electronics

MDPI

Article

Prex-Net: Progressive Exploration Network Using Efficient Channel Fusion for Light Field Reconstruction

Dong-Myung Kim [1], Young-Suk Yoon [2], Yuseok Ban [1] and Jae-Won Suh [1,*]

[1] School of Electronics Engineering, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea; dmkim@chungbuk.ac.kr (D.-M.K.); ban@cbnu.ac.kr (Y.B.)
[2] Content Research Division, Telecommunications and Media Research Laboratory, Electronics and Telecommunications Research Institute, Daegeon 34129, Republic of Korea; ys.yoon@etri.re.kr
* Correspondence: sjwon@chungbuk.ac.kr; Tel.: +82-43-261-3268

Abstract: Light field (LF) reconstruction is a technique for synthesizing views between LF images and various methods have been proposed to obtain high-quality LF reconstructed images. In this paper, we propose a progressive exploration network using efficient channel fusion for light field reconstruction (Prex-Net), which consists of three parts to quickly produce high-quality synthesized LF images. The initial feature extraction module uses 3D convolution to obtain deep correlations between multiple LF input images. In the channel fusion module, the extracted initial feature map passes through successive up- and down-fusion blocks and continuously searches for features required for LF reconstruction. The fusion block collects the pixels of channels by pixel shuffle and applies convolution to the collected pixels to fuse the information existing between channels. Finally, the LF restoration module synthesizes LF images with high angular resolution through simple convolution using the concatenated outputs of down-fusion blocks. The proposed Prex-Net synthesizes views between LF images faster than existing LF restoration methods and shows good results in the PSNR performance of the synthesized image.

Keywords: light field reconstruction; angular super-resolution; convolutional neural network

Citation: Kim, D.-M.; Yoon, Y.-S.; Ban, Y.; Suh, J.-W. Prex-Net: Progressive Exploration Network Using Efficient Channel Fusion for Light Field Reconstruction. *Electronics* **2023**, *12*, 4661. https://doi.org/10.3390/electronics12224661

Academic Editors: Jose Santamaria and Chunwei Tian

Received: 6 October 2023
Revised: 13 November 2023
Accepted: 13 November 2023
Published: 15 November 2023

1. Introduction

Unlike general cameras, micro-lens-based light field (LF) cameras have the advantage of being able to obtain multi-view images with a single shot. Therefore, they are used in various applications, such as image refocusing [1,2], 3D reconstruction [3], and depth estimation [4,5]. However, due to the limited resolution of the image sensor, the LF images have a small spatial and angular resolution, making them difficult to use in actual applications. To overcome these problems, research is actively being conducted to increase the spatial resolution or angular resolution of LF images [6–21]. In particular, the method of increasing angular resolution is called light field reconstruction.

LF reconstruction is generating novel views that do not exist between LF input images. With recent developments in deep learning, convolutional neural network (CNN)-based LF reconstruction methods have been proposed [9–12,15–21]. LF reconstruction methods can be generally divided into methods that use depth images and methods that do not use depth images. Methods using depth images [9–12] reconstruct the LF images through the two processes of predicting the depth image and restoring the details. Therefore, the synthesized target view image greatly depends on the accuracy of the predicted depth image. As a result, these methods cause distortions, mainly at the edges of objects, or occlusions due to inaccurately predicted depth images, which reduces the quality of the synthesized images.

To avoid these problems arising from the inaccuracy of the predicted depth images, CNN-based methods [15–21] that do not use depth images have been proposed. Because LF images have a small disparity, the correlation that exists between images can be effectively

extracted by applying a filter whose receptive field is larger than the disparity. However, these methods require relatively long execution times because they pre-process and post-process images or use complex networks to improve performance.

In this paper, we propose a progressive exploration network using efficient channel fusion for light field reconstruction (Prex-Net) to solve these problems arising from existing LF reconstruction methods and obtain high-quality LF images. In the previous algorithm [22], we used adjacent 2×2 input LF images to reconstruct the 3×3 LF images and sequentially generate center images in the horizontal, vertical, and diagonal directions. The algorithm proposed in this paper reconstructs 7×7 LF images by reconstructing the LF images using the 2×2 corner input LF images. The proposed network consists of three modules: initial feature extraction, channel fusion, and LF reconstruction. The initial feature extraction module extracts the initial feature map using successive 3D convolution layers. The 3D convolution is effective in extracting spatiotemporal information; thus, it is suitable for extracting correlations between input LF images. However, the proposed network uses 3D convolution only in the initial feature extraction module because 3D convolution requires a high computational cost. The channel fusion module is a modified deep back-projection structure [23], which comprises densely connected fusion groups. The successive fusion groups search for meaningful information required for LF image reconstruction. Finally, the LF reconstruction module concatenates the feature maps extracted from the channel fusion module and synthesizes the LF images with a simple convolution layer. The proposed Prex-Net generates high-quality LF images faster than existing methods. The contributions of this paper are as follows:

- The proposed Prex-Net not only does not perform warping using depth images but also does not require pre-processing for initial feature extraction or post-processing for LF image improvement; thus, it quickly reconstructs LF images;
- For high-quality LF reconstruction, we use 3D convolution for initial feature extraction and progressively explore successive fusion groups to efficiently retrieve meaningful information existing between channels.

The following paper is structured as follows. Section 2 analyzes existing LF reconstruction methods by dividing them into depth-dependent methods [6–11] and depth-independent methods [13–21]. Section 3 covers the structure of the proposed network and the specific LF image reconstruction method. Section 4 demonstrates the superiority of the proposed method over existing methods through various performance evaluations. Finally, Section 5 concludes by analyzing the pros and cons of the proposed method.

