



OPEN Learning to remove occlusions in light field images using multiscale receptive fields and feature pyramid networks

Mostafa Farouk Senussi^{1,2}, Mahmoud Abdalla¹, Mahmoud SalahEldin Kasem^{1,3}, Mohamed Mahmoud^{1,2} & Hyun-Soo Kang^{1✉}

Removal of occlusions in light-field (LF) images is strongly influenced by the receptive field of the neural network. Existing methods often suffer from limited receptive fields, restricting their ability to capture long-range dependencies and recover occluded regions effectively. To overcome this, we propose LF-PyrNet, a novel end-to-end deep learning model that enhances occlusion removal through multi-scale receptive field learning and hierarchical feature pyramid-based refinement. Our model consists of three key components. First, the feature extractor expands the receptive field by integrating Residual Atrous Spatial Pyramid Pooling (ResASPP) and a modified receptive field block (RFB). These components allow the model to capture broader context and multi-scale spatial dependencies. Next, the core occlusion reconstruction network consists of three cascaded Residual Dense Blocks (RDBs). Each block contains four densely connected layers. A Feature Pyramid Network (FPN) then performs multi-scale feature fusion and refines the representations effectively. Finally, the refinement module, which incorporates both separable and standard convolutions, enhances detailed structural consistency and improves texture restoration in occluded regions. Experimental results show that expanding the receptive field significantly enhances the occlusion removal performance, making LF-PyrNet a reliable solution for reconstructing occluded regions in LF images.

Keywords Occlusion Removal, Light Field Imaging, Receptive Field Block (RFB), Residual Atrous Spatial Pyramid Pooling (ResASPP), Residual Dense Blocks (RDB), Feature Pyramid Network (FPN), Seperable Convolution Blok (SepConv)

The ability to see through occlusions¹ is a key challenge in computer vision, with wide-reaching implications for tasks such as object detection, tracking, and scene reconstruction^{2–7}. Single-view image approaches have explored inpainting-based solutions^{8–10}. However, inpainting in the absence of explicit mask guidance becomes an ill-posed problem, as these methods rely on predefined masks to infer missing content. This constraint restricts their adaptability to real-world occlusions, which exhibit diverse and unpredictable patterns. Moreover, without leveraging auxiliary information, single-view methods struggle to reconstruct occluded regions with high fidelity, limiting their effectiveness.

LF imaging provides a promising alternative by capturing both spatial and angular information. This multi-view representation allows regions hidden in one view to be revealed in another, offering complementary cues that make occlusion removal more reliable^{11,12}. The concept of LF imaging originated with Faraday¹³ and was later formalized by Gershun¹⁴. Its computational foundation was established by Adelson et al.¹⁵ through the plenoptic function, which represents light as a seven-dimensional (7D) radiance field, modeling it at wavelength λ , time t , and spatial position (x, y, z) along direction (θ, ϕ) (Fig. 1A):

$$L = P(\theta, \phi, \lambda, t, x, y, z). \quad (1)$$

¹School of Information and Communication Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea. ²Information Technology Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt. ³Multimedia Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt. ✉email: hskang@cbnu.ac.kr

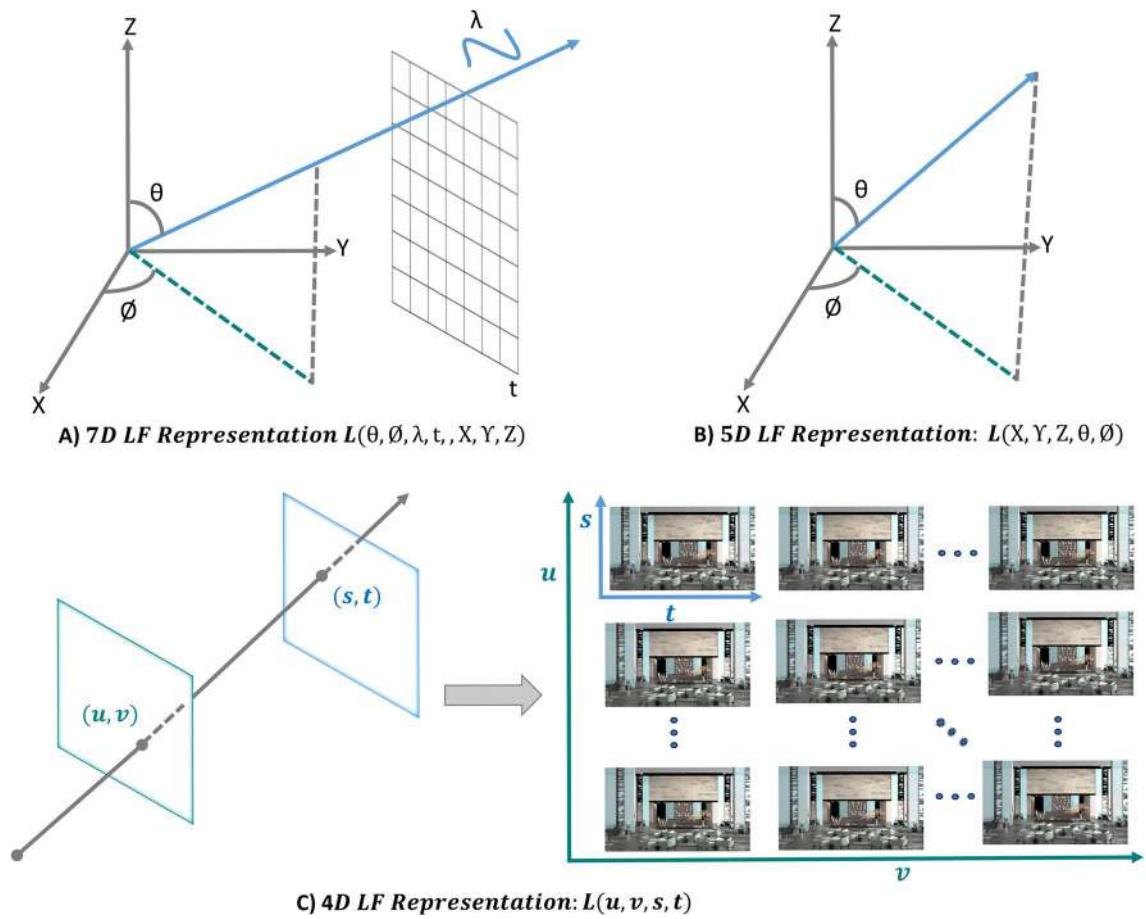


Fig. 1. Visualization of LF representations: (A) 7D $L(\theta, \phi, \lambda, t, X, Y, Z)$, (B) 5D $L(X, Y, Z, \theta, \phi)$, and (C) 4D $L(u, v, s, t)$ using two-plane parameterization, where light rays intersect parallel planes.

Due to its high dimensionality, a widely adopted simplification assumes that light radiance remains constant along its path and does not vary over time, leading to a five-dimensional (5D) representation^{16,17} (Fig. 1B):

$$L = P(x, y, z, \theta, \phi). \quad (2)$$

Levoy and Hanrahan¹⁸, building on earlier work by Curless and Levoy¹⁹ and Hanrahan²⁰, introduced the widely used four-dimensional (4D) two-plane representation. (Fig. 1C):

$$L = P(u, v, s, t), \quad (3)$$

enabling efficient multi-view capture, post-capture refocusing, and depth-aware imaging, laying the foundation for LF applications in computer vision, occlusion removal, and immersive technologies^{21,22}.

Early LF occlusion removal methods primarily relied on encoder-decoder architectures. For instance, DeOccNet²³ used ResASPP to capture multi-scale features but struggled with large occlusions, while Mask4D²⁴ introduced 4D convolutions to maintain spatial-angular coherence. GAN-based approaches²⁵ refined occluded areas via semantic inpainting, and Zhang et al.²⁶ proposed a specialized adaptive filter for restoring occluded regions in sparse LFs, though its high memory demands limited applicability to denser datasets. Modular frameworks, such as ISTEY²⁷, improved reconstruction accuracy by separating feature extraction, occlusion detection, and inpainting into distinct stages. More recent methods have emphasized expanding the receptive field to better capture both global context and local details. Hybrid CNN-Transformer architectures, such as those by Wang et al.²⁸ and SwinScNet²⁹, combine local feature extraction with global contextual modeling to handle complex occlusions efficiently. Additionally, approaches integrating CSPDarknet53 and BiFPN³⁰ have enhanced multi-scale feature fusion, improving occlusion removal across diverse LF datasets.

Despite recent progress, existing LF occlusion removal methods still face several critical limitations that motivate the development of our approach. First, they struggle to capture fine spatial details under large occlusions due to limited receptive fields and insufficient mechanisms for supplementing high-frequency information, making it challenging to detect all occluded points across sub-aperture images (SAIs). Second, effectively combining all SAIs in a grid-like angular sampling is non-trivial, as improper fusion can disrupt the intrinsic LF structure, hinder full exploitation of inter-view complementary information, and result in blurred

reconstructions and color inconsistencies. Third, many approaches treat occluded and non-occluded pixels indiscriminately, allowing invalid pixels from occluded regions to degrade reconstruction quality and introduce artifacts. These challenges underscore the need for a model that preserves LF structure, adaptively discriminates between pixel types, and enriches multi-scale spatial details through an enlarged receptive field for accurate and robust occlusion removal.

To this end, we propose LF-PyrNet, a novel end-to-end deep learning model designed for the efficient removal of occlusions in LF images. LF-PyrNet addresses these challenges by expanding the receptive field through the integration of ResASPP and a modified RFB, enabling multi-scale feature extraction and improving spatial dependency modeling. The occlusion reconstruction network consists of three cascaded RDBs, each with four densely connected layers, followed by a FPN for multi-scale feature fusion, ensuring robust occlusion reconstruction. Furthermore, the refinement module, integrating separable and standard convolutions, enhances texture details and structural consistency, leading to sharper and more realistic occlusion-free images. By explicitly addressing receptive field constraints, preserving LF structure, and enhancing multi-scale representation learning, LF-PyrNet demonstrates superior performance in reconstructing occluded regions, establishing it as a highly effective and reliable solution for complex LF occlusion scenarios.

