applied sciences

MDPI

Article

Occlusion Removal in Light-Field Images Using CSPDarknet53 and Bidirectional Feature Pyramid Network: A Multi-Scale Fusion-Based Approach

Mostafa Farouk Senussi [1,2] and Hyun-Soo Kang [1,*]

[1] School of Information and Communication Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea; mostafa.farouk@aun.edu.eg

[2] Information Technology Department, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt

[*] Correspondence: hskang@cbnu.ac.kr

Abstract: Occlusion removal in light-field images remains a significant challenge, particularly when dealing with large occlusions. An architecture based on end-to-end learning is proposed to address this challenge that interactively combines CSPDarknet53 and the bidirectional feature pyramid network for efficient light-field occlusion removal. CSPDarknet53 acts as the backbone, providing robust and rich feature extraction across multiple scales, while the bidirectional feature pyramid network enhances comprehensive feature integration through an advanced multi-scale fusion mechanism. To preserve efficiency without sacrificing the quality of the extracted feature, our model uses separable convolutional blocks. A simple refinement module based on half-instance initialization blocks is integrated to explore the local details and global structures. The network's multi-perspective approach guarantees almost total occlusion removal, enabling it to handle occlusions of varying sizes or complexity. Numerous experiments were run on sparse and dense datasets with varying degrees of occlusion severity in order to assess the performance. Significant advancements over the current cutting-edge techniques are shown in the findings for the sparse dataset, while competitive results are obtained for the dense dataset.

Keywords: light-field images; occlusion removal; CSPDarknet53; bidirectional feature pyramid network (BiFPN); end-to-end learning; separable convolutional blocks; half-instance initialization network (HINet); multi-scale fusion; sparse datasets; dense datasets

check for updates

Citation: 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.* **2024**, *14*, 9332. https://doi.org/10.3390/app14209332

Academic Editor: Pedro Couto

Received: 10 September 2024
Revised: 7 October 2024
Accepted: 11 October 2024
Published: 13 October 2024

1. Introduction

Occlusion removal in light-field (LF) images is a pivotal task in the realm of computer vision, particularly for applications involving object detection, recognition, and tracking [1–10]. Occlusions can lead to significant degradation in performance by masking essential object features, resulting in misclassification, decreased accuracy, and unreliable tracking. This challenge is particularly pronounced in dynamic environments where occlusions frequently occur, making it imperative to develop effective occlusion removal techniques.

Different from a single-view image, which captures a snapshot of a scene from a fixed viewpoint, in which the occlusion removal task is known as single-image inpainting [11–13], the light-field nature captures not just the intensity of light, but also its directionality across the entire scene by using light-field camera arrays [14–18]. The cameras are arranged in the angular direction and each camera capture a single-view image, as illustrated in Figure 1. This creates richer information that allows for post-capture effects like refocusing [15,19,20], scene depth estimation [21–26] , angular and spatial super-resolution [27–33], saliency detection [34,35], deblurring [36], reconstruction [37], and view synthesis [38–41].

Employing light-field imaging in occlusion removal tasks proves advantageous due to its ability to capture multiple views of a scene, thereby mitigating occlusion issues encountered in conventional single-view imaging. This enables algorithms to reconstruct occluded areas by leveraging information from unoccluded views, where obscured pixels in one view remain visible in others. But the challenge becomes significantly pronounced when dealing with large and complex occlusions that obscure critical details. Traditional methods, often struggle with these scenarios due to their limited ability to capture extensive contextual information and integrate multi-scale features effectively.

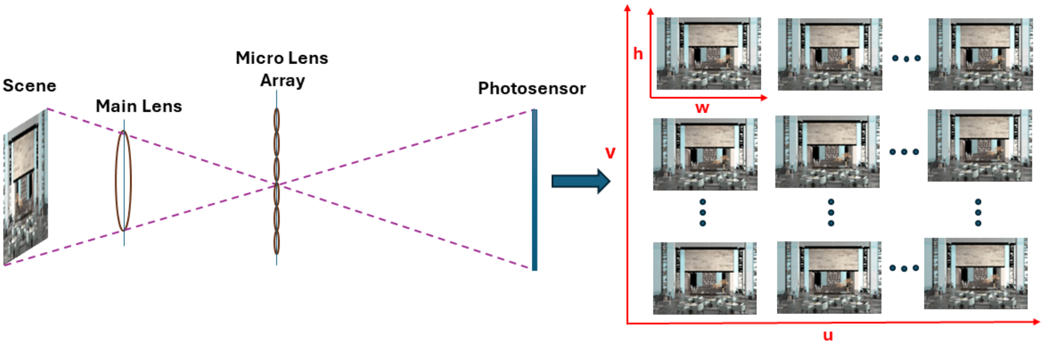

Figure 1. The left side shows a light-field camera system that focuses the scene through a microlens array onto a photosensor. The right side displays the resulting sub-aperture image grid, with u and v indicating angular resolution (different viewpoints) and w and h denoting spatial resolution within each sub-aperture image.

Recent advancements have sought to address these limitations by incorporating more sophisticated network architectures and multi-scale feature integration techniques. However, many of these methods still fall short when faced with large occlusions, primarily due to their inadequate handling of global and local feature dependencies. To overcome these challenges, a more holistic approach that combines robust feature extraction, efficient multi-scale fusion, and meticulous image refinement is essential. In response to these issues, we propose a novel architecture that synergistically combines CSPDarknet53 [42] and the bidirectional feature pyramid network (BiFPN) [43] for efficient light-field occlusion removal. CSPDarknet53, known for its robust and rich feature extraction capabilities across multiple scales, serves as the backbone of our network. It effectively captures detailed local features while maintaining a high degree of computational efficiency. To enhance the integration of these features, we incorporate BiFPN, which facilitates advanced multi-scale fusion and attention mechanisms. BiFPN's ability to aggregate and refine features across different scales ensures comprehensive integration, thereby improving the network's capability to handle occlusions of varying sizes and complexities. Additionally, we employ separable convolutional blocks within BiFPN to maintain computational efficiency without compromising the quality of feature extraction. To further refine the reconstructed images and address both local details and global structures, we integrate the Half-Instance Initialization Network (HINet) [44]. HINet meticulously refines the images, ensuring that even the smallest details are preserved while maintaining the integrity of the overall structure. This multi-faceted approach allows our network to effectively handle occlusions, providing a robust solution for comprehensive occlusion removal. We conducted many experiments on multiple sparse and dense datasets with different degrees of occlusion severity to verify the efficiency of the proposed approach. Our experimental results show significant improvements over the cutting-edge methods currently in use, with competitive quantitative metrics for PSNR and SSIM and useful effectiveness in real-world applications.

To summarize, our contributions are threefold:

1. We introduce a synergistic combination of CSPDarknet53 and BiFPN, enhanced with separable convolutional blocks, for robust and efficient feature extraction and multi-scale fusion.
2. The integration of HINet for meticulous image refinement ensures comprehensive occlusion removal, addressing both local and global features effectively.

3. Our method is extensively evaluated on diverse datasets, demonstrating significant improvements over state-of-the-art methods and practical effectiveness in real-world applications.

This comprehensive approach not only achieves superior quantitative results but also sets a new standard for future research in light-field occlusion removal. The source code will be made available upon publication, facilitating further research and development in this domain.

2. Related Work

Occlusion removal is a subset of image restoration that is focused on reconstructing the areas of an image that are blocked or missing. In this section, we briefly overview the related approaches for handling occlusion in single-view images, known as image inpainting, as well as techniques specifically designed for LF image occlusion removal. Table 1 provides a comparative summary that serves as a quick reference guide to the advancements in this field, highlighting their key features and limitations.

Table 1. Overview of methods for image inpainting and light-field occlusion removal, detailing their key features and limitations.

Name	Key Features	Limitations
	Image Inpainting	
Anisotropic Diffusion [45]	Effective for small occlusions	Produces over-blurred results for larger areas, sensitive to noise, and requires careful tuning of parameters.
PDEs [46]	Fills missing areas using PDEs	Struggles with large occlusions, less effective for complex textures, and computationally intensive for high resolutions.
PatchMatch [47]	Identifies and copies similar textures	Fails to create semantically coherent structures, limited by the availability of suitable patches, and can produce visible seams.
Partial Convolution [11]	Encodes contextual features	Limited by invalid pixel artifacts, may require careful initialization, and struggles with large occlusions.
RFR [48]	Hierarchical vector quantized VAE	Blurry results for large continuous holes, sensitive to the choice of hyperparameters, and can be slow in inference.
LBAM [49]	Attention mechanisms in encoder–decoder	Requires large datasets for training, potentially slow convergence, and struggles with complex occlusions due to semantic limitations.
Single Image Inpainting [50]	Propagates information to light-field views	Limited to central view inpainting, struggles with depth variability, and may not adequately capture occlusions from different angles.
	Light-Field Occlusion Removal	
Synthetic Aperture [51]	Resampling captured light	Struggles with large occlusions, limited in handling dynamic scenes, and requires extensive pre-processing.
Energy Minimization [52]	Distinguishes occlusion from background	Limited to specific depth ranges, sensitive to noise, and computationally expensive.
Layered Imaging [4]	Depth-independent all-in-focus imaging	Limited to specific scene types, requires accurate depth information, and struggles with occlusions in dense environments.
K-means Clustering [53]	Iterative reconstruction framework	Ineffective for complex structures, sensitive to initialization, and struggles with real-time applications.
DeOccNet [54]	End-to-end network with ASPP	Results often blurry, struggles with large occlusions, and requires substantial training data.
Mask4D [55]	4D convolution for spatial layout	Requires extensive computational resources, complex implementation, and potentially slow inference times.
GANs [56]	Synthesizes results with reconstructed backgrounds	Struggles with very large occlusions, sensitive to training instability, and requires large datasets.
Shifted Lenslet Filtering [57]	Extracts features from lenslet images	High memory and preprocessing requirements, struggles with dense light fields, and may fail in complex scenes.
ISTY Framework [58]	Modularizes feature extraction	Challenges with varying disparity ranges, requires careful parameter tuning, and still struggles with large occlusions.
Hybrid CNN-Transformer [59]	Combines CNNs and Transformers	Still performs poorly in big occlusions, complexity in training, and potential inefficiencies in real-time applications.
Ours	CSPDarknet53 with BiFPN + HINet Layer	Challenges with very complex occlusions, sensitivity to hyperparameters, and requires extensive fine-tuning.