2. Related Works

2.1. Depth-Dependent LF Reconstruction

Conventional LF reconstruction methods using depth information include the following methods [6–8]. Wanner and Goldluecke [6] proposed a variational model to synthesize novel views of LF images. In this method, the disparity map is estimated through the local epipolar plane image (EPI) of the LF images and the warp map required for view synthesis is generated using the obtained disparity map. This method has a fast execution time because it does not require the discretization of the disparity map. However, it is hard to synthesize regions with light reflection or occlusions.

Mitra and Veeraraghavan [7] proposed a patch-based algorithm using the Gaussian mixture model (GMM). They designed a GMM patch prior to using the disparity pattern of the LF images and integrated the patches to generate the LF super-resolution images. However, this method was not easy to synthesize if the patch size was smaller than the maximum disparity of the LF images.

Due to recent developments in deep learning, methods using the CNN have been proposed for LF image restoration [9–12]. Flynn et al. [9] proposed a deep learning framework that synthesizes new views using stereo images with a large disparity. However, since only two images are used, the viewpoint of the synthesized image is limited; thus, improvement was needed.

Kalantari et al. [10] proposed a framework that predicts depth information using densely sampled LF images, warps the input image with the predicted depth information, and then refines the warped intermediate synthesized image with a color estimation module. However, because the target views were synthesized one by one, the speed of reconstructing LF images was quite slow.

LFASR-geometry [11] solved the problem of the one-by-one reconstruction of LF images by predicting depth maps of multiple images at once in the middle of the framework. However, distortion occurred in the synthesized image due to the inaccurate estimation of depth information. To overcome this problem, LFASR-FS-GAF [12] adopted an attention map and plane sweep volume (PSV) to create a more accurate intermediate synthesized image. They can synthesize good-quality LF images; however, there was a possibility for improvement regarding predicting a more accurate depth map.

2.2. Depth-Independent LF Reconstruction

Shi et al. [13] proposed an approach to reconstruction that optimizes for sparsity in the continuous Fourier spectrum. They recovered the sparsity of the original continuous Fourier spectrum based on nonlinear gradient descent. But they needed specific sampling patterns to synthesize the LF images.

Vagharshakyan et al. [14] used an adaptive discrete shearlet transform to reconstruct LF images using the EPI inpainting. Since they restored LF images using EPIs, synthesizing was performed successively on an axis-by-axis basis. Because of this, a slight speed decrease occurred and improvements were needed in image synthesis in areas with occlusions where it was difficult to restore the epipolar line.

Because densely sampled LF images do not have large disparity, they have the characteristic that the distance of pixels that must be referenced when synthesizing images is small. From this perspective, deep learning networks with limited receptive fields are suitable for reconstructing densely sampled LF images. Recently, LF reconstruction methods using advanced deep learning technology have been proposed [15–21].

Yoon et al. [15] proposed a network that uses two adjacent views to generate an intermediate view. However, the network could not fully utilize the depth information inherent in LF images due to limited input. Therefore, there is a disadvantage in that input with a large disparity cannot be processed.

Gul and Gunturk [16] introduced a learning-based method to improve spatial and angular resolution by inputting lenslet image stacks into a network. Similar to this method [9], the angular resolution was increased to 3×3 from 2×2. However, quality deteriorated due to the simple network structure and, so, improvement is needed.

Wu et al. [19] reconstructed LF images using EPIs. They proposed a framework with a blur, detail reconstruction, and deblur structure. In this method, when the disparity exceeds a certain value, the quality of the resulting image deteriorates due to limitations in the size of the blur kernel. Additionally, Wu et al. [20] proposed a framework that can synthesize images with relatively larger disparity by sheared EPIs. To solve the problem, they convert the input image to an EPI, restore the sheared EPI, and then reconstruct the EPIs through the inverse shear operation. This method has a slightly slower processing time than the previous method because it uses a complex EPI reconstruction technique consisting of several steps.

DistgASR [21] was able to efficiently extract inherent spatial and angular information simultaneously by extracting features using the MacPI structure and various filters. However, despite the excellent quality of the resulting image, the time for inference tended to be slightly delayed due to the use of various filters in parallel.

3. Proposed Network

3.1. Overview

The goal of the proposed network is to reconstruct low angular resolution (LR) LF images $L^{LR} \in \mathbb{R}^{n^2 \times H \times W}$ to high angular resolution (HR) LF images $\hat{L}^{HR} \in \mathbb{R}^{N^2 \times H \times W}$ by

increasing the angular resolution of the LF images. The n^2 and N^2 are the low and high angular resolutions of the LF images. And the spatial resolution of the LF images is represented by $H \times W$, where H and W are the height and width of LF images, respectively. This can be defined as Equation (1); the proposed network receives the four corner images of the original LF images and reconstructs the LF images with angular high resolution:

$$\hat{L}^{HR} = f\left(L^{LR}\right). \tag{1}$$

The architecture of the entire network consists of three parts, as shown in Figure 1. The input image is a corner image of the original LF image, which is stacked in order from top left to bottom right and input into the network. In the initial feature extraction module, we extract the inherent correlation information of the stacking input images through the successive 3D convolution layers. The extracted initial feature map is input to the channel fusion module to fuse the features and output meaningful features necessary for LF image reconstruction. The channel fusion module consists of successive fusion groups and each fusion group has a pair of up-fusion block and down-fusion block. Finally, the feature maps from the down-fusion blocks are concatenated and fed into the LF reconstruction module to produce the reconstructed LF image with one simple convolution layer.