Our contributions are fourfold:

- **Multi-scale receptive field expansion:** We integrate ResASPP with a modified RFB to enhance receptive field learning, effectively capturing both local and global spatial dependencies for robust occlusion removal.
- **Hierarchical feature fusion and reconstruction:** Our network leverages three cascaded RDBs and a FPN to enable effective multi-scale feature fusion, ensuring fine-grained occlusion reconstruction with minimal artifacts.
- **Detail enhancement and texture refinement:** Our network uses a refinement module with separable and standard convolutions to improve texture details and maintain structural consistency, resulting in clearer and more realistic occlusion-free images.
- **Extensive evaluation:** Experiments on synthetic and real LF data show the superiority of LF-PyrNet in high-quality reconstruction and handling complex occlusions.

The remainder of this paper is organized as follows: Section [Related work](#) explores related work and sets the stage for our contributions. Section [Proposed method](#) describes the technical design of our architecture. Section [Experiments](#) provides a comprehensive evaluation, both quantitatively and qualitatively, along with a comparative analysis. Section [Ablation study](#) investigates the effect of individual components on performance through an ablation study. Finally, Section [Conclusion and future work](#) discusses challenges, suggests future research directions, and concludes with the key findings of the paper.

Related work

This section explores prior methods for handling occlusions, including classical techniques and modern deep learning approaches for LF occlusion removal. It highlights key advancements to provide a clear foundation for our work.

Conventional methods

The foundation of synthetic aperture focusing in LF imaging was laid by Vaish et al.³¹, who introduced a resampling approach to improve visibility through partial occlusions. Their method aligned 4D LF data to enable refocusing across multiple planes, enhancing background clarity while blurring occluders in the foreground. Expanding on this, Vaish et al.³² applied synthetic aperture techniques to 3D reconstruction, contrasting their approach with traditional stereo methods and introducing occlusion-aware multi-view techniques based on color and entropy metrics. To facilitate LF acquisition, Vaish et al.³³ developed a calibration method utilizing planar parallax to estimate camera positions and project images onto different planes, achieving greater accuracy than conventional techniques. Pei et al.³⁴ later proposed a pixel-labeling approach to detect and mask occlusions. In a subsequent work, they incorporated image matting to generate all-in-focus images³⁵. However, this method was limited by depth-specific focal ranges. Yang et al.⁴ addressed these limitations by segmenting scenes into layers, allowing refocusing at varying depths. Xiao et al.³⁶, who introduced an iterative reconstruction strategy that used clustering to separate occlusions from the background, refining the results through optimization. Challenges remain to handle large occlusions and improve depth consistency, which requires further advances in LF imaging.

Deep learning-based methods

To overcome the limitations of conventional occlusion removal techniques, Wang et al.²³ introduced DeOccNet, a deep learning architecture. The network integrates an encoder-decoder structure with ResASPP to expand the receptive field and enhance occlusion comprehension. It employs a mask embedding technique to simulate occluded LF images, but it faces challenges in handling large occlusions and suffers from blurring due to weak spatial dependency modeling in SAI stacking. Addressing these drawbacks, Li et al.²⁴ proposed Mask4D, leveraging 4D convolution to maintain spatial coherence and angular consistency, significantly improving occlusion handling. Similarly, Zhao et al.²⁵ explored generative adversarial networks (GANs) for occlusion removal, enabling semantic inpainting that integrates occluded regions with background content for more realistic reconstruction. Zhang et al.²⁶ developed a dynamic microlens filter to extract features from shifted lenslet images in sparse LFs; however, its reliance on rigid background assumptions and substantial memory usage limits its applicability to denser LFs. To address these constraints, Zhang et al.³⁷ introduced LFORNet. The network integrates Foreground Occlusion Location (FOL), Background Content Recovery (BCR), and

a refinement module. This design enables effective management of occlusions of varying scales using multi-angle view stacks (MVAS). Further advancements include Song et al.³⁸, who proposed a dual-pathway fusion network that separately synthesizes center views and predicts occlusions, later combining them for improved reconstruction accuracy. Hur et al.²⁷ introduced ISTY, a framework incorporating modules for feature extraction, occlusion detection, and inpainting, effectively handling occlusions in both sparse and dense datasets. However, its reliance on CNNs restricts its capacity to capture complex occlusion patterns due to limited receptive fields. To address this, Wang et al.²⁸ developed a hybrid model integrating CNNs for local feature extraction and Swin Transformers for capturing global structures, improving performance in large occlusion scenarios. Expanding on this, Zhang et al.²⁹ introduced SwinScNet, incorporating ScConv blocks for efficient feature compression and a Swin-Unet to optimize occlusion removal performance while maintaining computational efficiency. Building on this progress, Senussi et al.³⁰ integrated CSPDarknet53 for hierarchical features, BiFPN for multi-scale fusion, and HiNet refinement to improve occlusion robustness. Chen et al.³⁹ proposed ATM-OAC, combining adaptive threshold mask prediction, occlusion-aware convolution, and subpixel complementation for enhanced spatial-angular extraction and view reconstruction. To address disparity-based LF challenges, Zhang et al.⁴⁰ developed a progressive MPI method using occlusion priors and attention filtering to recover backgrounds accurately. Lastly, Zhang et al.⁴¹ introduced MANet, a lightweight framework jointly predicting occlusion masks and restoring backgrounds via gated spatial-angular aggregation and texture-semantic attention.

Proposed method

In this section, we introduce LF-PyrNet, a novel deep learning model designed to address the challenges of occlusion removal in LF images. Existing methods often face difficulties in capturing long-range dependencies and effectively reconstructing missing details. To overcome these limitations, LF-PyrNet enhances occlusion removal through multi-scale receptive field learning and hierarchical feature refinement. By expanding the receptive field and improving feature fusion, LF-PyrNet accurately reconstructs occluded regions while preserving fine structural details and texture fidelity.

As illustrated in Fig. 2, LF-PyrNet consists of three key components: feature extraction, occlusion reconstruction, and refinement module. The network processes a 5×5 grid of SAIs as input. The feature extraction module integrates ResASPP and a modified RFB to effectively capture both local and global contextual dependencies, ensuring a rich feature representation across multiple scales. These features form the foundation for occlusion-aware reconstruction. They are then processed by the occlusion reconstruction module, which consists of three cascaded RDBs, each composed of four densely connected layers, followed by a FPN for multi-scale feature fusion and progressive reconstruction of missing regions. Finally, the refinement module, integrating both separable and standard convolutions, enhances texture consistency and structural integrity, ensuring a seamless transition between reconstructed and non-occluded regions. The network is trained with center-view (CV) SAI as the reference, guiding the model towards precise reconstruction and supervision.

LF feature extractor

As shown in Fig. 2, the feature extraction process begins with an initial convolutional layer, followed by the combination of ResASPP and a modified RFB. This design expands the receptive field by leveraging atrous convolutions with varying dilation rates in ResASPP to capture multi-scale features. The RFB further broadens the receptive field, enhancing the model's ability to capture both local and long-range dependencies across the LF image.

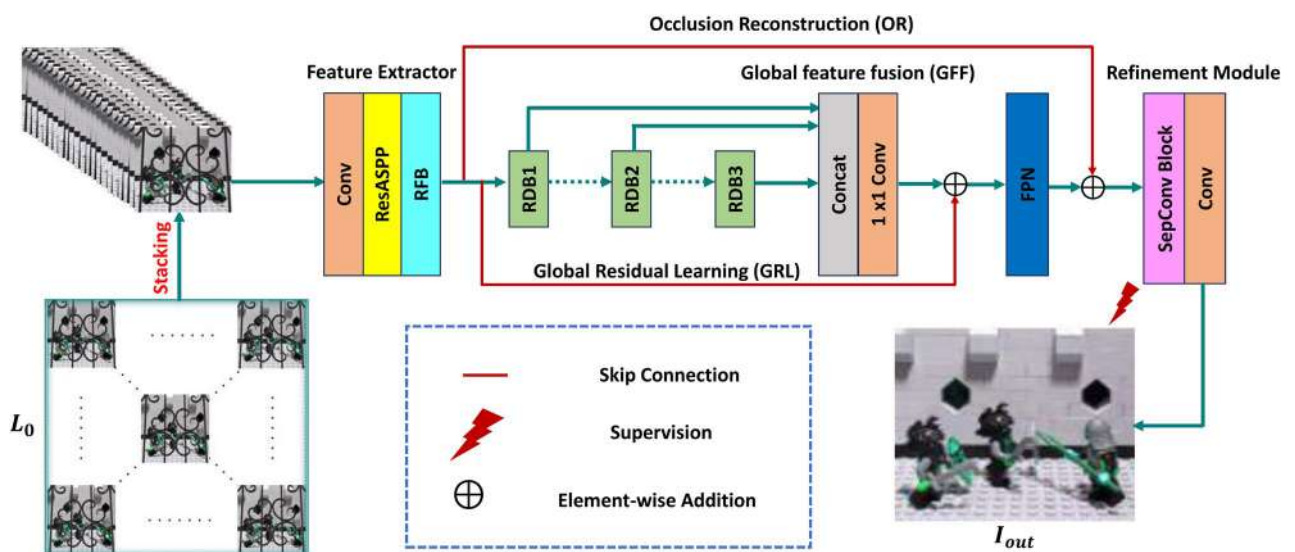


Fig. 2. Overall Architecture of LF-PyrNet for Occlusion Removal in Light Field Images, Consisting of Feature Extraction, Reconstruction, and Refinement.

The input tensor, denoted as $L_0 \in \mathbb{R}^{U \times V \times H \times W \times C_{in}}$, represents the LF image, where U and V correspond to the angular dimensions, H and W denote the spatial dimensions, and C_{in} represents the number of input channels. Specifically, for our LF images, the input tensor is structured as $L_0 \in \mathbb{R}^{5 \times 5 \times 256 \times 192 \times 3}$, capturing both the angular and spatial components, as well as the color channels of the LF data.

The convolutional layer initially processes the tensor by applying a 1×1 kernel, with a stride of 1 and padding of 1, to the input. This operation serves as an efficient mechanism to interact with the angular dimensions, effectively merging the information from the U and V angles across the channel dimension. By doing so, the network preserves the spatial and angular resolutions of the LF data while ensuring that features from all channels interact seamlessly. The output of this initial convolution is formally defined as follows:

$$F_C = \text{Conv}_{1 \times 1}(L_0), \quad (4)$$

where $F_C \in \mathbb{R}^{(3 \times U \times V) \times H \times W}$ represents the transformed tensor, rich with angular, spatial, and channel-wise features. Subsequently, the output tensor F_C is fed into the Residual Atrous Spatial Pyramid Pooling (ResASPP) module, as shown in Fig. 3. This design expands the receptive field by leveraging atrous convolutions with varying dilation rates, $d = \{1, 2, 4, 8\}$, to capture multi-scale features, enabling the model to effectively understand both local and global contextual information. The parallel atrous convolutions, each followed by a LeakyReLU activation with a leaky factor of 0.1, enrich the feature extraction process, making the model sensitive to diverse spatial patterns across the LF image.