2.1. Image Inpainting

Single-image inpainting aims to reconstruct the missing or masked regions in a single image with realistic content. Traditional techniques, such as anisotropic diffusion [45] and partial differential equations [46], have been effective for small occlusions but often produce over-blurred results when dealing with larger occlusions. Patch-based methods [47,60] have been developed to address this limitation by identifying and copying similar textures from other parts of the image. However, these methods can struggle with generating semantically coherent structures for larger missing areas. Recent advancements in deep learning algorithms, along with the availability of extensive datasets of single RGB images, have enabled the effective filling of these masked regions without needing information beyond the mask. Reference [11] introduced a specialized convolutional layer structure called Partial Convolution, which plays a significant role by encoding contextual features and avoiding artifacts from invalid pixels within the masked region through masked and re-normalized convolution. Building on this, ref. [48] introduced Recurrent Feature Reasoning (RFR), which is a two-stage model based on hierarchical vector-quantized variational auto-encoder for reconstructing large continuous holes by recurrently inpainting parts of the image and averaging the generated feature groups when they contain no invalid pixels. Another method, developed by [49], involves Learnable Bidirectional Attention Maps (LBAMs). This method replaces Partial Convolution with attention mechanisms in both the encoder and decoder, allowing the decoder to focus on filling only the masked regions. Unlike Partial Convolution, LBAM uses soft attention maps and differentiable mask updates, which provide greater flexibility and stability during model training.

Despite these improvements, single-image inpainting techniques often require large datasets for training and can struggle with reconstructing complex occlusions solely based on learned semantics. In the context of light-field imaging, single-image inpainting has been used to address light-field completion, which involves filling entire light-field views with consistent information. Instead of directly addressing the 4-D manifold, refs. [50,61] applied single-image inpainting to the central view (CV) image and propagated the information to the remaining views. Their approach used the single-image inpainting method for the CV image with a given inpainting mask, focusing on transferring the inpainted information to other views within the light field. Other methods have incorporated additional information such as edges and segmentation masks to improve the inpainting results.

2.2. Light-Field Occlusion Removal

LF imaging has gained traction due to the development of portable plenoptic cameras, which capture both the spatial and angular information of light rays, enabling a richer representation of scenes. This 4D light radiation field allows for the inference of depth information and reconstructing occluded objects more accurately than single-view methods. Early research leveraged the synthetic aperture focusing method, as proposed by [51], which involved resampling the captured light to blur the foreground and focus on the background. Despite enhancements to this approach, including alternative cost functions and dense multi-camera setups for recording light-field video, these methods struggled with large occlusions and lacked the ability to distinguish between foreground occlusion pixels and background occluded pixels. Ref. [52] addressed this limitation by labeling each pixel in every view through energy minimization to distinguish between occlusion and background. They later improved this technique by introducing a depth-free, all-in-focus synthetic aperture imaging method based on LF visibility analysis. However, these methods were limited to specific depth ranges, leaving objects at other depths blurred. Ref. [4] addressed this limitation by dividing scenes into visible layers, enabling depth-independent all-in-focus imaging. Ref. [53] developed an iterative reconstruction approach within a global optimization framework by using k-means clustering to classify occlusion and background pixels, refining results using a coarse-to-fine strategy.

Despite these advancements, traditional methods based on handcrafted features and stereo matching techniques struggled in scenes with complex structures and heavy occlusions, leading to the exploration of learning-based approaches. The advent of learning-based methods has significantly advanced LF occlusion removal. DeOccNet, proposed by [54], was the first end-to-end network for LF occlusion removal, utilizing a deep encoder–decoder model with residual atrous spatial pyramid pooling. This model also introduced a mask embedding approach to generate training datasets, allowing for effective supervised learning. However, DeOccNet's results were often blurry and struggled with large occlusions. Subsequent improvements came with Mask4D, presented by [55], which maintained the spatial layout of SAIs and employed 4D convolution to fully extract angular information, thereby improving de-occlusion performance. Ref. [56] leveraged Generative Adversarial Networks to semantically inpaint occluded regions, synthesizing the results with reconstructed backgrounds to produce occlusion-free images. Ref. [57] introduced a filter for extracting features from shifted lenslet images to reconstruct occluded regions. While effective for sparse LFs, this method faced challenges with dense LFs due to assumptions about background visibility and the high memory and preprocessing requirements.

To address both sparse and dense LF images, [58] introduced the ISTY framework for light-field de-occlusion, which modularizes the process into three distinct roles: extracting light-field features, defining occlusions, and inpainting occluded regions. The framework effectively manages the challenges posed by varying disparity ranges and enhances performance in both sparse and dense LF datasets. While these CNN-based methods showed promise, they were limited by the inherent local receptive field of CNNs, leading to incomplete occlusion removal and quality decline in complex scenes. To mitigite this limitation, a sophisticated approach proposed by [59] that synergistically combines Convolutional Neural Networks (CNNs) and Swin Transformers was developed to advance occlusion removal; CNNs were employed in the shallow layers to capture intricate local features and details, while Swin Transformers were used in the deeper layers to model the global patterns of large occlusions. The network leverages the global receptive field capabilities of Transformers to handle extensive occlusions, while CNNs mitigate the Transformers' challenges in fine-detail extraction. This hybrid architecture, integrating both global and local feature extraction, results in a comprehensive approach to restoring occlusion-free images. However, because of the way they handle local and global feature dependencies, many of these techniques still perform poorly in big occlusions. Robust feature extraction, effective multi-scale fusion, and careful image refining are some necessary components of a more comprehensive strategy to overcome these obstacles.

Unlike earlier techniques that relied on handcrafted features or limited CNN architectures, our approach synergizes the robust multi-scale feature extraction of CSPDarknet53 with the efficient multi-scale fusion capabilities of BiFPN, enhancing the network's ability to handle complex and large occlusions. Additionally, the incorporation of the Half-Instance Initialization Network blocks ensures meticulous image refinement, addressing both local and global structures more effectively than prior methods.

3. The Proposed Method

3.1. Architecture Overview

Our proposed architecture for light-field occlusion removal integrates CSPDarknet53 with the BiFPN and the HINet blocks, as detailed in Figure 2. The system comprises three key components: multi-scale feature extraction, multi-scale feature fusion, and image refinement sub-network. Reliable and multi-scale feature extraction is provided by the backbone, CSPDarknet53. This is supplemented by BiFPN, which improves feature integration via sophisticated multi-scale fusion and a swish attention mechanism. To further refine the occluded areas, the HINet block is utilized, focusing on both local details and global structures. The network's efficiency is maintained through separable convolutional blocks, ensuring high performance without compromising resources. We show significant improvements over current approaches, offering efficient occlusion removal across various scenarios. Because of this multi-perspective method, our network handles occlusions well and provides a reliable solution for accurate occlusion removal. The complete process can be summarized as

$$I_{\text{out}} = \tanh(\text{HINet}(\text{Conv}_{3\times3}(\text{BiFPN}(\text{CSPDarknet53}(L_0))_{X6}))) \tag{1}$$

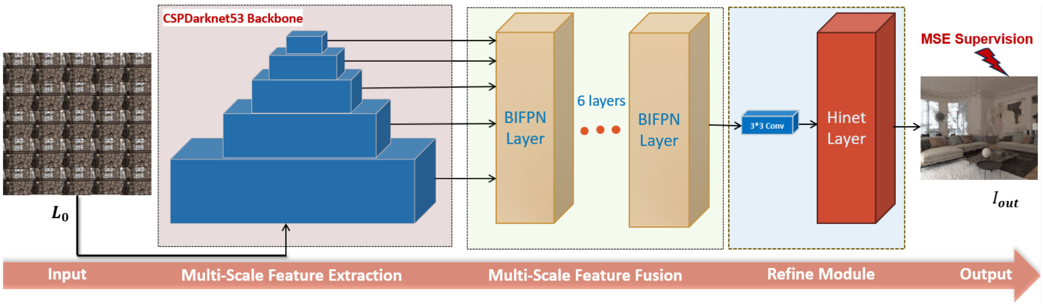

Figure 2. An overview of our three-part architecture: a multi-scale feature extraction, a multi-scale feature fusion, and an image refining subnetwork.