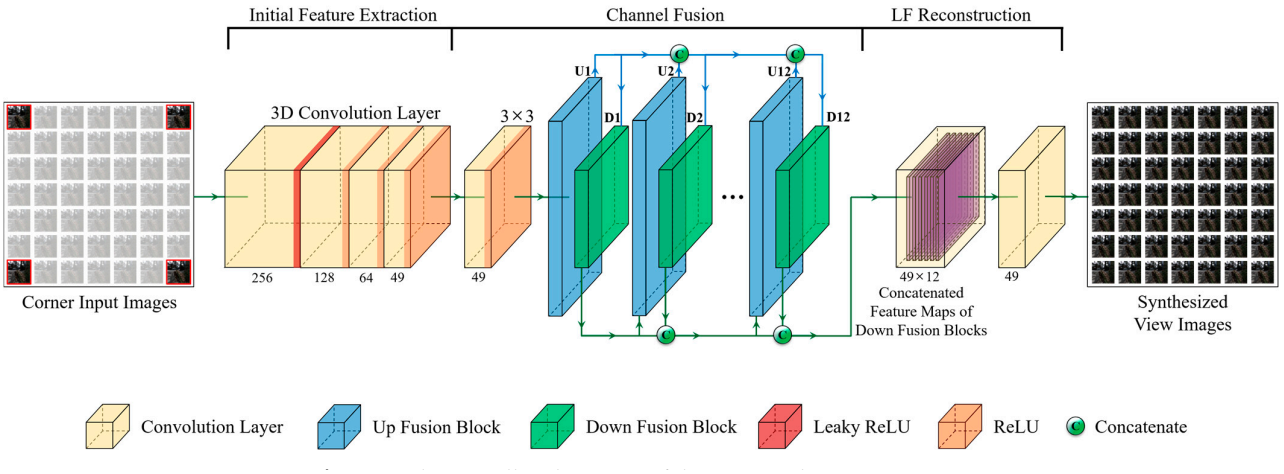

Figure 1. The overall architecture of the proposed Prex-Net.

3.2. Initial Feature Extraction

The initial feature extraction module consists of four 3D convolution layers. The 3D convolutional neural network [24] is mainly used to extract spatiotemporal features from multiple images, such as video data [25,26]. LF images have continuity like the frames of a video; thus, there is a deep relationship between LF images. Therefore, it can be naturally assumed that 3D convolution can show good performance in extracting the feature maps of LF images [11,17].

In the proposed network, the shape of the input feature map is reconstructed from (4, H, W) to (1, 4, H, W) and, then, 3D convolution is applied. In the initial feature extraction module, the first three layers use a $3 \times 3 \times 3$ 3D convolution with a stride of (1, 1, 1) and padding of (1, 1, 1). The last 3D convolution layer has a filter size of $4 \times 3 \times 3$, stride of (1, 1, 1), and padding of (0, 1, 1). The feature map that passes the final 3D convolution has a size of (49, 1, H, W) and is reshaped to a size of (49, H, W) before being input to the channel fusion module. After the first 3D convolution layer, we apply leaky ReLU [27] to extract as many features as possible and use ReLU after the remaining 3D convolution layers.

3.3. Channel Fusion Module

The channel fusion module consists of one 2D convolution layer and twelve channel fusion groups. The 2D convolution layer is used to refine the feature map extracted from

the last 3D convolution layer in the initial feature extraction module. The channel fusion group consists of an up-fusion block and a down-fusion block.

The illustration of the up-fusion block can be seen in Figure 2. The input of the first up-fusion block is the output of the 2D convolution layer within the channel fusion module. The inputs of the remaining up-fusion blocks are the concatenated outputs of the previous down-fusion blocks $\left[D^1, D^2, \cdots, D^{i-1}\right]$. This can be defined by Equation (2), where i represents the order of the fusion group. The concatenated output D^{i-1}_{feats} of the down-fusion blocks is passed through a 1×1 convolution [28] layer $p(.)$ to adjust the number of the channel to 49:

$$D^{i-1}_{feats} = \mathrm{CAT}\Big(\left[D^1, D^2, \cdots, D^{i-1}\right]\Big), \text{ where } i > 1, \tag{2}$$

where $\mathrm{CAT}(.)$ is the concatenate operator.

$$U^i_0 = \phi\Big(p\Big(D^{i-1}_{feats}\Big)\Big), \tag{3}$$

where $\phi(.)$ represents the activation function PReLU.

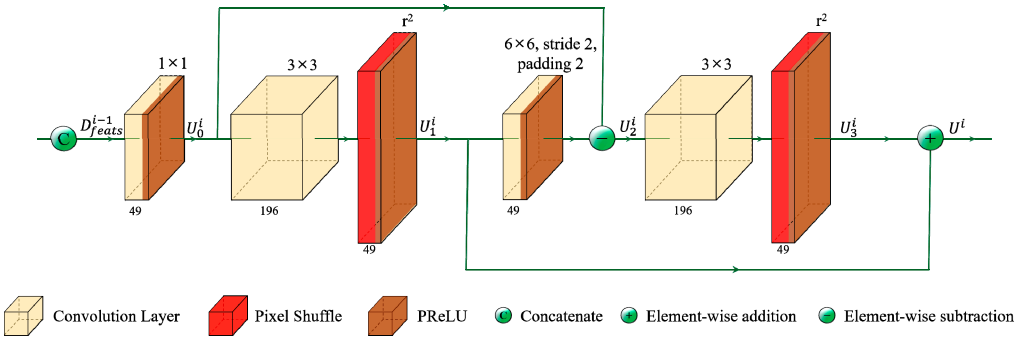

Figure 2. Up-fusion block.