The outputs of these convolutions are concatenated and processed through a 1×1 convolution for channel reduction, followed by a residual connection with the input tensor F_C . This residual connection ensures that the model retains essential details from the original features while enriching them with multi-scale contextual information.

The final output, F_R , is computed as:

$$F_R = F_C + \text{Conv}_{1 \times 1}(\text{Cat}(\{\text{LReLU}(\text{Conv}_d(F_C))\})), \quad (5)$$

where $d \in \{1, 2, 4, 8\}$ represents the dilation rates. This process results in the output tensor $F_R \in \mathbb{R}^{H \times W \times C_{out}}$, which effectively preserves multi-scale and context-aware features essential for accurate occlusion removal. By expanding the receptive field and capturing multi-scale patterns, the ResASPP module enhances the feature extraction process, ensuring better reconstruction of occluded regions.

Following the ResASPP module, the output tensor F_R is further refined through the modified RFB, as shown in Fig. 4, to enhance feature extraction with multi-scale receptive fields. The modified RFB module, consisting of four parallel convolutional branches, captures information at various spatial scales, using different dilation rates to extend the receptive field without introducing excessive computational complexity. Each branch of the RFB performs the following: Branch 0 applies a simple 1×1 convolution to capture fine local features, preserving detailed spatial information from the F_R tensor. Branch 1 first applies a 1×3 convolution followed by a 3×1 convolution, which captures medium-range dependencies. Then, a dilated 3×3 convolution is used to expand the receptive field and integrate broader contextual information, effectively capturing both local and intermediate spatial interactions. Branch 2 employs 1×5 and 5×1 convolutions to extract even larger-scale features, with a dilated 5×5 convolution applied afterward. This branch enhances the receptive field further, enabling the model to capture larger contextual information relevant to complex occlusions. Finally, Branch 3 uses 1×7 and 7×1 convolutions to extract long-range dependencies, followed by a dilated 7×7 convolution. This allows the model to capture global context, critical for handling large occlusions and preserving overall scene coherence.

The outputs of these branches are concatenated along the channel dimension and passed through a 3×3 convolution to reduce the dimensionality. The final result is added to the input tensor F_R through a residual connection, where a 1×1 convolution of F_R serves as the shortcut. This ensures that both fine-grained features from the input tensor and the multi-scale contextual information from the RFB are preserved. Mathematically, the output of the RFB is computed as:

$$F_M = \text{ReLU}(\text{Conv}_{3 \times 3}(\text{Cat}(\{x_0, x_1, x_2, x_3\}))) + \text{Conv}_{1 \times 1}(F_R), \quad (6)$$

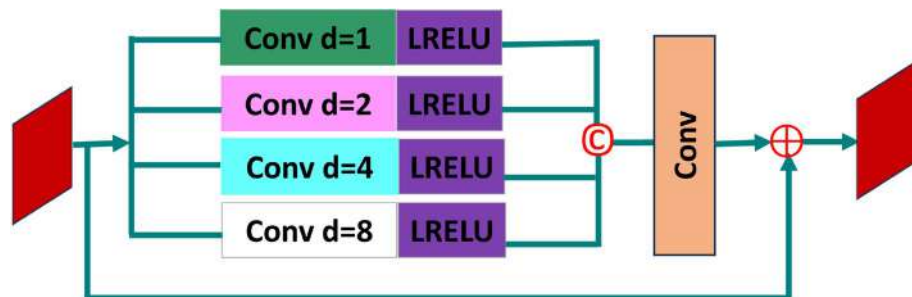


Fig. 3. The overall structure of the ResAspp Block.

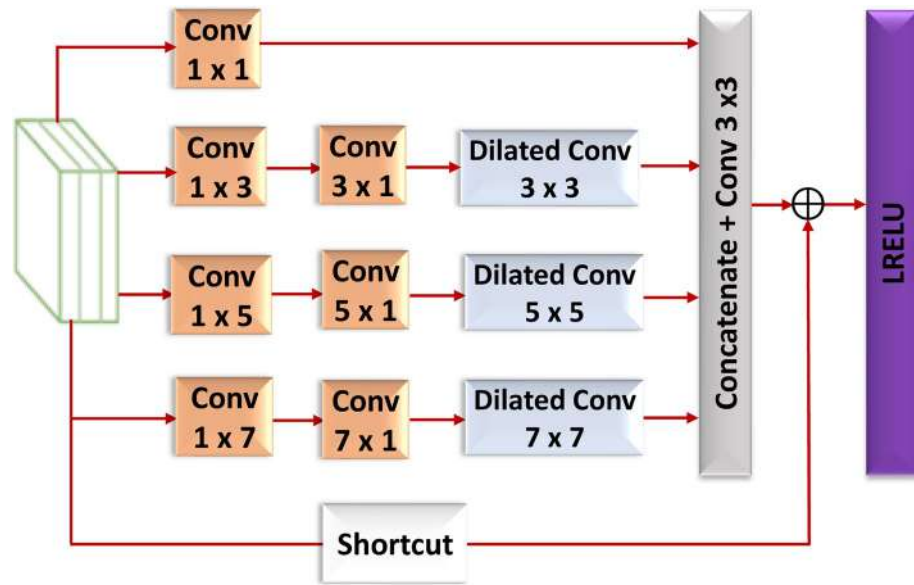


Fig. 4. Modified RFB architecture with parallel branches for multi-scale feature extraction and receptive field enhancement.

where $\{x_0, x_1, x_2, x_3\}$ represents the outputs from each branch, and $\text{Conv}_{3 \times 3}(F_R)$ is the residual connection. The resulting tensor $F_M \in \mathbb{R}^{H \times W \times C_{\text{out}}}$ contains enriched multi-scale features, capturing both local details and broader contextual dependencies, which are crucial for the next model stages.

Occlusion Reconstruction (OR)

The occlusion reconstruction subnetwork, shown in Fig. 2, is built with three cascaded Residual Dense Blocks (RDBs), each enriched with Local Feature Fusion (LFF) to refine spatial details. To enhance multi-scale representation and global coherence, we incorporate Dense Feature Fusion (DFF), which combines Global Feature Fusion (GFF) and Global Residual Learning (GRL) for robust contextual aggregation. A Feature Pyramid Network (FPN) is also utilized to facilitate hierarchical feature integration and progressive refinement of occluded regions for precise restoration.

The first RDB receives the multi-scale feature tensor F_M from the previous feature extraction stage. Through dense connectivity and local residual learning, it begins refining the occluded regions. Each subsequent RDB further enhances the feature representation by adaptively fusing newly extracted features with those from earlier layers, improving the network's capacity to handle complex occlusions.

More formally, let D represent the number of residual dense blocks. The output of the d -th RDB, denoted as F_d , is computed recursively as

$$F_d = H_{\text{RDB},d}(F_{d-1}) = H_{\text{RDB},d}(H_{\text{RDB},d-1}(\dots H_{\text{RDB},1}(F_M) \dots)), \quad (7)$$

where $H_{\text{RDB},d}$ represents the composite operations (e.g., convolutions and ReLU activations) in the d -th RDB. Each F_d encapsulates rich local features, leveraging all convolutional layers within its block. After obtaining these hierarchical features from the RDB cascade, we perform Dense Feature Fusion (DFF) to integrate global information. The fused feature map is computed as

$$F_{\text{DF}} = H_{\text{DFF}}(F_{-1}, F_0, F_1, \dots, F_D), \quad (8)$$

where F_{-1} denotes shallow feature maps from earlier stages, and H_{DFF} merges features across different levels.

Residual Dense Block (RDB)

As depicted in Fig. 5, the proposed RDB is designed to enhance feature representation through three key mechanisms: dense connectivity, local feature fusion (LFF), and local residual learning (LRL), which together establish a contiguous memory (CM) effect. In the d -th RDB, let F_{d-1} and F_d denote the input and output feature maps, respectively, each initially comprising G_0 channels.

Within each RDB, the output of the c -th convolutional layer is formulated as

$$F_{d,c} = \sigma(W_{d,c}[F_{d-1}, F_{d,1}, \dots, F_{d,c-1}]), \quad (9)$$

where σ is the ReLU activation function, $W_{d,c}$ represents the weights of the c -th convolutional layer, and the concatenation $[F_{d-1}, F_{d,1}, \dots, F_{d,c-1}]$ aggregates feature maps from the preceding RDB and all previous layers within the current block. This concatenation results in a feature set with $G_0 + (c-1) \times G$ channels, where G is

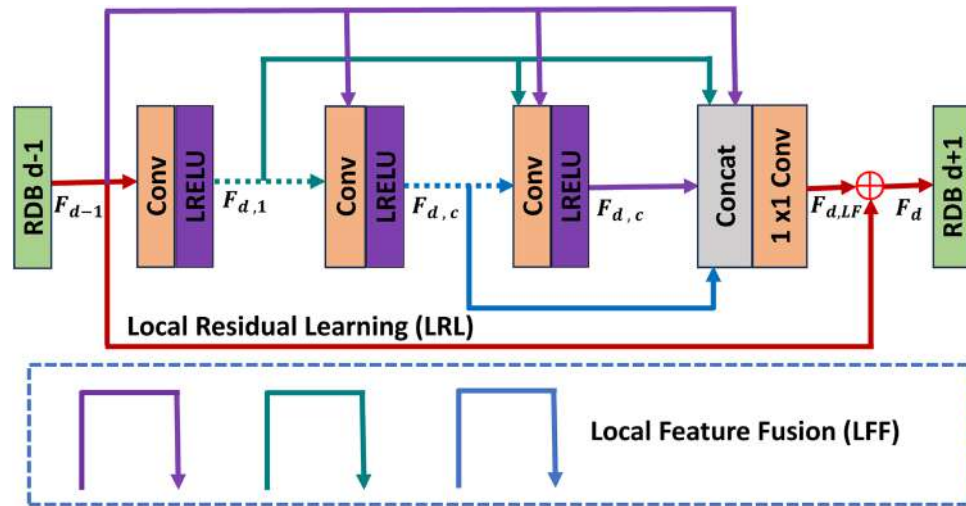


Fig. 5. Illustration of the RDB structure, which improves feature representation through dense connectivity, local feature fusion (LFF), and local residual learning (LRL).

the growth rate. To mitigate the rapid increase in feature dimensions, local feature fusion (LFF) is applied using a 1×1 convolution:

$$F_{d,LF} = H_{d,LF}([F_{d-1}, F_{d,1}, \dots, F_{d,C}]), \quad (10)$$

where $H_{d,LF}$ denotes the transformation induced by the 1×1 convolution across all C layers within the block. Finally, local residual learning (LRL) is incorporated to improve information flow, yielding the final output of the RDB as

$$F_d = F_{d-1} + F_{d,LF}. \quad (11)$$

This design effectively combines dense feature aggregation with residual connections, thereby preserving hierarchical information and enhancing gradient propagation, which are key to accurate occlusion reconstruction.

