electronics

MDPI

Article

Light Field Super-Resolution via Dual-Domain High-Frequency Restoration and State-Space Fusion

Zhineng Zhang, Tao Yan *, Hao Huang, Jinsheng Liu, Chenglong Wang and Cihang Wei

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China; 6223114033@stu.jiangnan.edu.cn (Z.Z.); 6223110021@stu.jiangnan.edu.cn (H.H.); 6223110026@stu.jiangnan.edu.cn (J.L.); 6223115020@stu.jiangnan.edu.cn (C.W.); 6243115022@stu.jiangnan.edu.cn (C.W.)

* Correspondence: yantao.ustc@gmail.com

Abstract: The current light field super-resolution methods mainly face the following challenges: difficulty in handling redundant information in light fields; heavy reliance on the spatial domain to recover details; and insufficient interaction of spatial and angular features. We propose a novel light field super-resolution (LF-SR) network, termed DHSFNet, which effectively enhances super-resolution performance from a dual-domain perspective, encompassing both the frequency and spatial domains. Our DHSFNet contains three key points. (1) A local sparse angular attention module (LSAA) is proposed to selectively capture relationships between adjacent sub-views using geometric prior information to reduce computational complexity. (2) We design a dual-domain high-frequency restoration sub-network, with a frequency-domain branch using mask-guided multi-scale discrete cosine transform (DCT) restoration and a spatial-domain branch employing multi-scale cross-attention to recover texture details. (3) A Mamba-based fusion module (MF) is introduced to efficiently facilitate global spatial–angular interaction, which achieves linear complexity and outperforms Transformer-based methods in both accuracy and speed. Comprehensive experiments conducted on three benchmark datasets demonstrate the superior performance of our method in the LF-SR task.

Keywords: light field; super-resolution; local sparse angular attention; dual-domain high-frequency restoration; Mamba-based fusion

check for updates

Academic Editor: Silvia Liberata Ullo

Received: 18 March 2025
Revised: 21 April 2025
Accepted: 22 April 2025
Published: 25 April 2025

Citation: Zhang, Z.; Yan, T.; Huang, H.; Liu, J.; Wang, C.; Wei, C. Light Field Super-Resolution via Dual-Domain High-Frequency Restoration and State-Space Fusion. *Electronics* **2025**, *14*, 1747. https://doi.org/10.3390/electronics14091747

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Light field (LF) image processing [1], which captures both the spatial and angular information of light rays, has emerged as a promising technique for 3D scene reconstruction [2–4], virtual reality [5,6], and computational photography [7,8]. Although LF image processing captures rich angular and spatial information, the inherent trade-off between these dimensions often degrades the spatial resolution, yielding low-resolution (LR) LF images. Since high-resolution images are essential for downstream tasks such as depth estimation and sub-view synthesis, effective super-resolution (SR) techniques are critically needed. Although Single-Image Super-Resolution (SI-SR) has achieved remarkable progress, light field super-resolution (LF-SR) remains challenging due to its unique requirement to jointly exploit spatial–angular correlations while preserving high-frequency details. The existing LF-SR approaches mainly operate in the spatial domain, leveraging convolutional neural networks (CNNs) or Transformers to model local textures and long-range dependencies. Although these methods demonstrate potential in super-resolution tasks, they suffer from two critical limitations: First, they inadequately utilize the physical

interpretability of frequency-domain representations, which inherently encode hierarchical texture patterns critical for high-frequency recovery. Second, the existing spatial–angular interaction modules—whether based on 4D convolutions [9], graph networks [10,11], or attention mechanisms—fail to balance computational efficiency with effective feature fusion, often introducing redundant computations.

To address these challenges, we propose DHSFNet, a frequency–spatial collaborative network for LF-SR. Specifically, the network first employs a decoupled feature extraction block (Distg-Block) [12] to initialize spatial, angular, and Epipolar Plane Image (EPI) features of the LF image separately. This preserves the structural information in the LF and prevents feature entanglement. Then, to capture the correlations among the sub-views in LF, we propose a local sparse angular attention module (LSAA), which selectively models relationships between adjacent sub-views using geometric priors, reducing computational complexity by 64% compared to dense attention mechanisms. Next, we propose a dual-domain high-frequency restoration sub-network to recover high-frequency detail information. The frequency-domain branch uses a mask-guided multi-scale DCT module (DCTM) to recover high-frequency details. The spatial-domain branch applies a detail-guided cross-attention module (DCA) to restore texture details. Finally, we design a Mamba-based fusion module (MF) to achieve efficient global spatial–angular interaction using forward and backward state-space models. This enables linear-complexity feature fusion and improves both accuracy and speed over traditional Transformers. Our network is the first to restore high-frequency details of light field images from both the frequency and spatial domains, leveraging a state-space model (SSM) to efficiently model spatial–angular correlations and globally fuse high-frequency information, achieving state-of-the-art performance in LF-SR tasks.