The feature map U^i_0 is the input of the 3×3 convolution layer $q(.)$, which extends the channels of the feature map to 196. After expanding the channels, we apply pixel shuffle $PS_r(.)$ with an upscale factor r to gather pixels from adjacent channels [29–31]. The pixel shuffle reshapes a feature map $\mathbb{R}$ of size $(r^2 \times C, H, W)$ into a feature map $\mathbb{R}$ of size $(C, r \times H, r \times W)$, where C is the number of channels. It can be expressed as:

$$\mathbb{R}^{C \times rH \times rW} = PS_r\Big(\mathbb{R}^{r^2 C \times H \times W}\Big). \tag{4}$$

$$U^i_1 = \phi\Big(PS_r\Big(q\Big(U^i_0\Big)\Big)\Big), \text{ where } r = 2. \tag{5}$$

The feature map U^i_1 is downsized by the 6×6 convolution layer $g(.)$ with a stride of 2 and padding of 2. The downsized feature map is subtracted from the U^i_0 to form the residual feature map U^i_2 by:

$$U^i_2 = U^i_0 - \phi\Big(g(U^i_1)\Big). \tag{6}$$

The residual feature map U^i_2 is, again, expanded into 196 channels through the 3×3 convolution layer and, then, follows the pixel shuffle operation. This can be described as:

$$U^i_3 = \phi(PS_2(q(U^i_2))). \tag{7}$$

Finally, the two feature maps U^i_3 and U^i_1 are added and used as input for the down-fusion block:

$$U^i = U^i_3 + U^i_1. \tag{8}$$

The illustration of the down-fusion block can be seen in Figure 3. The inputs of the down-fusion blocks are the concatenated outputs of the previous up-fusion blocks $\left[U^1, U^2, \cdots, U^i\right]$. The concatenated output U^i_{feats} of the up-fusion blocks is passed through a 1×1 convolution layer to adjust the number of the channel to 49, like the up-fusion block. It can be expressed as:

$$U^i_{feats} = \mathrm{CAT}\Big(\left[U^1, U^2, \cdots, U^i\right]\Big), \text{ where } i > 0. \tag{9}$$

$$D^i_0 = \phi\Big(p\Big(U^i_{feats}\Big)\Big). \tag{10}$$

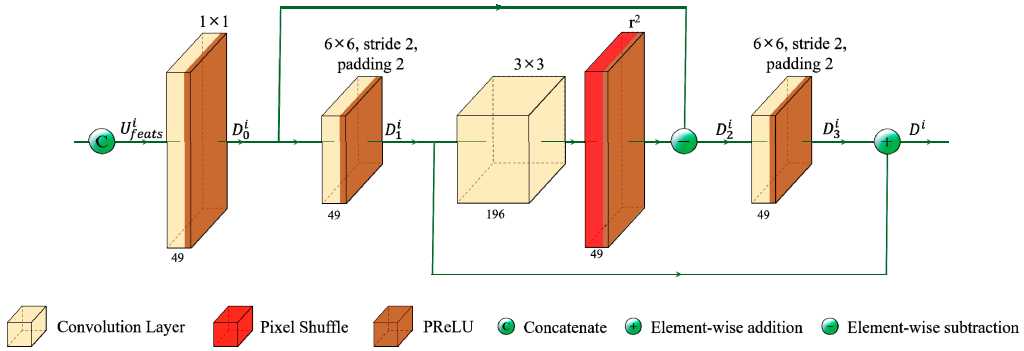

Figure 3. Down-fusion block.

Afterwards, the feature map D^i_0 is fused through the 6×6 convolution layer with a stride of 2 and padding of 2:

$$D^i_1 = \phi\Big(g\Big(D^i_0\Big)\Big). \tag{11}$$

The feature map D^i_1 is expanded to 196 channels through the 3×3 convolution layer and then follows the pixel shuffle operation. The feature map rearranged by pixel shuffle is subtracted from the previous feature map D^i_0. This can be described as:

$$D^i_2 = D^i_0 - \phi\Big(PS_2\Big(q\Big(D^i_1\Big)\Big)\Big). \tag{12}$$

The residual feature map D^i_2 is, again, downsized into 49 channels through a 6×6 convolution layer with a stride of 2 and padding 2. It can be expressed as:

$$D^i_3 = \phi\Big(g(D^i_2)\Big). \tag{13}$$

Finally, the output of down-fusion block D^i is obtained by:

$$D^i = D^i_3 + D^i_1. \tag{14}$$

3.4. LF Reconstruction

The LF reconstruction module receives the feature map D^i_{feats}, which is a concatenate of the output of the down-fusion blocks. The LF reconstruction module consists of one 3×3 convolution layer. As a result, we can obtain 49 reconstructed images:

$$\hat{L}^{HR} = q\Big(D^i_{feats}\Big). \tag{15}$$

We trained the proposed Prex-Net using the Charbonnier loss [32] because the Charbonnier loss not only mitigates the potential blur associated with L2 loss but also enhances robustness to outliers [33]. The penalty coefficient ε is set to 10^{-3}. The network is trained to

minimize the loss between the ground-truth LF image L^{HR} and the synthesized LF image $\hat{L}^{HR}$. The loss function is defined as:

$$\mathcal{L} = \sqrt{\left\| L^{HR} - \hat{L}^{HR} \right\|^2 + \epsilon^2}. \tag{16}$$

4. Experimental Results

In this section, we compare and evaluate the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and execution time to verify the superiority of the proposed Prex-Net.