3.2. Multi-Scale Feature Extraction

The CSPDarknet53 architecture is utilized by the multi-scale feature extraction module to extract a wide range of detailed multi-scale features from the input light-field images. Because of its intricate architecture that uses cross-stage partial connections to enhance performance, CSPDarknet53 is well known for its effective and efficient feature extraction capabilities. The 5D tensor $L_0 \in \mathbb{R}^{U\times V\times H\times W\times C}$ is used to describe the input light-field images. The angular dimensions U and V denote the number of viewpoints captured within the light field, while the spatial dimensions H and W correspond to the height and width of each viewpoint, respectively. For this study, the spatial resolution is set at $H = 256$ and $W = 192$, with an angular resolution of 5×5. The number of channels C indicates the color depth or feature channels of the images. In order to increase the channel depth and extract earlier features that will enhance the feature representation, these images are stacked along the channel dimension, leading to a total size of $5 \times 5 \times 256 \times 192 \times 75$ and pass through an initial convolution layer:

$$F_{\text{in}} = \text{Conv}(L_0) \tag{2}$$

In this initial layer, a 3×3 convolution is employed with a stride of 1 and padding of 1, followed by a ReLU activation function. This configuration allows for a comprehensive extraction of spatial features while preserving the spatial dimensions of the input tensor.

The CSPDarknet53 architecture comprises Five CSPBlocks, each designed to encapsulate features at different sizes and degrees of abstraction. Within each block, downsampling procedures are employed to decrease spatial dimensions while enhancing feature depth. This downsampling is executed through convolutional operations with a kernel size of 3×3 and a stride of 2, effectively halving the spatial dimensions.

$$F_{k+1} = \text{Downsample}(\text{CSPBlock}_k(F_k)) \tag{3}$$

for $k \in \{0,1,2,3,4\}$, where $F_0 = F_{in}$ and k represents each block number.

Each CSPBlock begins by splitting the input tensor F_k into two parts, $\text{Split}_L(F_k)$ and $\text{Split}_R(F_k)$. This splitting is typically achieved by channel-wise division, where the number of channels in F_k is evenly divided between the left and right segments, and processes each part independently through a series of ResBlocks, where each ResBlock consists of two layers of convolutions to capture complex features effectively and aids in gradient flow efficiency. This method minimizes computation and aids in gradient flow efficiency. The first convolution is a 1×1 convolution followed by batch normalization and a ReLU activation function. The second is a 3×3 convolution with padding of 1, again followed by batch normalization and a ReLU activation function. Mathematically, the operations in a ResBlock can be described as follows:

$$F_{res} = \text{ReLU}(\text{BN}(\text{Conv}_{1\times 1}(F_k))) + F_k \tag{4}$$

$$F_{out} = \text{ReLU}(\text{BN}(\text{Conv}_{3\times 3}(F_{res}))) \tag{5}$$

The output from each ResBlock is then concatenated as follows:

$$F_{concat} = \text{Concat}(\text{ResBlock}(\text{Split}_L(F_k)), \text{ResBlock}(\text{Split}_R(F_k))) \tag{6}$$

After concatenation, the combined features are processed through a final 1 × 1 convolution, which adjusts the feature dimensionality and prepares it for subsequent processing stages:

$$F_{out} = \text{Conv}_{1\times1}(F_{concat}) \tag{7}$$

The output F_{out} from each of the five CSPBlocks represent the features captured at varying scales, effectively encompassing both fine and coarse details extracted from the light-field images. The multi-scale representation achieved through this architecture is critical for accurately modeling complex visual patterns, particularly in scenarios where nuanced feature representation is paramount. The combination of downsampling and residual connections facilitates the maintenance of spatial coherence within the extracted features while enabling deeper levels of feature abstraction. After traversing the CSPBlocks, the enriched feature representations are directed to the subsequent BIFPN module for integration and enhancement.

3.3. Multi-Scale Feature Fusion

The multi-scale features retrieved by the CSPDarknet53, symbolized as $\{P_i\}, i \in \{3, 4, 5, 6, 7\}$, where each P_i represents a feature map at a different scale, are integrated by the multi-scale feature fusion module using six BiFPN layers. In this case, the greatest resolution is represented by P_3 and the lowest by P_7. The BiFPN layer uses these features as input, processes them, and fuses them to improve object representation at various sizes. To strike a compromise between computing efficiency and semantic richness, the BiFPN begins at P_3 and concludes at P_7. Feature maps from P_3 to P_7 avoid the high computational costs associated with lower-level features (such as P_1 and P_2) while capturing the high-level semantic information required for tasks such as occlusion removal. This design optimizes overall performance and feature integration by aligning with the CSPDarknet53 backbone.

The feature maps P_3 to P_7 have the following dimensions:

$$P_3 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_3},\ P_4 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_4},\ P_5 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C_5},\ P_6 \in \mathbb{R}^{\frac{H}{64}\times\frac{W}{64}\times C_6},\ P_7 \in \mathbb{R}^{\frac{H}{128}\times\frac{W}{128}\times C_7}$$

where C_k represents the channel depth of each feature map. These multi-scale features are effectively fused across BiFPN layers to enhance the model's ability to perform occlusion removal, capturing both low-resolution and high-level semantic features, as well as finer details.

The BiFPN architecture as shown in Figure 3 consists of several key components: lateral connections, weighted bidirectional feature fusion process, and both bottom-up and top-down pathways to combine features from different scales. This bidirectional strategy gathers both high-level semantic information and low-level details, ensuring comprehensive feature fusion. In the lateral connection phase, each input feature map is processed through a 1 × 1 convolution to adjust its channel dimensions. This step is crucial for ensuring that feature maps from different scales can be effectively combined. The operation is defined as

$$P_{\text{lat}k} = \text{Conv}_{1\times1}(P_k)$$

This transformation enables the integration of features from different scales while preserving their spatial information.

In the bottom-up and top-down pathways, each BiFPN layer uses separable convolution blocks in both upward and downward paths to process and integrate features.

Top-Down: To match the resolution of higher-resolution maps, low-resolution feature maps are upsampled. The process is outlined below:

$$\text{Upsample}(P_i) \rightarrow \text{Convolution}_{\text{up}} \rightarrow \text{Feature}_{\text{up}}$$

where $\text{Convolution}_{\text{up}}$ is a separable convolution block applied after upsampling. Bottom-Up: In order to match the resolution of lower-resolution maps, high-resolution feature maps are downsampled. The process is outlined below:

$$\text{Downsample}(P_{\text{lat}k}) \rightarrow \text{Convolution}_{\text{down}} \rightarrow \text{Feature}_{\text{down}}$$

where $\text{Convolution}_{\text{down}}$ is a separable convolution block applied after downsampling.

Each separable convolution block consists of (1) depthwise convolution, which operates on each input channel separately; (2) pointwise convolution, which merges the depthwise convolution's outputs and (3) batch normalization and swish activation, which are applied to the outputs of the pointwise convolution to normalize and activate the features.

The output feature map F_{sep} from the separable convolution blocks is defined as

$$F_{\text{sep}} = \text{DepthwiseConv}(\text{PointwiseConv}([P_i])) \quad (8)$$

where $[P_i]$ are the input feature maps to the BiFPN layer.

Weighted Bidirectional Feature Fusion Process: To integrate features from different levels in an adaptable manner, weighted feature fusion is employed by the BiFPN layers, wherein the significance of every input feature map is determined by dynamically learned weights. In order to achieve efficient multi-scale feature integration, the feature maps are first pushed to normalized weights. The following is a representation of the adaptive multi-scale feature fusion process:

$$\text{P}_{\text{fused}} = \text{Conv}_{\text{fusion}}\left(\text{Swish}\left(\sum_i w_i \cdot \text{P}_i\right)\right) \quad (9)$$

While swish is an attention mechanism used to improve feature representation and dynamically weight the contributions of various feature levels, w_i represents the learned weight for each feature map p_i and $\text{Conv}_{\text{fusion}}$ is a convolution operation applied to the combined feature map. Several essential elements are involved in this process: while swish activation is performed on the weighted sum of features to add non-linearity and improve feature representation, ReLU activation is applied to the learnable weights before normalization.

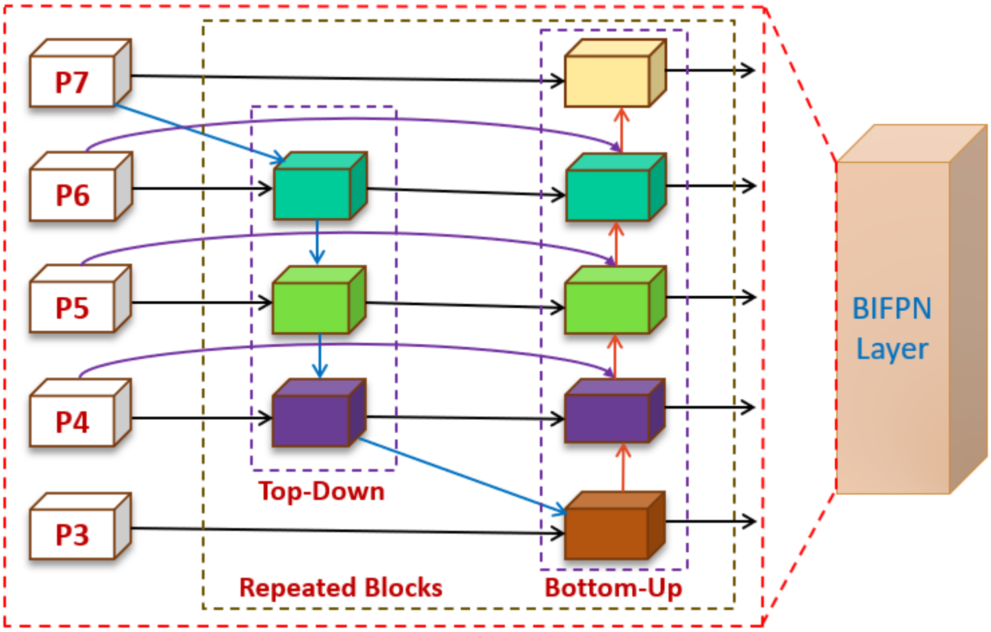

Figure 3. The structure of a bi-directional feature pyramid network layer.