Dense Feature Fusion (DFF)

After extracting local dense features using a series of RDBs, the network performs dense feature fusion (DFF) to exploit these hierarchical features on a global scale. The DFF module first applies Global Feature Fusion (GFF) by concatenating the outputs of all RDBs:

$$F_{GF} = H_{GF}([F_1, F_2, \dots, F_D]), \quad (12)$$

where H_{GF} is a composite function that typically involves a 1×1 convolution for adaptive fusion, followed by a 3×3 convolution to further extract salient features. Subsequently, global residual learning (GRL) is employed to integrate the shallow features F_{-1} with the fused global feature F_{GF} :

$$F_{DF} = F_{-1} + F_{GF}. \quad (13)$$

This hierarchical fusion of local dense features and global residual learning creates a multi-scale representation essential for subsequent accurate occlusion-free image reconstruction.

Feature Pyramid Network (FPN)

After executing Dense Feature Fusion (DFF), the occlusion reconstruction subnetwork incorporates a Feature Pyramid Network (FPN) to effectively aggregate multi-scale contextual features and reinforce hierarchical feature propagation. Designed to balance fine-grained spatial details with high-level semantic understanding, the FPN employs a structured fusion mechanism that integrates features across multiple scales. Operating through a bottom-up feature encoding process, the FPN extracts multi-resolution representations via a sequence of transformative ResBlocks, as shown in Fig. 6. This is followed by a top-down refinement stage, where hierarchical feature fusion is achieved through lateral connections, enabling effective cross-scale feature propagation. Using these pathways, the FPN preserves high-frequency details and broader context, improving the network's ability to reconstruct occluded regions more clearly.

Given an input feature map F^{DFF} from the preceding DFF module, the bottom-up hierarchical encoding process extracts multi-scale features through successive transformations, formulated as:

$$F^\ell = H_{\text{ResBlock},\ell}(F^{\ell-1}), \quad (14)$$

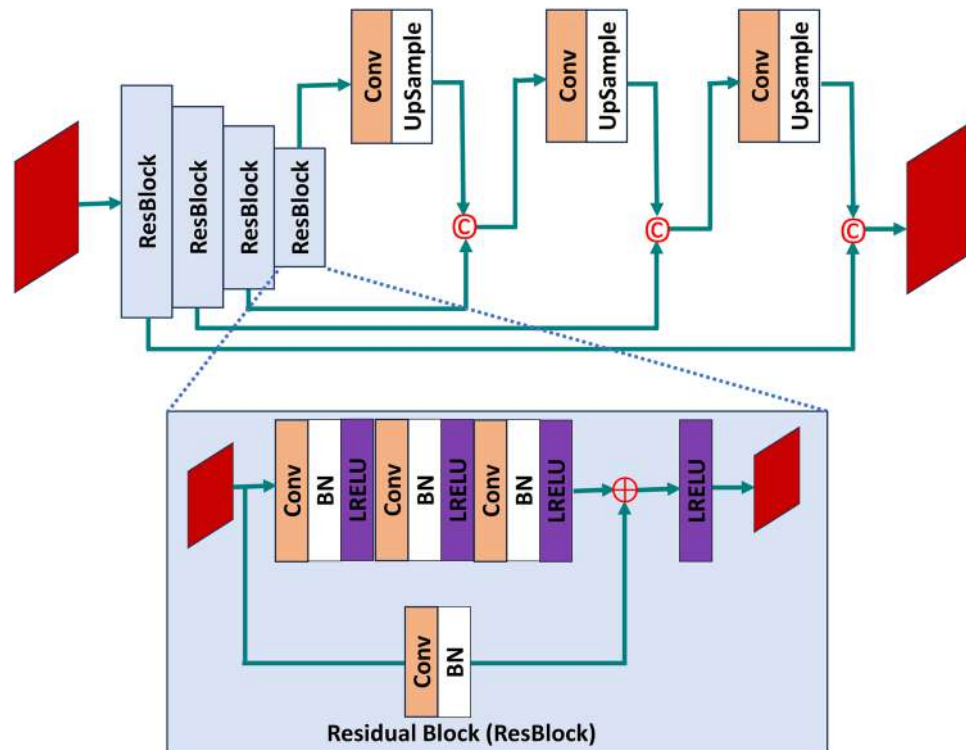


Fig. 6. Schematic representation of the FPN architecture, enhancing feature integration by fusing multi-scale features through pyramidal layers.

where $H_{\text{ResBlock},\ell}(\cdot)$ represents the residual transformations at level ℓ , typically implemented using convolutional layers with skip connections to maintain feature integrity and enhance gradient flow. The ResBlock is a fundamental component in this process, structured with three sequential convolutional layers, batch normalization, and Leaky ReLU activation functions. Batch normalization stabilizes training by normalizing feature distributions, while Leaky ReLU activation addresses vanishing gradients, enhancing feature representation. A key aspect of the ResBlock is its skip connection, which directly propagates the input feature map to the output through convolution and batch normalization layers. This mechanism helps retain spatial details, prevents degradation in deeper network layers, and facilitates more effective training. Through residual learning, the network focuses on refining occlusion-reconstructed features while preserving the underlying structural integrity. The final activation layer ensures robust feature propagation, maintaining consistency in reconstructed features across multiple scales.

To enable effective feature propagation, the network employs a top-down fusion mechanism, where high-resolution details are preserved by combining deeper feature representations with intermediate-level features via lateral connections. This is formally expressed as:

$$\hat{F}^\ell = H_{\text{lat},\ell}(F^\ell) + \text{Upsample}(F^{\ell+1}), \quad (15)$$

where $H_{\text{lat},\ell}(\cdot)$ is a 1×1 convolution applied to intermediate features for dimensionality alignment, and Upsample represents a spatial upsampling operation that enhances feature consistency across scales. The fused hierarchical features are further processed through a smoothing transformation to refine the occlusion-reconstructed representation, ensuring both local consistency and global contextual integrity:

$$F_{\text{FPN}} = H_{\text{smooth}}(\hat{F}^1, \hat{F}^2, \dots, \hat{F}^L), \quad (16)$$

where H_{smooth} consists of convolutional layers, typically a combination of 3×3 convolutions and nonlinear activations, which act to enhance local feature continuity while preserving multi-scale information. Thus, the FPN constitutes the final stage of the occlusion reconstruction subnetwork, consolidating multi-scale features into a robust occlusion-reconstructed representation F_{FPN} . This output is subsequently transmitted to the refinement subnetwork, which performs additional fine-grained enhancement to ensure superior occlusion-free image reconstruction.

Refinement module

The refinement subnetwork is designed to enhance the occlusion-reconstructed features at a fine-grained level, ensuring the generation of a high-quality, occlusion-free output image. As shown in Fig. 2, it consists of two key

components: a Separable Convolution Block (SeparableConvBlock) followed by a standard convolutional layer that reduces the channel dimension to three, representing an RGB image.

Central to the refinement module is the SeparableConvBlock, which decomposes the standard convolution operation into two distinct processes: a depthwise convolution followed by a pointwise convolution, as shown in Fig. 7. In the depthwise stage, a $k \times k$ convolution layer is applied, where $k \times k$ represents the kernel size (set to 3×3 in our work). A unique filter is applied independently to each input channel, enabling the block to capture spatial information while significantly reducing computational cost. The pointwise convolution, implemented as a 1×1 convolution, then fuses these per-channel responses, facilitating the integration of information across channels. Finally, the output from the SeparableConvBlock, which refines the 64-channel feature representation, is passed through a 3×3 convolutional layer that reduces the feature map to three channels, corresponding to the final occlusion-free RGB image. The full refinement operation can be expressed by the following equation:

$$I_{\text{out}} = \text{Conv}_{3 \times 3} (\text{SepConvBlock} (F_{\text{FPN}})). \quad (17)$$

This structured sequence of operations guarantees that the output image I_{out} is free from occlusions, with enhanced clarity and detail.

Loss function

To optimize image reconstruction, we employ a hybrid loss function comprising MAE, SSIM, and Perceptual Loss, balancing pixel-wise accuracy, perceptual fidelity, and structural consistency:

$$\mathcal{L} = k_1 L_{\text{MAE}} + k_2 L_{\text{SSIM}} + (1 - k_1 - k_2) L_{\text{PER}}, \quad (18)$$

where $k_1 = 0.30$ and $k_2 = 0.35$ are empirically set.

MAE loss minimizes pixel-wise discrepancies:

$$L_{\text{MAE}} = \frac{1}{HW} \sum_{i,j} |I_{i,j} - \hat{I}_{i,j}|. \quad (19)$$

SSIM loss enhances structural coherence:

$$L_{\text{SSIM}} = 1 - \text{SSIM}(I, \hat{I}), \quad (20)$$

where SSIM evaluates luminance, contrast, and structure.

Perceptual loss enforces high-level feature and texture consistency:

$$L_{\text{PER}} = L_{\text{FEAT}}^{\tau} + L_{\text{STYLE}}^{\tau}, \quad (21)$$

This formulation ensures fine detail preservation and robust occlusion handling.