4.1. Datasets and Implementation Details

The images used for network training are 100 LF images obtained from the Stanford Lytro Light Field Archive [34] dataset and Kalantari et al. [10]. All LF images in the dataset have an angular resolution of 14 × 14 and a spatial resolution of 376 × 541. Among the 14 × 14 LF images, we only used the centrally located 7 × 7 LF images, excluding the highly distorted LF images on the outside.

We changed the RGB domain of the LF images to the YCbCr domain and used only the Y value, which is the luminance value. We used randomly spatially cropped 96 × 96 patches to train the network. Our Prex-Net was optimized with the ADAM optimizer [35] with β_1 = 0.9, β_2 = 0.999. The initial learning rate was 1×10^{-4}, which decreased by 0.5 every 500 epochs, and a total of 3000 epochs were trained. All experiments were performed on a Nvidia RTX 4090 GPU.

4.2. Comparison with State-of-the-Art Methods

To evaluate the synthesized LF images of the proposed method, the PSNR and SSIM of the reconstructed image were compared with the existing methods, which were that of Kalantari et al. [10], LFASR-FS-GAF [12], and DistgASR [21]. Both PSNR and SSIM are values that compared only the Y channel of the original image and the reconstructed LF image. For testing, we used 30Scenes [10] consisting of 30Scenes, Occlusions [34] consisting of 25 scenes, and Reflective [34] consisting of 15 scenes. We used pre-trained networks to compare the performance with existing methods and, if there were no trained networks, we trained them in the same conditions. The PSNR and SSIM used for quantitative performance evaluation can be formulated as:

$$\text{PSNR} = \log_{10}\left(\frac{R^2}{MSE}\right), \tag{17}$$

$$\text{SSIM}(\text{x}, \text{y}) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_y^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{18}$$

where R represents the maximum value of the pixel and MSE means the mean squared error. Additionally, μ_x, μ_y represent the means of the images x and y, σ_x, σ_y represent the variances of the images x and y, σ_{xy} represents covariance, and C represents the contrast constant.

Table 1 shows the experimental results comparing the PSNR and SSIM. Generally, methods that do not use depth images, such as DistgASR [21] and the proposed Prex-Net, are superior to methods that use depth images [10,12]. The performance of the proposed Prex-Net improved as the number of utilized fusion groups increased and, when six groups were applied, it exhibited a performance superior to DistgASR [21]. The proposed Prex-Net with 12 fusion groups showed average PSNR improvements of 1.82 dB, 0.89 dB, and 0.23 dB over the Kalantari et al. [10] method, LFASR-FS-GAF [12], and DistgASR [21], respectively.

Table 1. PSNR/SSIM values achieved by different methods for $2 \times 2 \rightarrow 7 \times 7$ angular SR.

Method	30Scenes	Occlusions	Reflective	Average
Kalantari et al. [10]	41.42/0.984	37.46/0.974	38.07/0.953	38.98/0.970
LFASR-FS-GAF [12]	42.75/0.986	38.51/0.979	38.35/0.957	39.87/0.974
DistgASR [21]	43.61/0.995	39.44/0.991	39.05/0.977	40.70/0.988
Prex-Net (Group 6)	43.39/0.987	39.70/0.982	39.26/0.960	40.78/0.976
Prex-Net (Group 12)	43.49/0.987	40.00/0.983	39.30/0.961	40.93/0.977

The SSIM results show that the SSIM values of the proposed network are on average 0.011 lower than those of DistgASR [21] because the proposed Prex-Net performs 3D convolutions in the initial feature extraction module, which tends to slightly blur the extracted features. Additionally, the results of the proposed Prex-Net on the 30Scenes dataset showed a 0.12 dB reduction compared to DistgASR [21]. The reason is that DistgASR [21], which uses more filters in parallel, is more powerful for spatial data processing. However, the proposed Prex-Net performs better on the edges of occluded regions and areas with complex light reflections, such as Occlusions or Reflective datasets. This proves that the proposed algorithm extracts correlations better from the input image and that the search and fusion process of the features distributed between channels is effective.

Figure 4 shows the comparison results of visual quality for reconstructed LF images. It can be seen that the proposed Prex-Net expresses the details of the image better than the existing method. To demonstrate the superiority of the synthesized image quality, we display the error map at full image size and show two cropped images. The error map expresses the difference between the ground-truth image and the synthesized image as a value between 0 and 1. At this time, to facilitate error comparison, difference values greater than 0.1 are displayed as 0.1. Rock, occlusions_44, and reflective_3, used for image quality comparison, are images included in 30Scenes [10], Occlusions [34], and Reflective [34], respectively. In particular, occlusions_44 has complex occlusions, such as leaves and tree branches, and reflective_3 is difficult to synthesize because light is reflected from the car surface. In the Rock images, the contours of leaves appear clearly and the synthesized background is also much more distinct compared to conventional methods. Likewise, in the case of occlusions_44, it is difficult to compare visually due to the complex mixing of leaves but there are almost no artifacts that occur in existing methods; thus, an image of similar quality to the original image can be confirmed. Looking at the reflectvie_3 image, you can see that the proposed method predicts the light reflection area better than the existing method, which excessively reflects light blur in the image.