The swish attention mechanism can be represented mathematically as follows:

$$P_{i,\text{up}} = \text{Swish}\Big(W_i^T \cdot P_i + W_i^{\text{up}} \cdot \text{Upsample}(P_{i+1})\Big) \tag{10}$$

$$P_{i,\text{down}} = \text{Swish}\Big(W_i^T \cdot P_i + W_i^{\text{down}} \cdot \text{Downsample}(P_{i-1})\Big) \tag{11}$$

For each feature map P_i, W_i^{up} and W_i^{down} indicate the weights applied after the upsampling and downsampling procedures, respectively. W_i^T denotes the weight matrix applied to the feature map. The procedures that modify the feature maps' resolution are $\text{Upsample}(P_{i+1})$ and $\text{Downsample}(P_{i-1})$. For each layer of the six cascaded BIFPN layers, this fusion procedure is performed for each of the five feature maps $[P_i]$. A set of refined feature maps at different scales is output at each BiFPN layer $\{P_3^{\text{fused}}, P_4^{\text{fused}}, P_5^{\text{fused}}, P_6^{\text{fused}}, P_7^{\text{fused}}\}$. The final output of the BiFPN is a concatenation of all fused feature maps across the levels, which can be summarized as

$$F_{\text{output}} = \text{Concat}(P_3^{\text{fused}}, P_4^{\text{fused}}, P_5^{\text{fused}}, P_6^{\text{fused}}, P_7^{\text{fused}})$$

Finally, we send $\{F_{\text{output}}\}$ to the refine module for further image refining at the final BiFPN layer to enhance the quality of the output by leveraging the most representative features captured during multi-scale feature fusion.

3.4. Refinement Module

To refine the image, the refine module starts with a 3 × 3 convolution, which cuts down the number of channels to 3. These channels are then fed directly to a single stage of the HINet network. The single-stage of the HINet as illustrate in Figure 4 , is a U-Net-shaped architecture that employs an encoder–decoder designed to enhance feature extraction and image reconstruction. On the encoder side, a downsampling is carried out with a convolution of kernel size 4. Half-Instance Normalization Blocks (HIN Blocks) are used to process features while downsampling the channels twice.

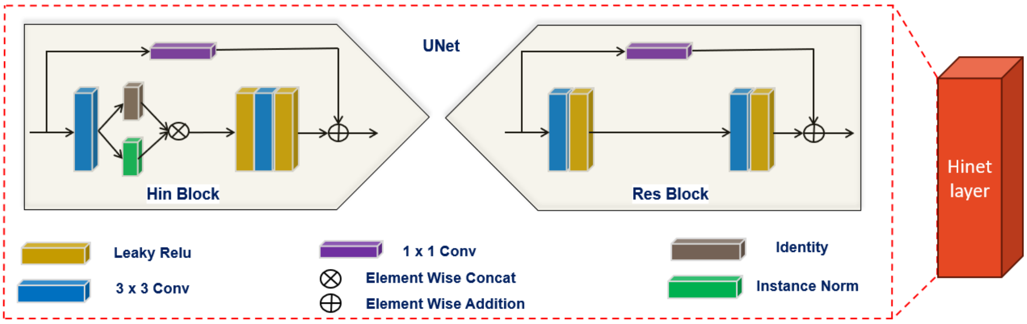

Figure 4. The structure of the Half-Instance Normalization Layer.

The HIN Block is essential for enhancing feature representation while maintaining efficiency. The input features $F_{\text{in}} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$ are processed and generate intermediate features $F_{\text{mid}} \in \mathbb{R}^{C_{\text{out}} \times H \times W}$, using a 3 × 3 convolutional layer. Intermediate features F_{mid} are split into two equal parts: F_{mid1} and F_{mid2}, each with $C_{\text{out}}/2$ channels. F_{mid1} undergoes instance normalization (IN) with learnable affine parameters. F_{mid2} remains unchanged by passing it through identity to retain contextual information. The concatenation of the two portions occurs over the channel dimension. The residual features R_{out} are obtained by processing the concatenated features via a 3 × 3 convolutional layer sandwiched between two leaky ReLU activations (with a leaky factor of 0.2). Taking the supplied features as input F_{in} for the convolutional layer of size 1, a shortcut link is made to them. By adding the shortcut features to the residual features R_{out}, the final output F_{out} is produced.

An upsampling operation was accomplished on the decoder side by using a kernel size of 2 in a transposed convolutional layer. To minimize information loss during resampling, features are refined using a single residual block [62]. The encoder's features are then combined with the upsampled features. After being processed via a 3 × 3 convolutional layer, the input features F_{in} with dimensions $\mathbb{R}^{C_{\text{in}} \times H \times W}$ are fed into the ResBlock via a leaky ReLU activation function with a leaky factor of 0.2. Introducing non-linearity enables the network to learn more intricate patterns. A second leaky ReLU activation is applied after the features flow through a third 3 × 3 convolutional layer. By collecting increasingly complex details and patterns in the incoming data, this further refines the features. To create a residual connection, the given features are used as input F_{in} for a convolutional layer of size 1, and the processed features from the second activation are summed to create the final output F_{out} of the ResBlock. The final occlusion-free light-field image is obtained by feeding the result of the refinement module through a Tanh activation function.

3.5. Loss Function

In our proposed architecture for light-field occlusion removal, Mean Squared Error (MSE) is employed as the primary loss function to optimize the performance of the network. The MSE loss function is defined as

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (I_{\text{out}}^{i} - I_{\text{gt}}^{i})^2, \tag{12}$$

where I_{out}^{i} represents the predicted pixel value at the i-th position and I_{gt}^{i} denotes the corresponding ground-truth pixel value, with N indicating the total number of pixels. The mean squared difference between true and anticipated pixel values is measured by this loss function, which gives a quantitative indicator of how well the occlusion removal occurred.

The integration of MSE within our architecture facilitates precise supervision during training and plays a crucial role in minimizing pixel-wise errors across various scales and levels of detail, directly contributing to the effectiveness of occlusion handling by penalizing deviations between the reconstructed and true images. This ensures that the model learns to produce accurate and high-quality results and enhances the practical applicability of the method in real-world scenarios.

4. Experiments

4.1. Experimental Setup

Using the methodology outlined in [58], we trained and tested our network as follows:

Training dataset: We train our LF occlusion removal network using a dataset that combines real-world occlusion scenarios with synthetic occlusion generation in a manner identical to the mask embedding method indicated by [54]. This method establishes occluded LF images by embedding occlusion masks into occlusion-free LF images, thereby simulating a range of occlusion conditions. A variety of disparity scenarios are represented in the LF images by the random placement of one to three occlusion masks during the mask-embedding procedure. In addition to the 80 mask images used in the [54] approach, 21 more thick and large real occlusion images were added to enhance the training dataset, tackling the challenge of removing large occlusions. We make sure that our LF images only include objects with negative disparity in order to provide ground-truth occlusion-free images.We used 1418 light-field images (out of 2957 total) from the DUTLF-V2 dataset [63], a dense LF dataset taken with the Lytro Illum camera [64], for this purpose. Our model is able to learn and generalize occlusion elimination over a wide range of real-world scenarios because of the precise selection and augmentation of the training data.

Testing dataset: We use four synthetic sparse LF scenes (4-Syn) and nine synthetic sparse LF scenes (9-Syn), which were created by [54,59], respectively, to assess our network's performance on sparse light fields (LFs). Because Stanford CD scene [65] is a real sparse LF image with ground truth accessible, we include it in our quantitative comparison. From the DUTLF-V2 test dataset [63], we chose 615 LFs for dense LFs and an extra 33 real occlusion images. For the purpose of evaluating multi-disparity occlusion scenarios, which we refer to as Single Occ and Double Occ, respectively, we employ a mask-embedding technique with a disparity range of [1, 4]. We employ several real-world sparse and dense occlusion LF scenes that are publicly available for qualitative comparison. A thorough assessment of our method's performance in sparse LF environments is given by the sparse LF dataset, which includes scenes taken by [59] and the Stanford CD scene [65]. The Stanford Lytro dataset [66] and EPFL-10 [67], both taken with the Lytro Illum camera, form the dense LF dataset. Our network's ability to handle intricate occlusions in dense LF environments is thoroughly evaluated because of the variety of occlusion scenarios and disparity levels provided by these datasets.

Training details: An angular and spatial resolution of (U × V × X × Y) = (9 × 9 × 600 × 400) is present in the LF images in the DUTLF-V2 dataset. The 300 × 200 spatial resolution is applied to the central 5 × 5 pictures for our needs. To achieve a resolution of (256 × 192), we arbitrarily center-crop and flip images horizontally during training. To implement occlusion embedding, we employ the mask-embedding technique to randomly pick, mix, and shuffle one or more masks in RGB images. We optimize our model using the ADAM optimizer with $(\beta_1, \beta_2) = (0.5, 0.9)$ and a batch size of 4. Because of the restricted GPU memory, the λ_1 and λ_2 parameter values are set to 0.01 and 120 , respectively. A learning rate of 0.001 is used initially and is halved every 150 epochs. Using the PyTorch framework, 500 epochs of training are completed in 20 h on a single 4090 Nvidia Geforce GPU.