Experiments

Experimental setup

To build upon the foundations laid in²⁷ and³⁰, we adopted their core principles for training and evaluating our network while making adjustments to better suit the specifics of our experimental setup. Our approach was guided by their comprehensive training techniques and evaluation criteria, ensuring a robust comparison while

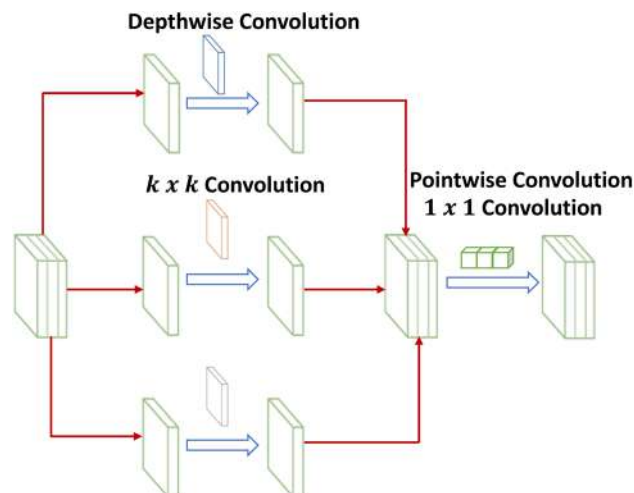


Fig. 7. Illustrative diagram of the Separable Convolution (SepConv) Block.

integrating our own modifications to optimize performance. The subsequent subsections detail the procedural steps taken during both training and testing phases.

Training dataset

For our LF-PyrNet network, we built a well-structured training pipeline using a carefully selected dataset that includes both real-world and synthetic occlusions. To create realistic occlusion cases, we applied the mask embedding strategy from²³, where occlusion masks are added to occlusion-free LF images. This method allows us to generate a variety of occlusion patterns, helping the model learn to handle different levels of complexity. To introduce variability in occlusion patterns, we randomly embed one to three occlusion masks per sample, simulating different disparity conditions within the LF images. While the original dataset included 80 synthetic masks from²³, we expanded it by adding 21 more masks, specifically designed to be larger and denser, making the occlusion removal task more challenging. These new masks were taken from real-world images to improve the model's ability to reconstruct missing details in practical scenarios. To ensure accurate training, we only included LF images where occluded objects had negative disparity, guaranteeing reliable ground-truth data. In total, we selected 1,418 LF images from the DUTLF-V2 dataset⁴², which was captured using the Lytro Illum camera⁴³. By integrating an enriched occlusion dataset and strategic augmentation, our approach enables the model to learn how to remove occlusions effectively, making it more adaptable to real-world challenges.

Testing dataset

To thoroughly evaluate our network's ability to handle occlusions in sparse LF images, we conduct experiments on both synthetic and real-world datasets. Specifically, we test our model on two synthetic sparse LF datasets: the 4-Syn and 9-Syn datasets^{23,28}, which are particularly challenging due to their angular sparsity. These datasets contain sparsely sampled LF images with occluded regions at multiple disparity levels, enabling us to analyze the model's ability to reconstruct missing details in complex scenarios. To further validate real-world performance, we use the Stanford CD dataset³³ for real sparse LF data, which provides a ground truth occlusion-free image, allowing us to measure how accurately our model reconstructs occluded regions. In addition, for dense LF scenarios, we extract 615 images from the DUTLF-V2 test dataset⁴², supplemented by 33 real occlusion cases, providing a realistic evaluation under complex occlusion patterns. To mimic occlusions with multiple depth levels, we employ a mask embedding strategy, generating Single Occ and Double Occ cases with disparities ranging from 1 to 4. This setup enables a controlled yet realistic evaluation of how the model restores occluded details across different depth layers. Beyond synthetic dataset evaluations, we conduct qualitative tests on publicly available real-world LF datasets. The Stanford Lytro dataset⁴⁴ and the EPFL-10 dataset⁴⁵ feature dense LF scenes with intricate occlusion patterns and varying depth complexities, providing a strong testbed for assessing real-world applicability. By using both synthetic and real-world datasets, we set up a complete evaluation framework to test our network on different types of occlusions in LF data.

Training details

For our experiments, we hire the DUTLF-V2 dataset⁴², which offers high-resolution LF images with $9 \times 9 \times 600 \times 400$ for angular and spatial dimensions. For our purposes, we focus on the central 5×5 views, reducing the spatial resolution to 300×200 . To enhance the training procedure, we center-crop and horizontally flip the images, resizing them to 256×192 . To introduce occlusions, we employ a mask embedding strategy that randomly selects, combines, and positions one to three RGB masks within the images. We train the model using the ADAM optimizer with settings $(\beta_1, \beta_2) = (0.5, 0.9)$ and a batch size of 18. Regularization is applied with $\lambda_1 = 0.01$ and $\lambda_2 = 120$, while the learning rate begins at 0.001 and is halved every 150 epochs. The entire training process spans 500 epochs, takes around 1 day on a single Nvidia GeForce 4090 GPU, using the PyTorch framework.

Experimental results

To evaluate our model, we perform a series of experiments on de-occluded LF images, benchmarking its performance against ten state-of-the-art LF occlusion removal techniques, including DeOccNet²³, Mask4D²⁴, Zhang et al.²⁶, LFORNet³⁷, Wang et al.²⁸, ISTY²⁷, Senussi et al.³⁰, ATM-OAC³⁹, Zhang et al.⁴⁰, and MANet⁴¹. To gain deeper insight into the role of angular information, we also compare our model with popular single-image inpainting methods, such as RFR⁴⁶ and LBAM⁴⁷. For fairness, DeOccNet²³ and Senussi et al.'s method³⁰ are retrained using our consistent training setup. The ISTY²⁷ model is evaluated using the authors' pretrained weights, while performance results for RFR⁴⁶, LBAM⁴⁷, and Zhang et al.²⁶ are adopted from ISTY²⁷. Other methods are evaluated using results from their original papers due to unavailable code.

Quantitative results

LF-PyrNet delivers outstanding performance on sparse LF datasets and exhibits competitive results on dense LF datasets. This is evidenced by the quantitative analysis in Table 1, where the PSNR and SSIM metrics highlight its strong effectiveness in occlusion removal.

For synthetic sparse LF datasets, LF-PyrNet achieves the highest PSNR score on 9-Syn (28.13 dB), outperforming all methods including Senussi et al.³⁰, the previous top performer in this category. Although LF-PyrNet's 4-Syn score (27.41 dB) ranks just outside the top three, behind Wang et al.²⁸ (27.87), MANet⁴¹ (29.17), and Zhang et al.⁴⁰ (29.70), these results highlight its effectiveness in managing occlusions across different disparity levels. Furthermore, on the real sparse LF dataset (CD), LF-PyrNet again achieves the best PSNR (25.79 dB), demonstrating its strong generalization to real-world occlusion cases. In terms of SSIM, which measures the structural similarity and perceptual quality of the reconstructed images, LF-PyrNet consistently achieves the highest scores on 9-Syn (0.870) and CD (0.887). In 4-Syn, it ranks third (0.872), yet still outperforming

Method	Sparse (syn)		Sparse (real)	Dense (syn)		Train Dataset	Image Size
	4-Syn	9-Syn	CD	Single Occ	Double Occ		
RFR ⁴⁶	19.89 / 0.668	20.69 / 0.672	21.13 / 0.646	26.28 / 0.867	23.25 / 0.801	DUTLF_V2	256×192
LBAM ⁴⁷	21.11 / 0.677	23.04 / 0.725	21.56 / 0.803	27.92 / 0.899	24.83 / 0.827	DUTLF_V2	256×192
DeOccNet ²³	23.74 / 0.701	23.70 / 0.715	22.70 / 0.741	28.67 / <u>0.914</u>	25.85 / 0.847	DUTLF_V2	256×192
Mask4D ²⁴	27.75 / 0.804	-	<u>25.78</u> / 0.739	-	-	DeOccNet_Train	224×224
Zhang et al. ²⁶	14.46 / 0.683	22.00 / 0.758	20.19 / 0.832	21.98 / 0.815	18.01 / 0.823	DUTLF_V2	256×192
LFORNet ³⁷	23.73 / 0.789	-	-	-	-	DeOccNet_Train	224×224
Wang et al. ²⁸	27.87 / 0.854	-	-	-	-	DeOccNet_Train	224×224
ISTY ²⁷	26.42 / 0.836	27.04 / 0.849	25.17 / 0.870	32.44 / 0.947	28.31 / 0.902	DUTLF_V2	256×192
Senussi et al. ³⁰	27.32 / 0.862	<u>27.48</u> / <u>0.853</u>	25.68 / <u>0.886</u>	30.70 / 0.838	<u>29.34</u> / <u>0.850</u>	DUTLF_V2	256×192
ATM-OAC ³⁹	26.74 / 0.892	-	-	-	-	DeOccNet_Train	224×224
Zhang et al. ⁴⁰	29.70 / 0.872	-	-	-	-	DeOccNet_Train	224×224
MANet ⁴¹	<u>29.17</u> / <u>0.874</u>	-	-	-	-	DeOccNet_Train	224×224
LF-PyrNet (Ours)	27.41 / 0.873	28.13 / 0.870	25.79 / 0.887	<u>31.96</u> / 0.826	30.11 / 0.828	DUTLF_V2	256×192

Table 1. Detailed quantitative evaluation on sparse and dense LF datasets using PSNR/SSIM. The best, second-best, and third-best results are shown in bold, underline, and bold-underline, respectively. A dash (–) indicates no evaluation of the model for that dataset.

most competing methods. This highlights its strong visual precision and preservation of the structure despite challenging occlusions. The model’s ability to surpass in sparse LF settings can be attributed to its innovative architectural components. The RFB enhances contextual understanding by capturing multi-scale features, allowing the model to process occlusions of varying sizes more effectively. The RDB facilitates feature reuse and long-range dependencies modeling, improving texture consistency and fine-detail reconstruction. The FPN strengthens multi-scale occlusion-aware feature fusion, ensuring that occluded regions are restored with higher accuracy. Unlike LF-PyrNet, methods like RFR ⁴⁶ and LBAM ⁴⁷ face limitations, as their single-image inpainting approach neglects the rich angular and background cues inherent in LF data, leading to suboptimal performance in occlusion removal. DeOccNet ²³ demonstrates reasonable performance but lacks consistency, particularly in more complex occlusions. Zhang et al. ²⁶ make notable strides in specific cases, but their approach is constrained by assumptions about background visibility, limiting its performance in broader scenarios. Conversely, ISTY ²⁷ and Senussi et al. ³⁰ face challenges in handling occlusions and sparse data due to its reliance on local receptive fields in CNNs. While Wang et al. ²⁸, MANet ⁴¹, and Zhang et al. ⁴⁰ previously achieved the best results, their methods did not evaluate performance on the 9-Syn or CD datasets, which limits their comparability.