4.3. Runtime Evaluation

In this section, we compare the average PSNR and execution time required to reconstruct a 7×7 LF image with existing methods. In Figure 5, it can be seen that methods using depth images have relatively low image quality and slow LF image reconstruction speed. In particular, the method of Kalantari et al. [10] is slow because it creates a depth image for each target image and synthesizes the images one by one. LFASR-FS-GAF [12] reconstructed LF images approximately 100 times faster, with an average speed of 0.89 dB higher than that of Kalantari et al. [10]. However, LFASR-FS-GAF [12] also uses PSV and attention maps to predict depth images; thus, it takes quite a long time to synthesize LF images.

The proposed Prex-Net shows a similar PSNR performance to DistgASR [21] but with a faster execution time. The reason is that DistgASR [21] uses various filters internally, causing some time delay. Prex-Net6, using six fusion groups, reconstructs LF images by an average of 0.08 dB greater than DistgASR [21] and about five times faster. Additionally, Prex-Net12, using 12 fusion groups, is 0.23 dB greater on average and reconstructs LF images approximately 2.8 times faster.

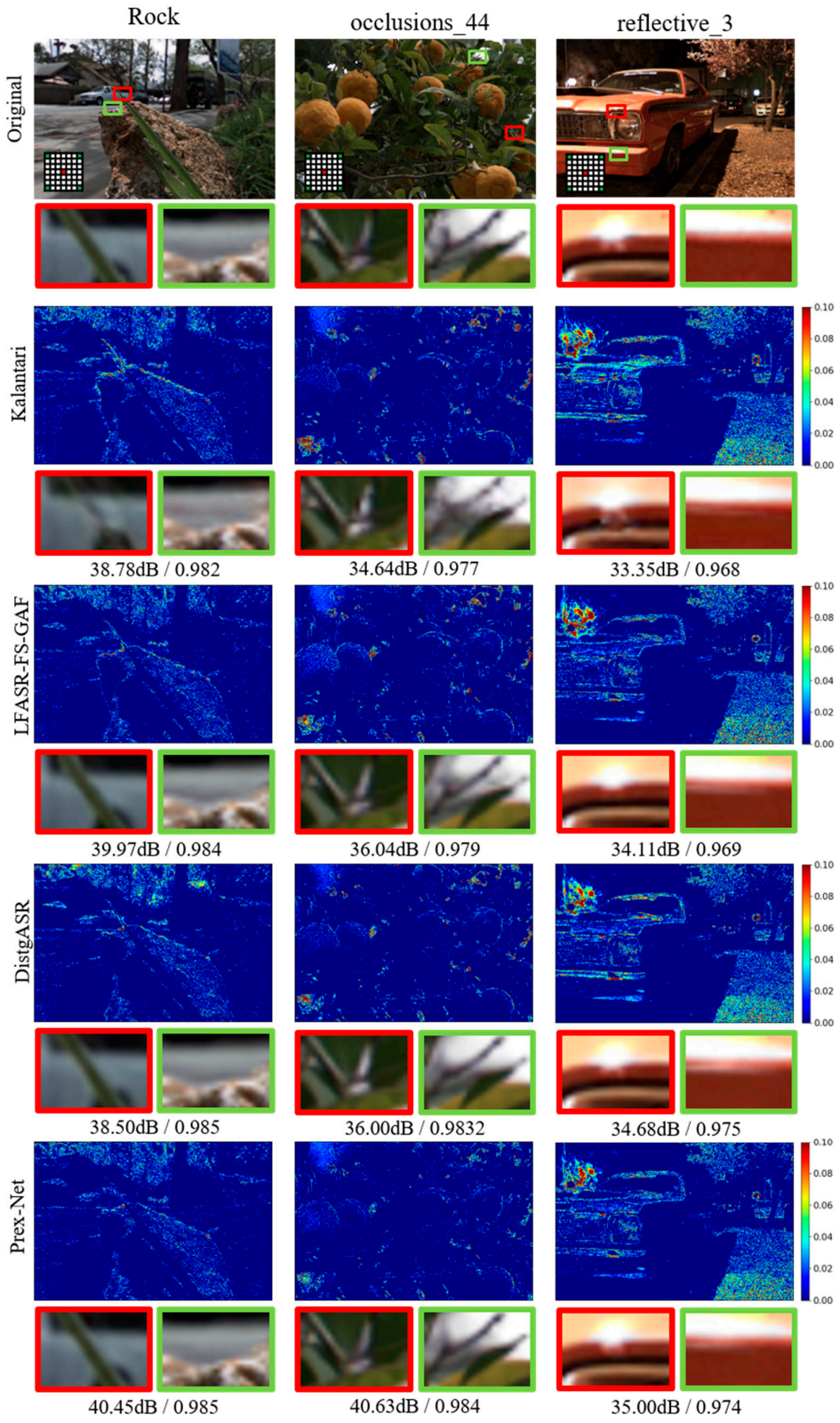

Figure 4. Comparison of the visual quality of reconstructed LF images.

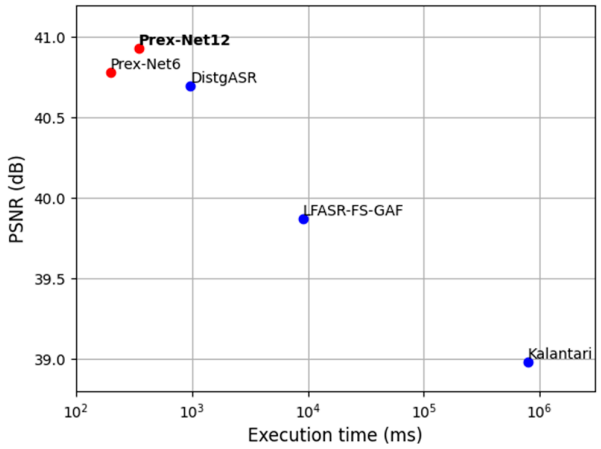

Figure 5. Performance (PSNR) and execution times comparison.