4.2. Experimental Results

4.2.1. Quantitative Results

We assess how accurate the de-occluded images are quantitatively by comparing our model to cutting-edge LF occlusion removal techniques: DeOccNet [54], as well as the approaches given by Zhang et al. [57] and ISTY [58]. We also analyze the information obtained from different views in light fields (LFs) by contrasting our model with single-image inpainting techniques, namely, RFR [48] and LBAM [49]. We used the same learning approach and mask embedding for a fair comparison inside the dense LF dataset and trained DeOccNet from scratch on our dataset. We applied the ISTY [58] approach using the provided authors' weights. Because of the restricted nature of the training data for RFR [48] and LBAM [49], and because Zhang et al. [57] do not provide access to their source codes, the quantitative findings for all three approaches are derived directly from ISTY [58]. The quantitative findings are presented in Table 2. For assessing the image quality, we utilized PSNR and SSIM, two well-known metrics used in LF occlusion-removal studies.

Table 2. Quantitative comparison on the sparse and dense LF dataset using PSNR and SSIM. ↑ indicates that a higher result is better; red indicates the best result, while blue indicates the second-best result.

LF Type	Sparse (Syn)		Sparse (Real)	Dense (Syn)	
Name	4-Syn	9-Syn	CD	Single Occ	Double Occ
			PSNR ↑		
RFR [48]	19.89	20.69	21.13	26.28	23.25
LBAM [49]	21.11	23.04	21.56	27.92	24.83
DeOccNet [54]	23.74	23.70	22.70	28.67	25.85
Zhang et al. [57]	14.46	22.00	20.19	23.15	18.01
ISTY [58]	26.42	27.04	25.17	32.44	28.31
Ours	27.32	27.48	25.68	30.70	29.34
			SSIM ↑		
RFR [48]	0.668	0.672	0.646	0.867	0.801
LBAM [49]	0.677	0.725	0.803	0.899	0.827
DeOccNet [54]	0.701	0.715	0.741	0.914	0.847
Zhang et al. [57]	0.683	0.758	0.832	0.900	0.823
ISTY [58]	0.836	0.849	0.870	0.947	0.902
Ours	0.862	0.853	0.886	0.838	0.850

While DeOccNet [54] shows limited performance overall, it achieves decent results across both datasets. With the exception of single-disparity occlusions in sparse LFs, Zhang et al. [57] do not consistently produce improved results. Although ISTY [58] shows a reasonable inference time and practical applicability, the pre-trained inpainting models suffer from catastrophic forgetting, and the inpainting knowledge used is currently suboptimal due to limited training data. Furthermore, it requires more parameters for dense light fields (LFs) than for sparse ones, making it twice as large as DeOccNet [54]. The single-image inpainting techniques RFR [48] and LBAM [49] perform better on dense LF images because they are adept at handling occlusion-free scenes. But, since these models cannot use the background information from LFs, they perform worse on sparse LFs. As quantitative measures show, our proposed solution outperforms existing LF occlusion removal and inpainting models in sparse light fields (LFs) and performs competitively in dense LFs.

4.2.2. Qualitative Results

We show some qualitative comparisons of real-world sparse light fields (LFs) between our approach and other cutting-edge techniques in Figure 5. Scenes 3 and 4 feature fewer occlusions and simpler textures, whereas the publicly accessible real-world CD scene and synthetic scenes 1 and 2, which have significant and complex occlusions, are compared. Occlusion artifacts are preserved, and hazy outputs are produced around occlusions by DeOccNet [54]. This restriction shows that it cannot precisely remove occlusions and restore distinct features. In comparison to DeOccNet, ISTY [58] shows higher performance, handling occlusions better in the sparse, dense LFs. But there remains space for improvement, as indicated by the final images' perceived occlusion artifacts. In the sparse LF dataset, the strategy that we propose shows strong de-occlusion performance through using information about occluded objects from various viewpoints in addition to effectively differentiating occlusions from background parts. In comparison to both DeOccNet and ISTY, our approach performs better because of its sophisticated feature extraction and integration strategies, which also lead to fewer occlusion artifacts and more precise reconstructions.

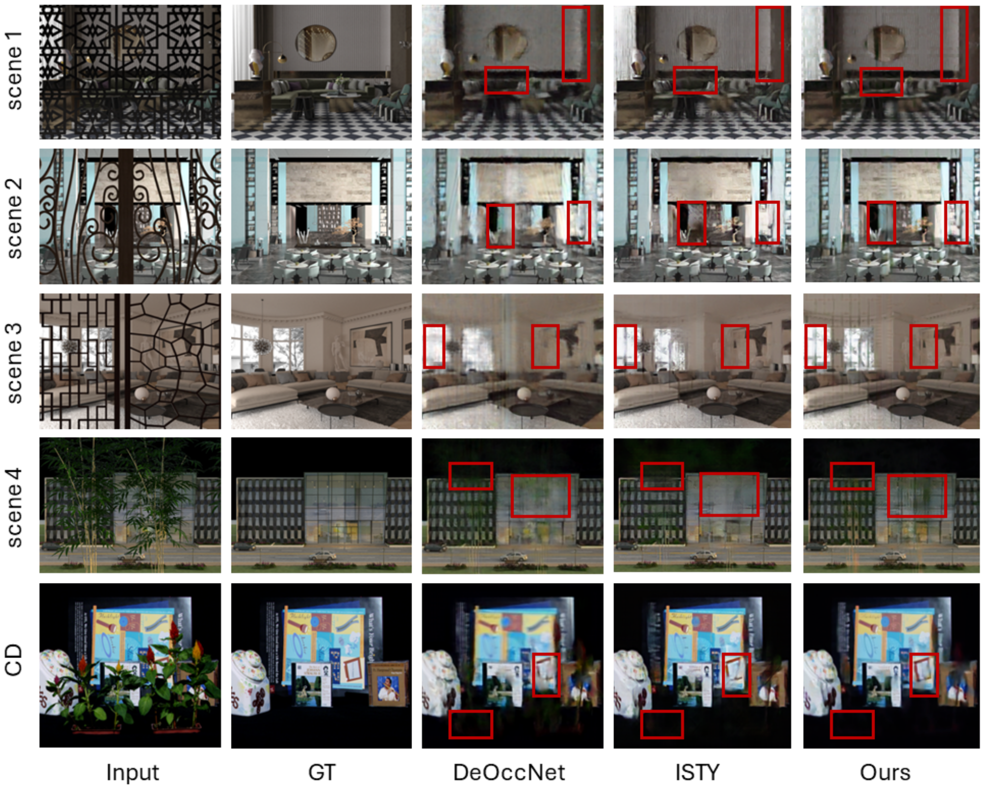

Figure 5. Qualitative comparisons of the sparse LF dataset. Selected areas of the outputs are highlighted with red boxes for detailed comparison. Our method reconstructs sharper occlusion-free CV images by effectively utilizing occluded background information from other views.

Figure 6's dense LF dataset illustrates how DeOccNet [54] struggles to handle dense and complicated occlusions, as seen by the outputs' frequent blurriness and retention of notable occlusion artifacts. When it comes to complex, multi-disparity occlusions, ISTY [58] performs better than DeOccNet and our approach, producing clearer outputs and fewer occlusion artifacts. Nevertheless, residual artifacts are still a problem for all of them. Although it falls short of ISTY, our suggested approach performs competitively in the dense LF dataset. This is mainly because our method does not include inpainting techniques, which are especially useful for dense LFs where inpainting expertise may greatly help reconstructing scenes without occlusion.

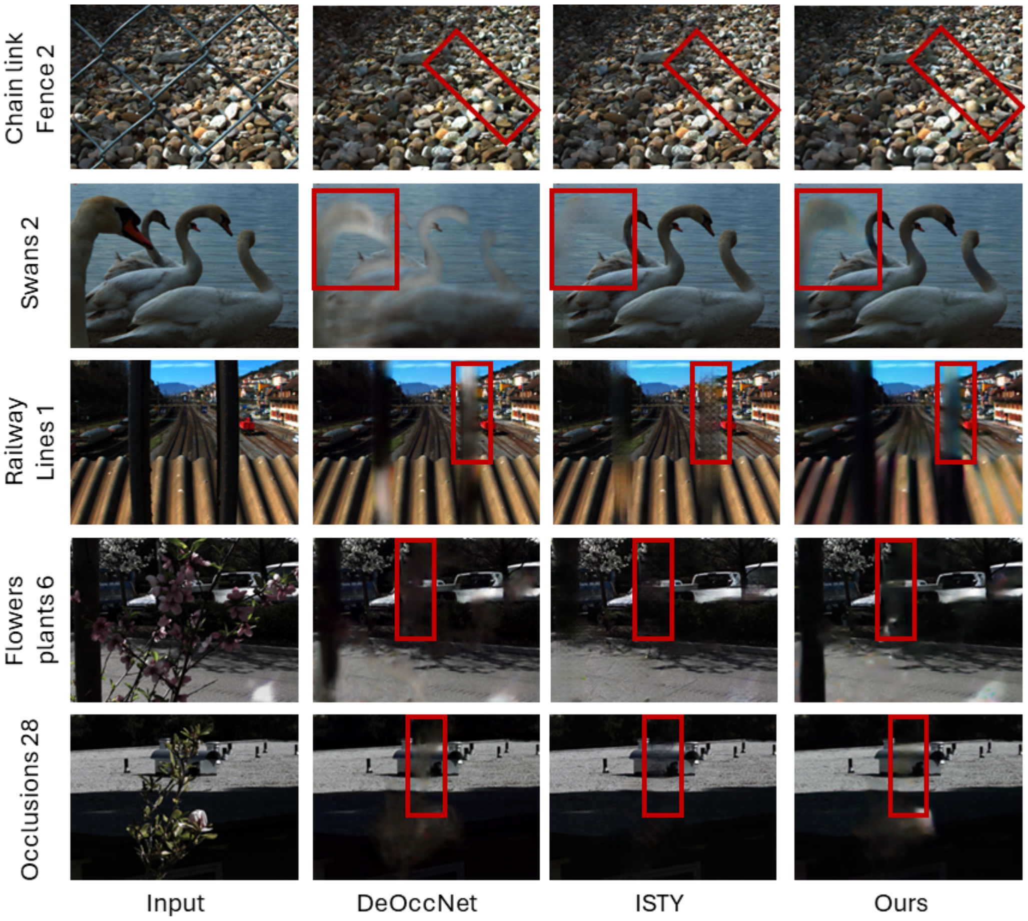

Figure 6. Qualitative comparisons on the dense LF dataset, with selected areas marked by red boxes for focused analysis, highlighting distinct differences among the methods.