In summary, the main contributions of this work are summarized as follows:

- First, we propose a novel local sparse angular attention module (LSAA) that sparsely samples informative sub-view pairs based on a local window mask mechanism, reducing computational complexity.
- Second, we propose a multi-scale discrete cosine transform (DCT) high-frequency feature recovery module, which effectively enhances the high-frequency information of images from the frequency domain by incorporating a masking mechanism.
- Third, we design a detail-guided cross-attention module (DCA), which restores spatial-domain texture detail features through a multi-scale detail enhancement module and a cross-attention mechanism.
- Finally, we design a Mamba-based fusion module (MF) that models long-range spatial–angular dependencies via hidden state transitions, achieving linear computational complexity compared to the Transformer-based self-attention mechanism.

2. Related Work

2.1. Light Field Super-Resolution

Early light field super-resolution (LF-SR) methods focused on spatial–angular consistency through geometric priors. For example, Wanner et al. [13] exploited Epipolar Plane Image (EPI) analysis to exploit angular correlations. Wang et al. [14] designed a patch-based optimization network with sub-pixel shifts to utilize the rich surrounding sub-views along some typical epipolar directions to explore the inter-view correlations. With the development of deep learning, Yeung et al. [15] introduced 4D convolutions to characterize the relationship among pixels, fully making use of the 4D structure information of LF data in both the spatial and angular domains. Yan et al. [16] resolved occlusion issues in disparity scaling and sub-view transformation by incorporating a smooth energy term, preserving texture details and enabling high-quality SR. Recent works like Zhou et al. [17]

regarded the sub-aperture image of each vertical or horizontal angle sub-view as a sequence and established the long-range geometric dependence through the space angle local enhancement self-attention layer while maintaining the locality of the sub-aperture image. However, these methods often fail to efficiently interact with angular features, resulting in substantial redundant computations. Moreover, the presence of occluded scene points adversely impacts the restoration of sub-aperture images.

2.2. Frequency-Domain Approaches for Image Restoration

Frequency-domain analysis has been explored in general image restoration tasks to address texture recovery. DCT-FANet [18] distinguished and adaptively enhanced the high-frequency information of low-resolution (LR) images and achieved success in recovering high-frequency details. Wavelet-SR [10] integrated wavelet transforms with CNNs to capture multi-scale frequency components. FreqFormer [19] attempted to fuse frequency features across sub-views but used fixed DCT bases, limiting adaptability to complex degradations. However, these methods were primarily designed for single images. Currently, no existing method recovers high-frequency features from the frequency domain for sub-aperture images (SAIs) in LF.

2.3. Spatial–Angular Interaction Mechanisms

Effective spatial–angular fusion is critical for LF-SR. Early efforts like Yoon et al. [20] employed 4D convolutions but faced prohibitive computational costs. V Khorasani Ghassab et al. [21] modeled angular relationships via graph networks yet struggled with real-time processing. Employing attention-based methods, Gao et al. [22] improved efficiency but overlooked angular redundancy in dense light field sub-views. Notably, Liu et al. [23] introduced sparse sampling to reduce sub-view redundancy, but its handcrafted sampling strategy limits generalization. These approaches highlight a persistent trade-off between computational efficiency and feature fusion effectiveness, underscoring the need for a local window sub-view interaction mechanism.

2.4. State-Space Models in Vision Tasks

State-space models (SSMs), particularly Mamba [24], have arisen as efficient alternatives to Transformers by leveraging hidden state transitions for long-range dependency modeling. In vision, VMamba [25] applied SSMs to image classification, while VideoMamba [26] extended them to video understanding. However, SSMs have not yet been explored in the field of LF image processing. The linear complexity of SSMs can effectively address the substantial computational challenges during the fusion of spatial–angular features in light fields. The existing LF-SR methods rely on self-attention with quadratic complexity, making Mamba-based architectures a promising solution for LF-SR tasks.

2.5. Super-Diffraction Imaging

Super-diffraction imaging techniques fundamentally enhance image super-resolution by breaking the classical diffraction limit through advanced optical designs or materials. These methods provide higher-frequency image information and enable the acquisition of finer details beyond conventional optical limits. For example, Pendry et al. [27] proposed the concept of constructing a “perfect lens” using negative-index materials, enabling imaging resolution beyond the diffraction limit. This theoretical breakthrough laid the foundation for subsequent developments in super-resolution imaging techniques. In addition, Liu et al. [28] proposed a macroscopic Fourier ptychographic imaging method that combines synthetic aperture techniques with deep learning. By integrating wave optics modeling with computational imaging, this method effectively overcomes diffraction limi-

tations and provides a complementary solution for enhancing imaging resolution from the perspective of optical modeling.