4.4. Ablation Study

In this section, we investigate the effect of the number of fusion groups and fusion block filter size on the quality of the synthesized LF images and experimentally verify the efficiency of the designed network.

4.4.1. Number of Fusion Groups

In general, it is a well-known fact that network performance increases as the network deepens. However, when the network becomes deeper than a certain level, issues like gradient vanishing or exploding can occur. To prevent this problem, methods such as ResNet [36] and DenseNet [37] were introduced to stably deliver weights to the end of the network. A network with this structure can avoid the problems mentioned above; however, as the network deepens, performance reaches saturation.

Therefore, in order to increase the efficiency of the network compared to its computational cost, the number of groups used in the network was experimentally determined. We compared PSNR by increasing the number of fusion groups to three, six, nine, twelve, and fifteen. As you can see in Table 2, it can be observed that PSNR performance converges with a difference of approximately 0.05 dB when the number of fusion groups reaches 12 and 15. Furthermore, when we examine the changes in execution time and the number of parameters, as shown in Table 3, it can be observed that the increase in the number of fusion groups results in a marginal improvement in PSNR performance compared to the increase in execution times and parameters.

Table 2. PSNR comparison according to the number of fusion groups.

Number of Groups	30Scenes	Occlusions	Reflective	Average
Group 3	43.13	39.37	38.83	40.44
Group 6	43.39	39.70	39.26	40.78
Group 9	43.42	39.86	39.25	40.84
Group 12	43.49	40.00	39.30	40.93
Group 15	43.49	40.03	39.41	40.98

Table 3. Comparison of execution times and parameters according to the number of fusion groups.

Number of Groups	Execution Time (ms)	Parameters (M)
Group 3	140	2.89
Group 6	198	4.58
Group 9	254	6.31
Group 12	347	8.08
Group 15	435	9.90

4.4.2. Filter Size

Fusion block increases spatial resolution by gathering pixels from adjacent channels of the feature map through pixel shuffle; then, 2D convolution is performed with a stride of 2 to fuse the features. In this process, the filter size determines how much information from adjacent channels will be used.

However, increasing the filter size can lead to an increase in the number of parameters, potentially reducing the efficiency of the network. Therefore, it is important to choose an appropriate filter size. As you can see in Table 4, there was a slight performance improvement with increasing filter size, with average performance differences of 0.04 dB and 0.05 dB, respectively. Therefore, in this paper, a filter size of 6×6 was selected considering efficiency versus performance.

Table 4. PSNR comparison according to filter size of up- and down-fusion blocks.

Filter Size	30Scenes	Occlusions	Reflective	Average
4×4	43.55	39.96	39.15	40.89
6×6	43.49	40.00	39.30	40.93
8×8	43.48	40.04	39.43	40.98

5. Conclusions

In this paper, we proposed the Prex-Net, which progressively explores information by fusing channels to reconstruct LF images. In the initial feature extraction module, we used 3D convolution to extract strong correlations between input LF images. Afterward, the network explored and fused LF features into successive fusion groups in a back-projection structure. The proposed fusion groups focused on using the information inherent between channels without missing the deep correlation between the LF images. The fusion blocks in the fusion group reshape the size of the feature map using pixel shuffle, efficiently searching for the features required for LF reconstruction that are widespread in the channel. In addition, in the process of reducing the reshaped feature map through convolution, the features are fused, effectively improving the related information extracted between channels. And the fusion block was able to synthesize LF images at high speed by reducing the computation amount of the overall network. By changing the filter size during the fusion process, we experimentally confirmed that the quality of the synthesized LF images improves as more information from adjacent channels is referenced.

In a performance comparison with existing methods, it was confirmed that methods using depth images generally tend to deteriorate the quality of synthesized LF images due to the inaccurate prediction of depth images. Additionally, when compared to existing methods that do not use depth images, the reconstructed image was generated faster than the existing method and synthesized LF images of good quality were shown. In particular, the proposed network shows good performance, even in occluded areas and light reflection areas that are difficult to synthesize well. However, the proposed network has a relatively large number of parameters; thus, there is a need to reduce the required memory.

Author Contributions: Conceptualization, D.-M.K.; methodology, D.-M.K. and Y.-S.Y.; software, D.-M.K.; formal analysis, D.-M.K. and Y.-S.Y.; investigation, Y.-S.Y. and J.-W.S.; resources, J.-W.S.; data curation, D.-M.K.; writing original draft preparation, D.-M.K. and Y.-S.Y.; writing—review and editing, D.-M.K., Y.-S.Y., Y.B. and J.-W.S.; validation, D.-M.K. and J.-W.S.; visualization, D.-M.K.; supervision, Y.B. and J.-W.S.; project administration, J.-W.S.; funding acquisition, J.-W.S. 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 grant funded by the Korean government (MSIT) (No. 2022R1A5A8026986).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this paper are public datasets. We also provide the proposed method's training and evaluation codes at: https://github.com/dmkim17/Prex-Net, which was created (accessed on 6 October 2023).