4.2.3. Performance Evaluation on Real-World Scene Data

In Figure 7, we test our method on real-world scenes provided by Wang [54]. Our method demonstrates significant improvements in occlusion removal compared to existing methods such as DeOccNet and ISTY. The evaluation is based on two factors: (1) the preservation integrity of non-occluded areas, and (2) the restoration accuracy in occluded regions. Our method performs particularly well in scenes with thin and repetitive occlusions, while it exhibits limitations in handling larger, more irregular occlusions.

In the first row (bike_01), our method is particularly effective in removing thin occlusions, such as the basket grid in the foreground. The grid structure, which occludes part of the background (i.e., the cars in the scene), is accurately removed without introducing significant artifacts in the unoccluded areas. The red boxes highlight the areas of focus where our method shows clear improvements. The multi-scale feature extraction and refinement in the BiFPN allow for more accurate distinction between foreground occlusions and background content. Compared with DeOccNet, which loses important background information and generates blurred results, our method restores the background with a higher level of detail. This is supported by the higher PSNR (13.51) and SSIM (0.536) values, which indicate better recovery quality. Moreover, ISTY produces slightly better results than DeOccNet but still fails to maintain the clarity of finer details, particularly in regions with complex textures, such as the ground and parked vehicles.

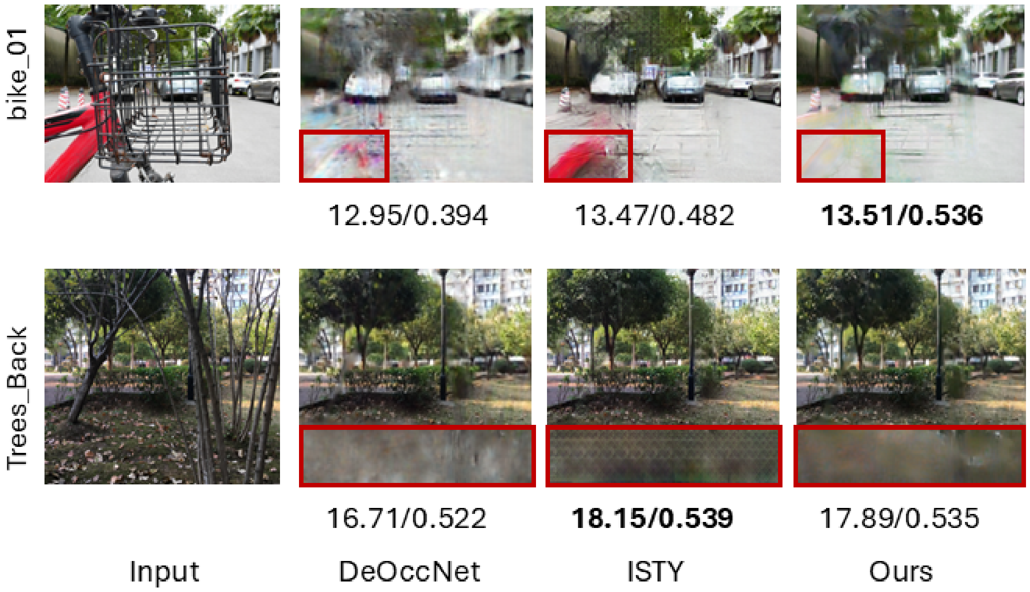

Figure 7. Analyzing occlusion removal effectiveness across real-world scenes. The first row (bike_01) highlights our method's success with thin occlusions, achieving higher PSNR and SSIM, as indicated by the red boxes. In contrast, the second row (Trees_Back) reveals limitations, as the method encounters challenges with larger, irregular occlusions, resulting in visible residues.

However, in the second row (Trees_Back), where the occlusions consist of large and irregular objects like tree trunks and branches, the performance of our method decreases. While our approach performs slightly better than both DeOccNet, the results are still not ideal. The occluded background, in this case, is partially recovered, but there are visible occlusion residues, as shown in the red box. The PSNR (17.89) and SSIM (0.535) values reflect this, indicating an improvement over one of the competing methods but still showing that the recovery quality is not optimal. Both our method and DeOccNet face challenges in dealing with dense occlusions, leading to blurred and incomplete reconstructions in the occluded regions. This limitation can be attributed to the complexity and size of the foreground occlusions. Large occlusions obscure significant portions of the scene, causing a serious loss of background information. Despite our method's multi-scale feature fusion strategy, which helps mitigate this issue to some extent, it still struggles to fully recover the background in heavily occluded areas. Future work may address this challenge by enhancing the model's global perception capabilities and potentially leveraging more comprehensive camera array setups for better occlusion removal.

4.2.4. Evaluation of Computational Efficiency and Scalability

Assessing computational efficiency and scalability is crucial for evaluating the performance of occlusion removal methods. Our proposed model, as detailed in Table 3, demonstrates a balance between complexity and performance, though certain trade-offs are evident. With 52.59 M parameters, our model is larger than Zhang et al. (2.7 M) and DeOccNet (39.0 M), while being smaller than larger architectures like LBAM (69.3 M) and ISTY (80.6 M). Notably, despite having more parameters than DeOccNet and Zhang et al., our model is still relatively efficient compared to LBAM and ISTY, indicating an optimized approach that achieves comparable or better performance with fewer parameters. In terms of inference time, although models such as DeOccNet (10 ms), LBAM (12 ms), and ISTY (24 ms) process faster, this speed often comes at the expense of reduced feature extraction capacity or lower performance. Our model, with an average inference time of 138.8 ms, strikes a balance between lightweight models like DeOccNet and LBAM and computationally heavier ones like Zhang et al. (3050 ms). This demonstrates that, while not the fastest, our model's design enables comprehensive multi-scale feature extraction, enhancing occlusion removal without overwhelming computational resources.

Table 3. Overview of each model's parameters and average inference time (Inf) for processing 256 × 192 light-field images on an Nvidia Geforce RTX 4090 GPU. ↓ indicates that smaller results are better.

Model	LBAM [49]	DeOccNet [54]	Zhang et al. [57]	ISTY [58]	Ours
Params ↓	69.3 M	39.0 M	2.7 M	80.6 M	52.59 M
Inf ↓	12 ms	10 ms	3050 ms	24 ms	138.8 ms

Regarding scalability, the model's computational demand (138.8 ms) and parameter count (52.59 M) suggest that it remains scalable for higher-resolution tasks or larger datasets without overburdening the system. This is crucial for applications that require a balance between accuracy and performance, particularly in real-time or resource-constrained environments. Although our method introduces a higher computational load compared to simpler architectures, its multi-scale processing capabilities and refinement layer contribute to superior occlusion reconstruction. Overall, the added complexity from CSPDarknet53, six BiFPN layers, and the HiNet refinement module enhances the model's capacity to handle detailed and high-resolution inputs. This leads to superior occlusion region reconstruction, justifying the marginal increase in computational cost. Importantly, the model remains scalable and efficient for more demanding tasks, striking a balance between accuracy and reasonable resource usage.

4.3. Ablation Study

In this section, we conduct a thorough ablation study to assess the contributions of various components in our proposed architecture, as illustrated in Table 4 and Figure 8. Our ablation study aims to identify the optimal number of BiFPN layers required for model convergence and evaluate the effects of selectively omitting key modules, including the complete removal of the BiFPN layers and the exclusion of the HiNet refinement module on the network's overall performance in light-field occlusion removal. The network is retrained using the same training data.

Our study begins by evaluating the model's performance across configurations with 3, 5, 6 (our baseline), and 7 BiFPN layers, as well as a variant that omits all BiFPN layers. The baseline model, featuring 6 BiFPN layers, consistently outperformed the others, achieving the highest performance metrics. As marked in red in the table, it attained an average PSNR of 28.104 dB and an SSIM of 0.858. This result underscores the model's effective balance between complexity and feature extraction, enabling it to capture rich, multi-scale contextual information crucial for accurately reconstructing occluded regions. Reducing the number of BiFPN layers to 3 or 5 resulted in a noticeable decline in performance metrics. This decrease can be attributed to the model's reduced capacity to capture rich contextual features, as fewer layers limited its ability to aggregate multi-scale information. The lack of sufficient multi-scale representations led to inadequate feature aggregation and impaired the model's ability to discern occluded regions effectively. Interestingly, increasing the number of BiFPN layers to 7 did not yield a proportional improvement in performance. This suggests that, while additional layers can enhance feature interactions, they may also introduce complexity and noise, resulting in diminishing returns and potential overfitting. This highlights the importance of balancing model depth with effective generalization across diverse input conditions.

We also investigated the contribution of the HiNet refinement module by training a version of our model without it. The exclusion of the HiNet layer led to a significant decrease in both PSNR and SSIM, emphasizing its critical role in enhancing detail and texture in the output. The HiNet layer enables selective refinement of feature maps produced by the preceding BiFPN layers, particularly enhancing intricate textures and edges, which significantly improves the clarity of reconstructed occluded regions. In summary, our ablation study demonstrates that the baseline model with 6 BiFPN layers is pivotal for achieving optimal performance in occlusion removal. Both the appropriate number of BiFPN layers and the inclusion of the HiNet refinement module are crucial for the model's effectiveness in reconstructing occluded regions, validating our architectural design choices.