For dense LF datasets, LF-PyrNet remains competitive. In the Single Occ scenario, it attains a PSNR of 31.96 dB, which is slightly lower than ISTY²⁷ (32.44 dB) but still superior to other competing methods. Similarly, in the Double Occ setting, LF-PyrNet achieves 30.11 dB, demonstrating effective handling of occluded regions, even under multi-occlusion conditions. Although its SSIM scores (0.826 for Single Occ and 0.828 for Double Occ) are slightly lower than those of some methods, its higher PSNR in most other cases suggests it recovers more accurate pixel values, indicating stronger pixel-wise reconstruction.

Overall, LF-PyrNet achieves state-of-the-art performance, leading in PSNR across most cases and delivering robust occlusion removal, even when slightly behind in SSIM for dense scenarios.

Qualitative results

By effectively reconstructing occluded regions with enhanced structural integrity and precise texture details, LF-PyrNet surpasses existing methods on sparse LF datasets, as seen in Fig. 8. In the first four rows, which correspond to synthetic sparse scenes, single-image inpainting methods struggle to recover fine details. RFR ⁴⁶ and LBAM ⁴⁷ produce imperfect reconstructions, exhibiting noticeable blurring and missing textures in the occluded regions. While DeOccNet ²³ shows a partial enhancement, it still introduces structural inconsistencies, particularly in complex patterns, such as the lattice structures in rows 2 and 3. Zhang et al. ²⁶ yields outputs with excessive smoothness, which blurs crucial details, whereas ISTY²⁷ achieves better structural restoration but fails to eliminate residual distortions. Senussi et al. ³⁰ offers notable results, yet it exhibits difficulties in restoring intricate patterns and textures, as seen in rows 3 and 4. LF-PyrNet clearly outperforms other methods in preserving textural details and geometric consistency. The integration of the RFB allows for more effective multi-scale context aggregation, enhancing occlusion recovery in structured patterns, as demonstrated in row 2, where fine details in the interior scene are reconstructed more accurately. The RDB strengthens feature propagation, enabling the model to recover high-frequency details, which is particularly evident in the restoration of lattice structures in rows 3 and 4. Additionally, the FPN enables better multi-scale learning, improving the handling of both thin and small occlusions, as seen in row 1. In the last row, corresponding to a real-world scene, the complexity of occlusions increases due to natural textures and variations in illumination. RFR ⁴⁶, LBAM ⁴⁷, and Zhang et al. ²⁶ produce reconstructions with severe blurring, while ²⁷ and Senussi et al. ³⁰ manage to retain some details but still exhibit residual occlusions and fail to preserve texture continuity. In contrast, LF-PyrNet

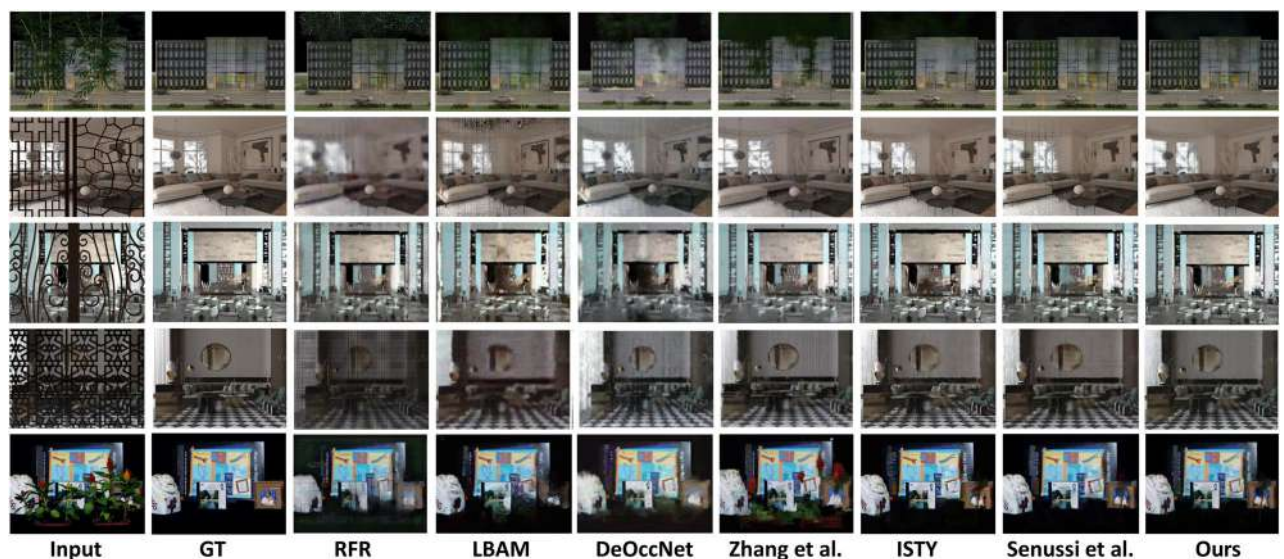


Fig. 8. Visual comparison of LF-PyrNet and existing methods on the sparse LF dataset.

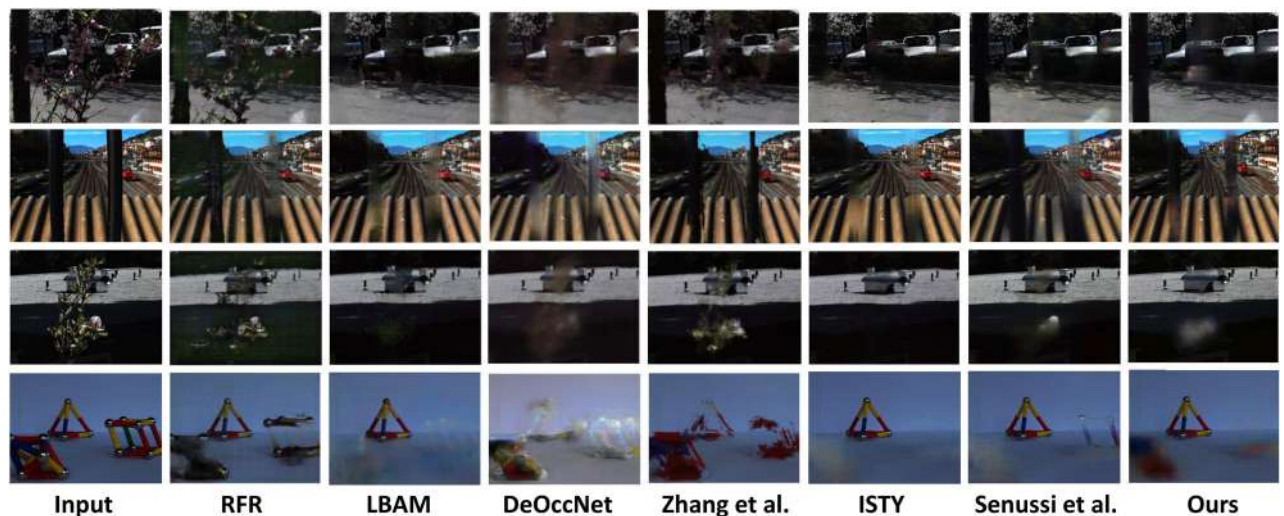


Fig. 9. Evaluation of LF-PyrNet and occlusion removal methods on the dense LF dataset, illustrating key visual differences.

successfully restores intricate textures and maintains the original scene characteristics, demonstrating its effectiveness in occlusion removal for both synthetic and real-world scenarios.

In the dense LF dataset, as depicted in Fig. 9, the visual comparison highlights the strengths and weaknesses of various state-of-the-art occlusion removal methods, particularly in single and double occlusion scenarios. LF-PyrNet, while performing well in recovering fine details, shows a trade-off when it comes to maintaining structural consistency. In the single occlusion cases (rows 1 and 3), LF-PyrNet effectively reconstructs key visual elements with sharpness and reduced artifacts. For instance, in the first row, it successfully recovers the overall shape and texture of the branches with impressive clarity, avoiding the haziness observed in RFR⁴⁶ and LBAM⁴⁷, as well as the color bleeding seen in DeOccNet²³. However, when compared closely with ISTY²⁷ and Senussi et al.³⁰, subtle inconsistencies emerge. These include slightly irregular edges and a mild misalignment of fine patterns. While the pixel-wise accuracy is high, as reflected by a competitive PSNR of 31.96 dB, the perceptual integrity of the reconstructed content, indicated by a lower SSIM of 0.826, is slightly compromised. This is more apparent in the third row, where LF-PyrNet generates a clean fill of the occluded region in the flower scene, but the continuity of shadows and spatial alignment fall short compared to the smoother outputs of ISTY²⁷.

Moving to the double occlusion scenarios (rows 2 and 4), where challenges intensify due to overlapping regions, LF-PyrNet continues to excel in clarity and pixel-level restoration. In the second row, the method succeeds in reestablishing the repetitive vertical structures of the bridge, outperforming other approaches that either blur or warp these patterns. High-frequency textures, such as pole shadows and gaps, are better preserved

than in most competing models, further supporting its PSNR result of 30.11 dB. However, compared to ISTY²⁷, its spatial coherence is slightly weaker—while ISTY maintains a more perceptually accurate scene layout, LF-PyrNet prioritizes detail fidelity, sometimes at the expense of structural uniformity. This subtle imbalance is also evident in the fourth row, where LF-PyrNet produces crisp shapes, avoiding the distortions seen in methods like DeOccNet²³ and LBAM⁴⁷. Nevertheless, close observation reveals minor geometric distortions around object boundaries, once again highlighting the PSNR-SSIM trade-off.

Performance evaluation on real-world data

Restoring occluded regions in real-world LF images poses a significant challenge, requiring both structural accuracy and texture preservation. Figure 10 presents a comparative visual analysis of three occlusion removal methods: DeOccNet²³, Senussi et al.³⁰, and our LF-PyrNet, which highlights the effectiveness of LF-PyrNet in reconstructing missing details. The visual comparison begins with the original occluded scenes in the first column, followed by the results of DeOccNet²³, Senussi et al.³⁰, and LF-PyrNet in the subsequent columns. In the first row, where the occlusion involves a bicycle wheel obstructing the background, LF-PyrNet effectively reconstructs the missing details with high clarity. It ensures a smooth and natural transition between occluded and non-occluded regions. In contrast, DeOccNet²³ exhibits significant blurring, failing to recover fine details, while Senussi et al.³⁰ leave residual traces of occlusion, impacting the overall visual coherence of the scene.