Acknowledgments: This research was partly supported by the KOCCA in the CT R&D Program [RS-2023-00227409, the development of an indoor semantic 3D modeling technology based on a sketch for user-centric space creation in metaverse content].

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

References

1. 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]
2. Zhang, C.; Hou, G.; Zhang, Z.; Sun, Z.; Tan, T. Efficient auto-refocusing for light field camera. *Pattern Recognit.* **2018**, *81*, 176–189. [CrossRef]
3. Kim, C.; Zimmer, H.; Pritch, Y.; Sorkine-Hornung, A.; Gross, M.H. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.* **2013**, *32*, 73:1–73:12. [CrossRef]
4. Jeon, H.G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.W.; So Kweon, I. Accurate depth map estimation from a lenslet light field camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1547–1555. [CrossRef]
5. Wang, T.C.; Efros, A.A.; Ramamoorthi, R. Occlusion-aware depth estimation using light-field cameras. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3487–3495. [CrossRef]
6. Wanner, S.; Goldluecke, B. Variational light field analysis for disparity estimation and super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 606–619. [CrossRef] [PubMed]
7. Mitra, K.; Veeraraghavan, A. Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 22–28. [CrossRef]
8. Le Pendu, M.; Guillemot, C.; Smolic, A. A fourier disparity layer representation for light fields. *IEEE Trans. Image Process.* **2019**, *28*, 5740–5753. [CrossRef] [PubMed]
9. Flynn, J.; Neulander, I.; Philbin, J.; Snavely, N. Deepstereo: Learning to predict new views from the world's imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5515–5524. [CrossRef]
10. Kalantari, N.K.; Wang, T.C.; Ramamoorthi, R. Learning-based view synthesis for light field cameras. *ACM Trans. Graph.* **2016**, *35*, 1–10. [CrossRef]
11. Jin, J.; Hou, J.; Yuan, H.; Kwong, S. Learning light field angular super-resolution via a geometry-aware network. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11141–11148. [CrossRef]
12. Jin, J.; Hou, J.; Chen, J.; Zeng, H.; Kwong, S.; Yu, J. Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1819–1836. [CrossRef] [PubMed]
13. Shi, L.; Hassanieh, H.; Davis, A.; Katabi, D.; Durand, F. Light field reconstruction using sparsity in the continuous fourier domain. *ACM Trans. Graph.* **2014**, *34*, 1–13. [CrossRef]
14. Vagharshakyan, S.; Bregovic, R.; Gotchev, A. Light field reconstruction using shearlet transform. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 133–147. [CrossRef] [PubMed]
15. Yoon, Y.; Jeon, H.G.; Yoo, D.; Lee, J.Y.; Kweon, I.S. Light-field image super-resolution using convolutional neural network. *IEEE Signal Process. Lett.* **2017**, *24*, 848–852. [CrossRef]
16. Gul, M.S.K.; Gunturk, B.K. Spatial and angular resolution enhancement of light fields using convolutional neural networks. *IEEE Trans. Image Process.* **2018**, *27*, 2146–2159. [CrossRef] [PubMed]
17. 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. [CrossRef]
18. Yeung, H.W.F.; Hou, J.; Chen, J.; Chung, Y.Y.; Chen, X. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 137–152. [CrossRef]
19. Wu, G.; Zhao, M.; Wang, L.; Dai, Q.; Chai, T.; Liu, Y. Light field reconstruction using deep convolutional network on EPI. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6319–6327. [CrossRef]
20. 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] [PubMed]
21. Wang, Y.; Wang, L.; Wu, G.; Yang, J.; An, W.; Yu, J.; Guo, Y. Disentangling light fields for super-resolution and disparity estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 425–444. [CrossRef] [PubMed]

22. Kim, D.M.; Kang, H.S.; Hong, J.E.; Suh, J.W. Light field angular super-resolution using convolutional neural network with residual network. In Proceedings of the 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), Zagreb, Croatia, 2–5 July 2019; pp. 595–597. [CrossRef]
23. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1664–1673. [CrossRef]
24. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497. [CrossRef]
25. Sakkos, D.; Liu, H.; Han, J.; Shao, L. End-to-end video background subtraction with 3d convolutional neural networks. *Multimed. Tools Appl.* **2018**, *77*, 23023–23041. [CrossRef]
26. Maqsood, R.; Bajwa, U.I.; Saleem, G.; Raza, R.H.; Anwar, M.W. Anomaly recognition from surveillance videos using 3D convolution neural network. *Multimed. Tools Appl.* **2021**, *80*, 18693–18716. [CrossRef]
27. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical evaluation of rectified activations in convolutional network. *arXiv* **2015**, arXiv:1505.00853.
28. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
29. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883. [CrossRef]
30. Wang, L.; Wang, Y.; Lin, Z.; Yang, J.; An, W.; Guo, Y. Learning a single network for scale-arbitrary super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4801–4810. [CrossRef]
31. Zhao, H.; Kong, X.; He, J.; Qiao, Y.; Dong, C. Efficient image super-resolution using pixel attention. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; pp. 56–72. [CrossRef]
32. Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; Barlaud, M. Two deterministic half-quadratic regularization algorithms for computed imaging. In Proceedings of the 1st International Conference on Image Processing, Austin, TX, USA, 13–16 November 1994; pp. 168–172. [CrossRef]
33. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632. [CrossRef]
34. Raj, A.S.; Lowney, M.; Shah, R.; Wetzstein, G. *Stanford Lytro Light Field Archive*; Stanford Computational Imaging Lab: Stanford, CA, USA, 2016.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
36. 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. [CrossRef]
37. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. [CrossRef]