Table 4. Conducting ablation studies on our method and its variants across sparse and dense light-field datasets. Performance is evaluated using PSNR and SSIM (PSNR/SSIM), with top results marked in red. ↑ indicates that a higher result is better.

LF Type	Sparse (Syn)		Sparse (Real)	Dense (Syn)	
Name	**4-Syn**	**9-Syn**	**CD**	**Single Occ**	**Double Occ**
			PSNR ↑		
Ours w/o BiFPN	20.04	20.56	20.28	28.31	27.62
Ours w/ 3 BiFPN Layers	26.73	26.99	24.90	26.36	27.06
Ours w/ 5 BiFPN Layers	26.47	27.27	25.17	30.59	29.25
Ours w/ 7 BiFPN Layers	24.59	25.42	23.71	29.05	27.56
Ours w/o Refinement	26.53	27.16	24.85	27.84	26.51
Ours	27.32	27.48	25.68	30.70	29.34
			SSIM ↑		
Ours w/o BiFPN	0.547	0.534	0.701	0.789	0.812
Ours w/ 3 BiFPN Layers	0.828	0.841	0.833	0.795	0.822
Ours w/ 5 BiFPN Layers	0.832	0.849	0.863	0.835	0.847
Ours w/ 7 BiFPN Layers	0.748	0.784	0.805	0.830	0.835
Ours w/o Refinement	0.829	0.838	0.849	0.824	0.845
Ours	0.862	0.853	0.886	0.838	0.850

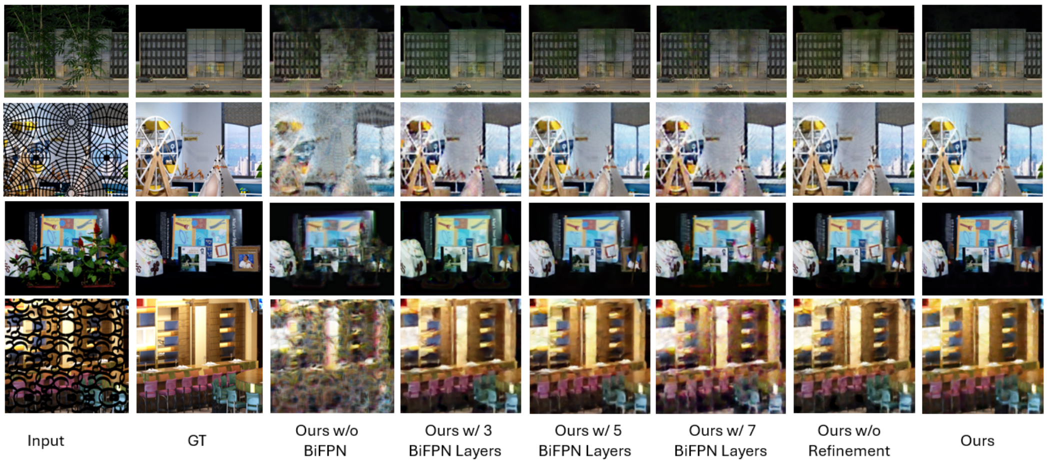

Figure 8. A visual exploration of model elements: unveiling the role of each component in our ablation study.

5. Limitations and Future Directions

While the proposed architecture represents a significant advancement in occlusion removal for light-field images, several limitations must be addressed to further enhance its robustness and practical applicability, particularly in handling large foreground occlusions that obscure significant scene areas, leading to background information loss. Although the multi-scale feature fusion strategy offers some mitigation, background recovery in heavily occluded regions remains inadequate. Future work should enhance global context awareness and advanced depth estimation techniques, as well as explore more extensive camera array configurations to improve recovery in complex scenes.

One limitation is the high computational complexity associated with the architecture, which can hinder deployment in resource-constrained environments. Future work should focus on reducing the number of network parameters without compromising performance. Techniques such as network pruning, quantization, and knowledge distillation could be employed to compress the model. Additionally, exploring lightweight backbone networks architectures could yield a more efficient configuration. Another challenge is the limited generalization of the model across complex occlusion scenarios. The model may struggle with unseen occlusion patterns that were not well represented in the training dataset. To address this, expanding the training datasets to include a wider range of real-world occlusions is essential. Incorporating self-supervised or unsupervised learning approaches could further enhance the model's ability to learn from unlabelled data. Moreover, the model may struggle to handle large disparities effectively.

In cases where occlusions span multiple depth planes, the current architecture may not adequately capture the necessary context. Introducing 3D or volumetric approaches that account for spatial relationships within the light-field data could significantly enhance the model's ability to manage large disparities. Finally, while the model excels at removing occlusions, it does not currently incorporate inpainting techniques to fill in areas left by these removals. This can lead to incomplete reconstructions in certain cases. Future work should explore integrating inpainting mechanisms or generative adversarial networks to generate plausible content in occluded regions, thereby enhancing the completeness and quality of the output. In summary, while this architecture marks a significant advancement in occlusion removal, addressing these limitations will improve its robustness and practical applicability in diverse real-world scenarios.

6. Conclusions

For efficient occlusion removal in light-field images, we have presented a comprehensive architecture in this work that combines CSPDarknet53 and the BiFPN. Using sophisticated feature extraction and multi-scale fusion techniques, this architecture tackles the significant difficulty of handling large and complicated occlusions. Robust feature extraction is provided by CSPDarknet53, while feature integration is improved by BiFPN. Efficiency is guaranteed without sacrificing feature extraction quality because of the incorporation of separable convolutional blocks. Furthermore, thorough image refining is made possible by the HINet, which efficiently addresses both local and global details. The network can thoroughly handle occlusions of different sizes and complexity thanks to this multi-perspective method. The originality of the proposed method lies in its efficient multi-scale feature extraction through CSPDarknet53, which reduces computational overhead while effectively managing occlusions. The dynamic feature fusion introduced by BiFPN ensures that both fine and coarse details are preserved, even in large occlusions. Additionally, the integration of HINet provides meticulous refinement by addressing local occlusion details and global structures, leading to smoother and more accurate reconstructions. The use of separable convolutions also enhances computational efficiency without sacrificing performance, making the method scalable for larger datasets and applicable in real-world scenarios. Numerous studies on diverse datasets with different degrees of occlusion severity show notable gains over cutting-edge techniques. Overall, the proposed architecture represents a significant advancement in occlusion removal, enhancing the accuracy of light-field imaging. By effectively addressing complex occlusions, it offers promising applications across various fields and sets the stage for further development in real-world scenarios.

Author Contributions: Conceptualization, M.F.S.; Methodology, M.F.S.; Software, M.F.S.; Validation, M.F.S. and H.-S.K.; Formal analysis, M.F.S.; Investigation, H.-S.K.; Resources, H.-S.K.; Data curation, M.F.S.; Writing—original draft, M.F.S.; Writing—review and editing, M.F.S. and H.-S.K.; Visualization, M.F.S. 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: The Basic Science Research Program partially supported this research through the National Research Foundation of Korea (NRF), funded by the Ministry of Education under Grant 2020R1I1A3A04037680, and partly by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2024-2020-0-01462, 30%).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Joshi, N.; Avidan, S.; Matusik, W.; Kriegman, D.J. Synthetic aperture tracking: Tracking through occlusions. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
2. 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.* **2017**, *7*, 1309. [CrossRef]
3. Yang, T.; Zhang, Y.; Tong, X.; Zhang, X.; Yu, R. A new hybrid synthetic aperture imaging model for tracking and seeing people through occlusion. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 1461–1475. [CrossRef]
4. Yang, T.; Zhang, Y.; Yu, J.; Li, J.; Ma, W.; Tong, X.; Yu, R.; Ran, L. All-in-focus synthetic aperture imaging. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part VI 13; Springer International Publishing: Cham, Switzerland, 2014; pp. 1–15.
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
6. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Online, 31 October 2017; pp. 7263–7271.
7. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
8. Kasem, M.; Abdallah, A.; Berendeyev, A.; Elkady, E.; Mahmoud, M.; Abdalla, M.; Hamada, M.; Vascon, S.; Nurseitov, D.; Taj-Eddin, I. Deep learning for table detection and structure recognition: A survey. *Acm Comput. Surv.* **2022**, *56*, 305.
9. Kasem, M.S.; Mahmoud, M.; Kang, H.S. Advancements and Challenges in Arabic Optical Character Recognition: A Comprehensive Survey. *arXiv* **2023**, arXiv:2312.11812.
10. Lin, Y.; Xie, Z.; Chen, T.; Cheng, X.; Wen, H. Image privacy protection scheme based on high-quality reconstruction DCT compression and nonlinear dynamics. *Expert Syst. Appl.* **2024**, *257*, 124891. [CrossRef]
11. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 85–100.
12. Mahmoud, M.; Kang, H.S. Ganmasker: A two-stage generative adversarial network for high-quality face mask removal. *Sensors* **2023**, *23*, 7094. [CrossRef]
13. Mahmoud, M.; Kasem, M.S.; Kang, H.S. A Comprehensive Survey of Masked Faces: Recognition, Detection, and Unmasking. *arXiv* **2024**, arXiv:2405.05900. [CrossRef]
14. Lin, X.; Wu, J.; Zheng, G.; Dai, Q. Camera array based light field microscopy. *Biomed. Opt. Express* **2015**, *6*, 3179–3189. [CrossRef]
15. 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, Washington, DC, USA, 27 June–2 July 2004; Volume 1, p. I.
16. Venkataraman, K.; Lelescu, D.; Duparré, J.; McMahon, A.; Molina, G.; Chatterjee, P.; Mullis, R.; Nayar, S. Picam: An ultra-thin high performance monolithic camera array. *ACM Trans. Graph. (TOG)* **2013**, *32*, 166. [CrossRef]
17. Wilburn, B.; Joshi, N.; Vaish, V.; Levoy, M.; Horowitz, M. High-speed videography using a dense camera array. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, Washington, DC, USA, 27 June–2 July 2004; Volume 2, p. II.
18. Wilburn, B.; Joshi, N.; Vaish, V.; Talvala, E.V.; Antunez, E.; Barth, A.; Adams, A.; Horowitz, M.; Levoy, M. High performance imaging using large camera arrays. *ACM Trans. Graph.* **2005**, *24*, 765–776. [CrossRef]
19. Ng, R.; Levoy, M.; Brédif, M.; Duval, G.; Horowitz, M.; Hanrahan, P. Light Field Photography with a Hand-Held Plenoptic Camera. Ph.D. Thesis, Stanford University, Stanford, CA, USA, 2005.