The second row introduces a more convoluted occlusion, where a metal fence partially obstructs a natural scene. In this case, LF-PyrNet demonstrates its ability to recover occluded regions while preserving structural consistency. Meanwhile, DeOccNet²³ produces overly smooth textures that detract from the natural appearance, while Senussi et al.³⁰ leave subtle occlusion remnants, leading to visible structural distortions in the scene. The third row, characterized by dense and thin occlusions from overlapping tree branches, presents an even greater challenge. While DeOccNet²³ struggles to differentiate between occlusion and background, resulting in texture loss and excessive smoothing, and Senussi et al.³⁰ fail to maintain spatial consistency, LF-PyrNet effectively reconstructs the fine details of the background, preserving natural textures with minimal distortion and fewer inconsistencies.

Evaluation of computational efficiency

Table 2 provides a detailed comparison, showing how LF-PyrNet performs compared to state-of-the-art methods in terms of computational efficiency, particularly in model size, inference speed, and training time on a Nvidia RTX 4090 GPU. LF-PyrNet stands out with just 16.97M parameters, making it one of the most efficient architectures in terms of model size among competing methods. It is significantly smaller than RFR⁴⁶ (30.59M), LBAM⁴⁷ (69.3M), ISTY²⁷ (80.6M) and Senussi et al.³⁰ (52.59M), demonstrating its ability to maintain strong performance while using fewer parameters. Even compared to DeOccNet²³ (39.0M) and Wang²⁸ (44.02), which are also known for their efficient parameter usage, LF-PyrNet offers a more optimized structure. Although Zhang et al.²⁶ (2.7M), ATM-OAC³⁹ (5.48M), and MANet⁴¹ (2.4M) are designed to be lightweight, their efficiency is severely compromised by the long inference time, making them impractical for real-time applications.

Regarding inference time, while DeOccNet²³ (0.01 s), LBAM⁴⁷ (0.012 s), and ISTY²⁷ (0.024 s) exhibit faster inference times, these models require significantly more parameters, leading to higher memory demands and potential scalability issues. In contrast, LF-PyrNet achieves an inference time of 0.043 s, outperforming RFR⁴⁶ (6.76 s), Senussi et al.³⁰ (0.138 s), Wang²⁸ (2.63 s), ATM-OAC³⁹ (1.04 s), MANet⁴¹ (1.74 s) and Zhang et al.²⁶ (3050 ms) by a considerable margin, offering a strong balance between computational cost and processing

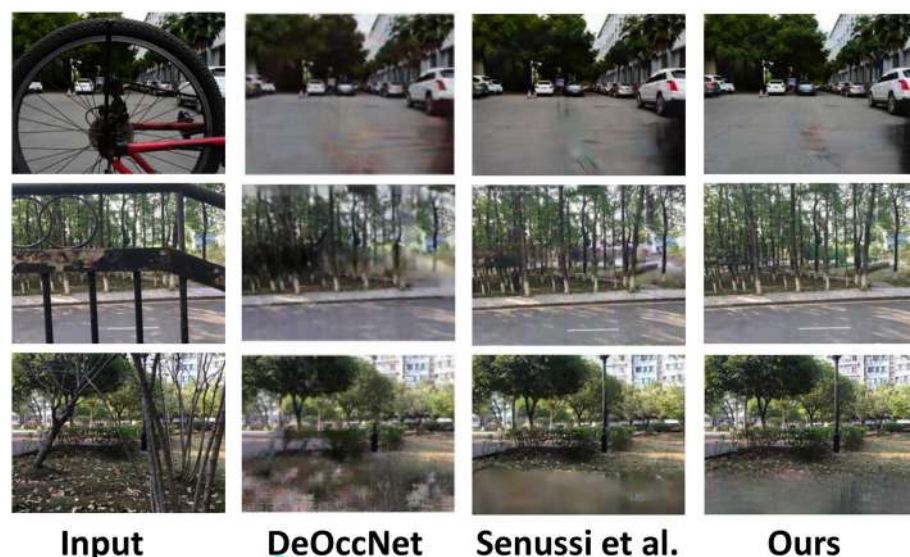


Fig. 10. Real-world scene restoration with LF-PyrNet, showcasing its superiority in occlusion removal compared to DeOccNet²³ and Senussi et al.³⁰.

Model	Param (M) ↓	Inf (S/Img) ↓	Train (Days) ↓	Img Size
RFR ⁴⁶	30.59	6.78	-	256x192
LBAM ⁴⁷	69.3	0.012	-	256x192
DeOccNet ²³	39.0	0.01	2	256x192
Zhang ²⁶	2.7	3.05	3	256x192
Wang ²⁸	44.02	2.63	2	224x224
ISTY ²⁷	80.6	0.024	1	256x192
Senussi ³⁰	52.59	0.138	0.8	256x192
ATM-OAC ³⁹	5.48	1.04	-	224x224
MANet ⁴¹	2.4	1.74	9.7	224x224
Ours	16.97	0.043	1	256x192

Table 2. Comparison of model size, inference time, and training time, where lower values (↓) indicate better performance and ‘-’ denotes unreported data.

LF Type	disparity	LF Methods				
		RFR ⁴⁶	LBAM ⁴⁷	DeOccNet ²³	Zhang et al. ²⁶	Ours
4-Syn	(-12, 9)	25.73 / 0.815	25.00 / 0.809	24.56 / 0.810	27.09 / 0.866	26.89 / 0.852
4-Syn	(-9, -4)	19.89 / 0.608	21.11 / 0.677	23.74 / 0.701	14.46 / 0.683	27.41 / 0.873
4-Syn	(-6, 4.5)	24.19 / 0.772	23.39 / 0.766	25.06 / 0.812	26.96 / 0.823	27.13 / 0.841
9-Syn	(-5, 5)	27.94 / 0.881	27.85 / 0.876	20.92 / 0.738	28.39 / 0.913	28.52 / 0.894
9-Syn	(-9, -4)	20.69 / 0.672	23.04 / 0.725	23.70 / 0.715	22.00 / 0.75	28.13 / 0.870

Table 3. Performance analysis under varying disparity ranges, best results highlighted in bold.

speed. In terms of training efficiency, LF-PyrNet is likewise efficient, requiring only 1 day of training, which is substantially faster than Zhang et al. ²⁶ (3 days), MANet ⁴¹ (9.7 days), and other efficient methods such as DeOccNet ²³ and Wang ²⁸ (2 days), demonstrating rapid convergence without sacrificing performance. While LF-PyrNet may not be the fastest model overall, its efficient design strikes an optimal balance between parameter efficiency and inference speed, making it ideal for real-world applications.

Performance analysis under varying disparity ranges

To further demonstrate the robustness of our approach, we evaluate performance under five disparity ranges using both single-image inpainting methods (RFR ⁴⁶, LBAM ⁴⁷) and LF-specific methods (DeOccNet ²³, Zhang et al. ²⁶). For consistency of evaluation, part of the comparative results are directly adopted from Zhang et al. ²⁶.

As shown in Table 3, our method consistently achieves superior performance across different LF types and disparity ranges. Specifically, our model attains the highest PSNR and SSIM in four out of five cases, with notable improvements in large disparity ranges such as (−9, −4), where it surpasses DeOccNet ²³ by more than 3.5 dB in 4-Syn and nearly 5 dB in 9-Syn settings. In contrast, inpainting methods (RFR ⁴⁶ and LBAM ⁴⁷) generate plausible semantics but fail to reconstruct complex occluded backgrounds, while LF-based methods DeOccNet ²³ and Zhang et al. ²⁶ underexploit angular correlations, leading to artifacts and degraded textures. Moreover, the sharp performance drop of DeOccNet ²³ and Zhang et al. ²⁶ cases highlights their limited generalization ability. By comparison, our method more effectively leverages spatial-angular consistency to recover fine textures, reduce artifacts, and preserve structural fidelity across diverse disparity ranges. These results confirm that our method scales more effectively to challenging disparity variations than both inpainting-based and LF-specific baselines.

Ablation study

The ablation study presented in Table 4 evaluates the impact of different network components on the performance of LF-PyrNet. While the full LF-PyrNet model generally achieves the highest performance across most cases, some ablated versions surpass it in specific scenarios, highlighting the exact role of individual components. Omitting the ResApp module leads to a consistent drop in performance across all scenarios. For instance, in the 9-Syn Sparse case, the PSNR decreases from 28.13 (full model) to 27.99, while SSIM reduces from 0.872 to 0.858, demonstrating the contribution of ResApp to improving feature extraction and reconstruction quality. Interestingly, removing the RFB slightly improves performance in the 9-Syn Sparse scenario, where the SSIM increases from 0.870 (LF-PyrNet) to 0.876 (w/o RFB). This indicates that RFB enhances global feature integration and helps extend the receptive field. However, in certain cases, removing it allows the network to focus on more localized details. This can lead to marginal gains. While RDB effectively reconstructs fine textures and preserves occlusion boundaries, improving detail restoration in most cases, its presence may introduce slight over-smoothing in heavily occluded dense scenarios. Excluding the RDB results in the highest SSIM for Dense Single Occlusion (0.832) and Dense Double Occlusion (0.832), surpassing the full model SSIM of 0.826 and 0.828, respectively.

LF Type	Sparse (syn)		Sparse (real)	Dense (syn)	
Name	4-Syn	9-Syn	CD	Single Occ	Double Occ
w/o ResASpp	26.42 / 0.854	27.69 / 0.865	25.01 / 0.864	29.33 / 0.815	27.23 / 0.812
w/o RFB	26.45 / 0.850	27.99 / 0.876	24.85 / 0.863	28.73 / 0.810	29.54 / 0.809
w/o RDB	26.24 / 0.845	26.77 / 0.847	25.05 / 0.867	29.00 / 0.832	27.60 / 0.830
w/o FPN	25.96 / 0.831	26.80 / 0.841	24.41 / 0.832	27.94 / 0.819	26.37 / 0.821
w/o Refinement	26.21 / 0.854	27.06 / 0.857	25.04 / 0.860	28.36 / 0.806	27.07 / 0.802
LF-PyrNet (Ours)	27.41 / 0.872	28.13 / 0.870	25.79 / 0.887	31.96 / 0.826	30.11 / 0.828

Table 4. Ablation Analysis: Highlighting the Best Performance in bold.



Fig. 11. Visual analysis of the ablation study, showcasing the influence of each model component on overall performance.

