	**32.44/0.947**
	Double Occ	26.44/0.881	27.76/0.896	**28.35**/0.901	28.30/0.899	28.31/**0.902**

Model	LBAM [32]	Zhang et al.[37]	DeOccNet [30]	Ours
Params	69.3M	2.7M	39.0M	80.6M
T_{inf}	12ms	3050ms	10ms	24ms

Table 3. The number of parameters and the average inference time T_{inf} of each model. T_{inf} is measured when calculating the LFs with spatial resolution of 256×192 using a TITAN XP GPU.

Fusion). The attention based fusion methods also outperform existing methods, but the performance improvement is marginal compared to the 1×1 convolution we used, even though attention requires more parameters and computational powers. A detailed implementation of each fusion method is provided in supplementary materials.

4.5. Limitations and Future Works

The inpainting knowledge is necessary for LF-DeOcc in the dense LFs, requiring more parameters compared to only dealing with the sparse LF images. Thus, our model is twice as large as DeOccNet [30] (Table 3). However, ours shows reasonable inference time compared to other methods, which is appropriate for real-world applications (Table. 3). For future work, more parameter-efficient LF-DeOcc methods are expected with efficient inpainting methods. Furthermore, the inpainting knowledge of ours is sub-optimal because the inpainter is trained on a relatively small number of datasets compared to RGB dataset. Pre-trained inpainter model does not alleviate this problem due to the catastrophic forgetting. Some continual learning approaches [9, 5, 34] may be effective for this problem with a trade-off between the performance and memory, parameters, or training times. Additionally, we expect combining our proposed framework with the LF completion methods [10] expands the LF-DeOcc from reconstructing single occlusion-free CV image to the entire occlusion-free LF image and gives new perspective to LF-DeOcc task.

5. Conclusion

In this paper, we propose a deep learning-based LF-DeOcc framework, ISTY, which considers the various occlusion scenarios to work on both sparse and dense LF images. By explicitly defining the occlusion mask and fusing background information from LF images into a single image inpainting model, the proposed framework can remove occlusions not only in the foreground but also in the arbitrary depth plane. Various experimental results show that the proposed framework outperforms previous LF-DeOcc methods in both sparse and dense LF images, reconstructing clear occlusion-free images. **Expected Societal Impact.** Since the proposed framework can remove the objects in arbitrary depth plane, without affecting other objects, it could be abused to conceal the crime scene or hide crucial clues.

Acknowledgements. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00866, Development of cyber-physical manufacturing base technology that supports high-fidelity and distributed simulation for large-scalability) and National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C201270611).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481–2495, 2017.

[2] Yunsu Bok, Hae-Gon Jeon, and In So Kweon. Geometric calibration of micro-lens-based light field cameras using line features. IEEE transactions on pattern analysis and machine intelligence, 39(2):287–300, 2016.

[3] Terence Broad and Mick Grierson. Light field completion using focal stack propagation. In ACM SIGGRAPH 2016 Posters, pages 1–2. 2016.

[4] Ke-Wei Chen, Ming-Hsu Chang, and Yung-Yu Chuang. Light field image editing by 4d patch synthesis. In 2015 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2015.

[5] Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. Advances in Neural Information Processing Systems, 33:16481–16494, 2020.

[6] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? ACM Transactions on Graphics, 31(4), 2012.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[8] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

[9] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526, 2017.

[10] 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.

[11] J. Y. Lee and R.-H. Park. Depth estimation from light field by accumulating binary maps based on foreground–background separation. IEEE Journal of Selected Topics in Signal Processing, 11(7):955–964, 2017.

[12] J. Y. Lee and R.-H. Park. Separation of foreground and background from light field using gradient information. OSA Applied Optics, 56(4):1069–1078, 2017.

[13] J. Y. Lee and R.-H. Park. Reduction of aliasing artifacts by sign function approximation in light field depth estimation based on foreground–background separation. IEEE Signal Processing Letters, 25(11):1750–1754, 2018.

[14] Jae Young Lee and Rae-Hong Park. Complex-valued disparity: Unified depth model of depth from stereo, depth from focus, and depth from defocus based on the light field gradient. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(3):830–841, 2021.

[15] Jae Young Lee, Rae-Hong Park, and Junmo Kim. Occlusion handling by successively excluding foregrounds for light field depth estimation based on foreground-background separation. IEEE Access, 9:103927–103936, 2021.

[16] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7757–7765, 2020.

[17] 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 Proceedings of the European Conference on Computer Vision (ECCV), pages 85–100, 2018.

[18] Ren Ng. Digital light field photography. PhD thesis, Stanford University, 2006.

[19] Yongri Piao, Zhengkun Rong, Shuang Xu, Miao Zhang, and Huchuan Lu. Dut-lfsaliency: Versatile dataset and light field-to-rgb saliency detection. arXiv preprint arXiv:2012.15124, 2020.

[20] A. S. Raj, M. Lowney, and R. Shah. Light-field database creation and depth estimation, 2016.

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[22] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7263–7271, 2017.

[23] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[24] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. 2016.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[27] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. Computer Graphics Laboratory, Stanford University, 6(7), 2008.

[28] V. Vaish, B. Wilburn, N. Joshi, and M. Levoy. Using plane + parallax for calibrating dense camera arrays. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., volume 1, pages I–I, 2004.

[29] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12250–12259, 2019.

[30] Yingqian Wang, Tianhao Wu, Jungang Yang, Longguang Wang, Wei An, and Yulan Guo. Deoccnet: Learning to see through foreground occlusions in light fields. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 118–127, 2020.

[31] Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo. Light field image super-resolution using deformable convolution. IEEE Transactions on Image Processing, 30:1057–1071, 2020.

[32] 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 8857–8866, 2019.

[33] Liron Yatziv, Guillermo Sapiro, and Marc Levoy. Light-field completion. In 2004 International Conference on Image Processing, 2004. ICIP'04., volume 3, pages 1787–1790. IEEE, 2004.

[34] Mengyao Zhai, Lei Chen, Jiawei He, Megha Nawhal, Frederick Tung, and Greg Mori. Piggyback gan: Efficient lifelong learning for image conditioned generation. In European Conference on Computer Vision, pages 397–413. Springer, 2020.

[35] Fang-Lue Zhang, Jue Wang, Eli Shechtman, Zi-Ye Zhou, Jia-Xin Shi, and Shi-Min Hu. Plenopatch: Patch-based plenoptic image manipulation. IEEE transactions on visualization and computer graphics, 23(5):1561–1573, 2016.

[36] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In International conference on machine learning, pages 7354–7363. PMLR, 2019.

[37] Shuo Zhang, Zeqi Shen, and Youfang Lin. Removing foreground occlusions in light field using micro-lens dynamic filter. In Zhi-Hua Zhou, editor, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pages 1302–1308. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[38] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6):1452–1464, 2017.