20. Wang, Y.; Yang, J.; Guo, Y.; Xiao, C.; An, W. Selective light field refocusing for camera arrays using bokeh rendering and superresolution. *IEEE Signal Process. Lett.* **2018**, *26*, 204–208. [CrossRef]
21. Lee, J.Y.; Park, R.H. Complex-valued disparity: Unified depth model of depth from stereo, depth from focus, and depth from defocus based on the light field gradient. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 830–841. [CrossRef]
22. Zhou, W.; Zhou, E.; Liu, G.; Lin, L.; Lumsdaine, A. Unsupervised monocular depth estimation from light field image. *IEEE Trans. Image Process.* **2019**, *29*, 1606–1617. [CrossRef]
23. Peng, J.; Xiong, Z.; Liu, D.; Chen, X. Unsupervised depth estimation from light field using a convolutional neural network. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 295–303.
24. Shin, C.; Jeon, H.G.; Yoon, Y.; Kweon, I.S.; Kim, S.J. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Online, 1 July 2018; pp. 4748–4757.
25. Tsai, Y.J.; Liu, Y.L.; Ouhyoung, M.; Chuang, Y.Y. Attention-based view selection networks for light-field disparity estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12095–12103.
26. Schilling, H.; Diebold, M.; Rother, C.; Jähne, B. Trust your model: Light field depth estimation with inline occlusion handling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Online, 1 July 2018; pp. 4530–4538.
27. Jin, J.; Hou, J.; Chen, J.; Kwong, S. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 24 June 2020; pp. 2260–2269.
28. Liu, D.; Wu, Q.; Huang, Y.; Huang, X.; An, P. Learning from EPI-volume-stack for light field image angular super-resolution. *Signal Process. Image Commun.* **2021**, *97*, 116353. [CrossRef]
29. Wang, Y.; Liu, F.; Zhang, K.; Hou, G.; Sun, Z.; Tan, T. LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Trans. Image Process.* **2018**, *27*, 4274–4286. [CrossRef] [PubMed]
30. Yeung, H.W.F.; Hou, J.; Chen, X.; Chen, J.; Chen, Z.; Chung, Y.Y. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Trans. Image Process.* **2018**, *28*, 2319–2330. [CrossRef]
31. Zhang, S.; Lin, Y.; Sheng, H. Residual networks for light field image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 26 November 2019; pp. 11046–11055.
32. Salem, A.; Ibrahem, H.; Kang, H.S. Learning epipolar-spatial relationship for light field image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 27 June 2023; pp. 1336–1345.
33. Salem, A.; Ibrahem, H.; Kang, H.S. Light Field Image Super-Resolution Using Deep Residual Networks on Lenslet Images. *Sensors* **2023**, *23*, 2018. [CrossRef]
34. Zhang, J.; Liu, Y.; Zhang, S.; Poppe, R.; Wang, M. Light field saliency detection with deep convolutional networks. *IEEE Trans. Image Process.* **2020**, *29*, 4421–4434. [CrossRef]
35. Zhang, M.; Ji, W.; Piao, Y.; Li, J.; Zhang, Y.; Xu, S.; Lu, H. LFNet: Light field fusion network for salient object detection. *IEEE Trans. Image Process.* **2020**, *29*, 6276–6287. [CrossRef]
36. Lumentut, J.S.; Kim, T.H.; Ramamoorthi, R.; Park, I.K. Deep recurrent network for fast and full-resolution light field deblurring. *IEEE Signal Process. Lett.* **2019**, *26*, 1788–1792. [CrossRef]
37. Salem, A.; Elkady, E.; Ibrahem, H.; Suh, J.W.; Kang, H.S. Light Field Reconstruction with Dual Features Extraction and Macro-Pixel Upsampling. *IEEE Access* **2024**, *12*, 121624–121634. [CrossRef]
38. Wang, Y.; Liu, F.; Wang, Z.; Hou, G.; Sun, Z.; Tan, T. End-to-end view synthesis for light field imaging with pseudo 4DCNN. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 333–348.
39. Wu, G.; Liu, Y.; Fang, L.; Dai, Q.; Chai, T. Light field reconstruction using convolutional network on EPI and extended applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1681–1694. [CrossRef]
40. Wu, G.; Liu, Y.; Dai, Q.; Chai, T. Learning sheared EPI structure for light field reconstruction. *IEEE Trans. Image Process.* **2019**, *28*, 3261–3273. [CrossRef]
41. Yagoub, B.; Kasem, M.S.; Kang, H.S. Enhancing X-ray Security Image Synthesis: Advanced Generative Models and Innovative Data Augmentation Techniques. *Appl. Sci.* **2024**, *14*, 3961. [CrossRef]
42. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Online, 24 June 2020; pp. 390–391.
43. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 24 June 2020; pp. 10781–10790.
44. Chen, L.; Lu, X.; Zhang, J.; Chu, X.; Chen, C. Hinet: Half instance normalization network for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 18 October 2021; pp. 182–192.
45. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000; pp. 417–424.
46. Ballester, C.; Bertalmio, M.; Caselles, V.; Sapiro, G.; Verdera, J. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Process.* **2001**, *10*, 1200–1211. [CrossRef]

47. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **2009**, *28*, 24. [CrossRef]
48. 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, Seattle, WA, USA, 13–19 June 2020; pp. 7760–7768.
49. Xie, C.; Liu, S.; Li, C.; Cheng, M.M.; Zuo, W.; Liu, X.; Wen, S.; Ding, E. Image inpainting with learnable bidirectional attention maps. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8858–8867.
50. Zhang, F.L.; Wang, J.; Shechtman, E.; Zhou, Z.Y.; Shi, J.X.; Hu, S.M. Plenopatch: Patch-based plenoptic image manipulation. *IEEE Trans. Vis. Comput. Graph.* **2016**, *23*, 1561–1573. [CrossRef]
51. Vaish, V.; Garg, G.; Talvala, E.; Antunez, E.; Wilburn, B.; Horowitz, M.; Levoy, M. Synthetic aperture focusing using a shear-warp factorization of the viewing transform. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops, San Diego, CA, USA, 21–23 September 2005; pp. 129–129.
52. Pei, Z.; Zhang, Y.; Chen, X.; Yang, Y.H. Synthetic aperture imaging using pixel labeling via energy minimization. *Pattern Recognit.* **2013**, *46*, 174–187. [CrossRef]
53. 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.* **2017**, *11*, 979–991. [CrossRef]
54. Wang, Y.; Wu, T.; Yang, J.; Wang, L.; An, W.; Guo, Y. DeOccNet: Learning to see through foreground occlusions in light fields. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 118–127.
55. Li, Y.; Yang, W.; Xu, Z.; Chen, Z.; Shi, Z.; Zhang, Y.; Huang, L. Mask4D: 4D convolution network for light field occlusion removal. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2480–2484.
56. 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.* **2021**, *111*, 107669. [CrossRef]
57. Zhang, S.; Shen, Z.; Lin, Y. Removing Foreground Occlusions in Light Field using Micro-lens Dynamic Filter. In Proceedings of the IJCAI, Montreal, QC, Canada, 19–27 August 2021; pp. 1302–1308.
58. 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, Waikoloa, HI, USA, 2–7 January 2023; pp. 229–238.
59. Wang, X.; Liu, J.; Chen, S.; Wei, G. Effective light field de-occlusion network based on Swin transformer. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 2590–2599. [CrossRef]
60. Wexler, Y.; Shechtman, E.; Irani, M. Space-time completion of video. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 463–476. [CrossRef] [PubMed]
61. Le Pendu, M.; Jiang, X.; Guillemot, C. Light field inpainting propagation via low rank matrix completion. *IEEE Trans. Image Process.* **2018**, *27*, 1981–1993. [CrossRef]
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
63. Piao, Y.; Rong, Z.; Xu, S.; Zhang, M.; Lu, H. DUT-LFSaliency: Versatile dataset and light field-to-RGB saliency detection. *arXiv* **2020**, arXiv:2012.15124.
64. 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.* **2016**, *39*, 287–300. [CrossRef]
65. Vaish, V.; Adams, A. The (new) stanford light field archive. *Comput. Graph. Lab. Stanf. Univ.* **2008**, *6*, 3.
66. Raj, A.S.; Lowney, M.; Shah, R. *Light-Field Database Creation and Depth Estimation*; Stanford University: Palo Alto, CA, USA, 2016.
67. Rerabek, M.; Ebrahimi, T. New light field image dataset. In Proceedings of the 8th International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 6–8 June 2016.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.