The most severe performance drop is observed when the FPN is removed, particularly in the CD and dense (Double Occ) cases, where the PSNR drops from 25.79 to 24.41 and from 30.11 to 26.37, respectively. This highlights the crucial role of FPN in multi-scale feature fusion, ensuring that both fine and coarse details are effectively captured. Without the refinement module, performance takes a sharp hit, especially in the Dense (Single and Double Occ) cases. In Single Occ, PSNR drops from 31.96 to 28.36, while in Double Occ, the decrease is even more pronounced, falling from 30.11 to 27.07. This decline in performance underscores the crucial role of the refinement module, particularly the SepConv block, which aids in enhancing texture restoration and accurately recovering occlusion boundaries. Without it, the model struggles to refine the fine details of occluded regions, leading to a diminished decline in overall quality. The visual results in Fig. 11 align with the quantitative findings, highlighting the significant enhancements achieved when the complete architecture is employed.

Conclusion and future work

This study presents LF-PyrNet, a new deep learning model designed to improve occlusion removal in LF images. By employing multi-scale receptive field learning through the integration of ResASPP and a modified RFB, LF-PyrNet expands the receptive field, capturing long-range dependencies and enhancing spatial feature extraction. The core network structure consists of three cascaded RDBs followed by an FPN, facilitating effective occlusion removal by progressively refining features at multiple scales. Additionally, our refinement module, which incorporates both separable and standard convolutions, ensures the restoration of even the finest details in occluded regions, enhancing the overall quality of the reconstruction. However, while LF-PyrNet performs well on sparse LF datasets, it faces challenges with denser datasets where occlusions are more complex. This limitation arises from the model's current inability to rely on mask prediction prior to inpainting occluded regions, which makes handling dense occlusions with high accuracy more difficult. Future work will focus on incorporating mask prediction mechanisms to guide the inpainting process, allowing the model to better isolate and reconstruct occluded regions, especially in dense LF datasets. Furthermore, advanced techniques for feature fusion will be explored to improve performance on large and complex occlusions. These enhancements will make the model more reliable and adaptable in real-world applications.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 11 August 2025; Accepted: 17 September 2025

Published online: 22 October 2025

References

- Senussi, M. F. et al. A comprehensive review on light field occlusion removal: Trends, challenges, and future directions. *IEEE Access* (2025).
- Joshi, N., Avidan, S., Matusik, W. & Kriegman, D. J. Synthetic aperture tracking: Tracking through occlusions. In *2007 IEEE 11th International Conference on Computer Vision*, 1–8 (IEEE, 2007).
- Ren, M., Liu, R., Hong, H., Ren, J. & Xiao, G. Fast object detection in light field imaging by integrating deep learning with defocusing. *Appl. Sci.* **7**, 1309 (2017).
- Yang, T. et al. All-in-focus synthetic aperture imaging. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI* **13**, 1–15 (Springer, 2014).
- Kasem, M. S. et al. HttD: A hierarchical transformer for accurate table detection in document images. *Mathematics* **13**, 266 (2025).
- Mahmoud, M. et al. Two-stage video violence detection framework using gmflow and cbam-enhanced resnet3d. *Mathematics* **13**, 1226 (2025).
- Abdalla, M. et al. Receiptqa: A question-answering dataset for receipt understanding. *Mathematics* **13**, 1760 (2025).
- Li, W. et al. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10758–10768 (2022).
- Phutke, S. S. & Murala, S. Nested deformable multi-head attention for facial image inpainting. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 6078–6087 (2023).
- Suvorov, R. et al. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159 (2022).
- Senussi, M. F., Abdalla, M., Kasem, M. S., Mahmoud, M. & Kang, H.-S. Spectral normalized u-net for light field occlusion removal. In *International conference on future information & communication engineering* **16**, 294–297 (2025).
- Han, K. *Light field reconstruction from multi-view images*. Ph.D. thesis, James Cook University (2022).
- Faraday, M. Liv. thoughts on ray-vibrations. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **28**, 345–350 (1846).
- Gershun, A. The light field. *J. Math. Phys.* **18**, 51–151 (1939).
- Adelson, E. H., Bergen, J. R. et al. *The plenoptic function and the elements of early vision*, vol. 2 (Vision and Modeling Group, Media Laboratory, Massachusetts Institute of ..., 1991).
- Wong, T.-T., Fu, C.-W., Heng, P.-A. & Leung, C.-S. The plenoptic illumination function. *IEEE Trans. Multimedia.* **4**, 361–371 (2002).
- León-López, K. M., Galvis-Carreño, L. V. & Arguello-Fuentes, H. Reconstruction of multispectral light field (5d plenoptic function) based on compressive sensing with colored coded apertures from 2d projections. *Revista Facultad de Ingeniería Universidad de Antioquia* 131–141 (2016).
- Levoy, M. & Hanrahan, P. Light field rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 441–452 (Association for Computing Machinery, 2023).
- Cutless, B. & Levoy, M. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 303–312 (1996).
- Hanrahan, M. L. P. Light field rendering. *SIGGRAPH96, Computer Graphics Proceeding* (1996).
- Chen, C.-C., Lu, Y.-C. & Su, M.-S. Light field based digital refocusing using a dsrl camera with a pinhole array mask. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 754–757 (IEEE, 2010).
- Dayan, S. B., Mendlovic, D. & Giryes, R. Deep sparse light field refocusing. *arXiv preprint arXiv:2009.02582* (2020).
- Wang, Y. et al. Deocnet: Learning to see through foreground occlusions in light fields. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 118–127 (2020).
- Li, Y. et al. Mask4d: 4d convolution network for light field occlusion removal. In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2480–2484 (IEEE, 2021).
- Pei, Z., Jin, M., Zhang, Y., Ma, M. & Yang, Y.-H. All-in-focus synthetic aperture imaging using generative adversarial network-based semantic inpainting. *Pattern Recognit.* **111**, 107669 (2021).
- Zhang, S., Shen, Z. & Lin, Y. Removing foreground occlusions in light field using micro-lens dynamic filter. In *IJCAI*, 1302–1308 (2021).
- Hur, J., Lee, J. Y., Choi, J. & Kim, J. I see-through you: A framework for removing foreground occlusion in both sparse and dense light field images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 229–238 (2023).
- Wang, X., Liu, J., Chen, S. & Wei, G. Effective light field de-occlusion network based on swin transformer. *IEEE Trans. Circuits Syst. Video Technol.* **33**, 2590–2599 (2022).
- Zhang, Q. et al. Swinccnet: Swin-unet encoder-decoder structured-light field occlusion removal network. *Opt. Eng.* **63**, 104102–104102 (2024).
- Senussi, M. F. & Kang, H.-S. Occlusion removal in light-field images using cspdarknet53 and bidirectional feature pyramid network: A multi-scale fusion-based approach. *Appl. Sci.* **14**, 9332 (2024).
- Vaish, V. et al. Synthetic aperture focusing using a shear-warp factorization of the viewing transform. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, 129–129 (IEEE, 2005).
- Vaish, V., Levoy, M., Szeliski, R., Zitnick, C. L. & Kang, S. B. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2, 2331–2338 (IEEE, 2006).
- Vaish, V., Wilburn, B., Joshi, N. & Levoy, M. 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.*, vol. 1, I–I (IEEE, 2004).
- Pei, Z., Zhang, Y., Chen, X. & Yang, Y.-H. Synthetic aperture imaging using pixel labeling via energy minimization. *Pattern Recognit.* **46**, 174–187 (2013).
- Pei, Z., Chen, X. & Yang, Y.-H. All-in-focus synthetic aperture imaging using image matting. *IEEE Trans. Circuits Syst. Video Technol.* **28**, 288–301 (2016).
- Xiao, Z., Si, L. & Zhou, G. Seeing beyond foreground occlusion: A joint framework for sap-based scene depth and appearance reconstruction. *IEEE J. Sel. Top. Signal Process.* **11**, 979–991 (2017).
- Zhang, S., Chen, Y., An, P., Huang, X. & Yang, C. Light field occlusion removal network via foreground location and background recovery. *Signal Process.: Image Commun.* **109**, 116853 (2022).
- Song, C., Li, W., Pi, X., Xiong, C. & Guo, X. A dual-pathways fusion network for seeing background objects in light field. In *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2022)*, vol. 12247, 339–348 (SPIE, 2022).
- Chen, J., An, P., Huang, X. & Yang, C. Adaptive threshold mask prediction and occlusion-aware convolution for foreground occlusions in light fields. In *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 1–5 (IEEE, 2024).
- Zhang, S., Chang, S., Shi, Z. & Lin, Y. Progressive multi-plane images construction for light field occlusion removal. *IEEE Transactions on Visualization and Computer Graphics* (2025).
- Chen, J. et al. Mask-aware light field de-occlusion with gated feature aggregation and texture-semantic attention. *IEEE Trans. Multimedia.* (2025).
- Piao, Y., Rong, Z., Xu, S., Zhang, M. & Lu, H. Dut-lfsaliency: Versatile dataset and light field-to-rgb saliency detection. *arXiv preprint arXiv:2012.15124* (2020).

43. Bok, Y., Jeon, H.-G. & Kweon, I. S. Geometric calibration of micro-lens-based light field cameras using line features. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 287–300 (2016).
44. Raj, A. S., Lowney, M. & Shah, R. *Light-field database creation and depth estimation* (Stanford University, Palo Alto, USA, 2016).
45. Rerabek, M. & Ebrahimi, T. New light field image dataset. In *8th international conference on quality of multimedia experience (QoMEX)* (2016).
46. Li, J., Wang, N., Zhang, L., Du, B. & Tao, D. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7760–7768 (2020).
47. Xie, C. et al. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8858–8867 (2019).

Author contributions

Conceptualization, M.F.S. and H.-S.K.; methodology, M.F.S. and M.A.; software, M.F.S. and M.A.; formal analysis, M.F.S. and M.S.K.; investigation, M.M. and H.-S.K.; resources, M.M. and H.-S.K.; data curation, M.F.S. and M.A.; writing-original draft preparation, M.F.S.; writing-review and editing, M.F.S., M.A., M.S.K., M.M., and H.-S.K.; validation, M.M. and H.-S.K.; visualization, M.M. and H.-S.K.; supervision, H.-S.K.; project administration, H.-S.K.; funding acquisition, H.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (RS-2023-NR076833), and partly by the Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government [Ministry of Science and ICT (MSIT)] (IITP-2025-RS-2020-II201462, 50%).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.-S.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025