2.6. Module Adaptation and Novel Design for Light Field Super-Resolution

In this work, we incorporate both adapted and newly designed modules to enhance light field super-resolution. Specifically, the serial network architecture and MF module are modified based on the LFT framework [29] and Vision Mamba [30], respectively, to better exploit spatial–angular correlations. Additionally, we propose the LSAA module to capture high-frequency details via sparse attention among adjacent views and design the DCTM and DCA modules to recover high-frequency information from the frequency and spatial domains, respectively.

3. Method

In this section, we first describe the overall architecture of the proposed network. Next, we introduce the sub-view interaction for high frequency. Then, we introduce the dual-domain high-frequency detail recovery sub-network in detail. Finally, we present the spatial–angular feature fusion process.

3.1. Overall Architecture of Our Network

The overall network architecture is illustrated in Figure 1. Our framework first transforms the 4D light field (LF) image into a 2D macro-pixel representation, enabling the efficient extraction of angular and spatial LF features using standard 2D convolutional kernels. Subsequently, the Spatial Feature Extraction Module (SFE) [12] is employed to capture shallow spatial features from the LF image, while the decoupled feature extraction block (Distg-Block) [12] comprehensively explores structural LF characteristics by separately extracting spatial, angular, and Epipolar Plane Image (EPI) [31] features.

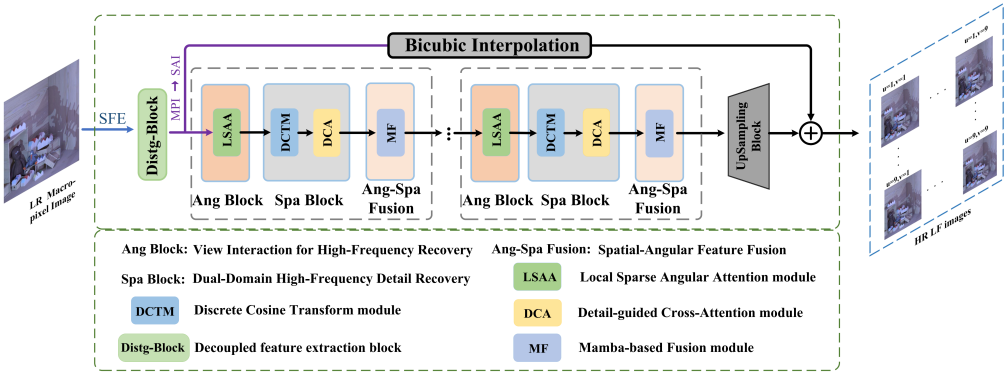


Figure 1. The overview of our network mainly consists of three parts: sub-view interaction for high-frequency recovery part, dual-domain high-frequency detail recovery sub-network part, and spatial–angular feature fusion part.

To specifically enhance high-frequency detail restoration, our approach adopts the following processing: Firstly, the local sparse angular attention module (LSAA) selectively models relationships between key sub-aperture images, effectively suppressing redundant angular information while restoring high-frequency details. Secondly, in frequency domain, the discrete cosine transform module (DCTM) recovers frequency-domain high-frequency components through a masked coefficient matrix. In spatial domain, the detail-guided cross-attention module (DCA) complements this process by preserving texture information. Finally, the Mamba-based fusion module (MF) achieves global consistency of high-frequency details across both spatial and angular dimensions with linear computational complexity.

3.2. Sub-View Interaction for High-Frequency Recovery

Local Sparse Angular Attention

Transformers [32] exhibit powerful global modeling capabilities for capturing high-frequency information by leveraging angular relationships in light field (LF) images. During image restoration, some sub-aperture images [23] may lose high-frequency details due to occlusions or resolution limitations, while complementary information from other sub-views can fill these gaps. However, adjacent LF sub-views exhibit strong local correlations due to disparity continuity, and distant sub-views contribute minimally due to occlusion or depth variations. To balance efficiency and accuracy, we propose a local sparse angular attention mechanism that confines self-attention computation to local windows of 3×3 adjacent sub-views.

Specifically, as illustrated in Figure 2, the input light field (LF) feature map $F \in \mathbb{R}^{C\times U\times V\times H\times W}$ is first normalized using layer normalization to accelerate the learning of LF features. Subsequently, an alternating sine–cosine angular positional encoding P_A is applied to preserve the relative positional relationships between sub-aperture images. The encoding process can be described as follows:

$$P_A(p, 2i) = \sin(\frac{p}{\alpha^{2i/C}}), \tag{1}$$

$$P_A(p, 2i+1) = \cos(\frac{p}{\alpha^{2i/C}}), \tag{2}$$

here, $p \in \{1, 2, \ldots, UV\}$ denotes the positional index of the input, representing the relative position of the sub-aperture image (SAI), UV is the total number of SAIs, i is the channel dimension index, C represents the number of positional channels, and α is the scaling parameter that controls the periodicity of the sine and cosine curves.

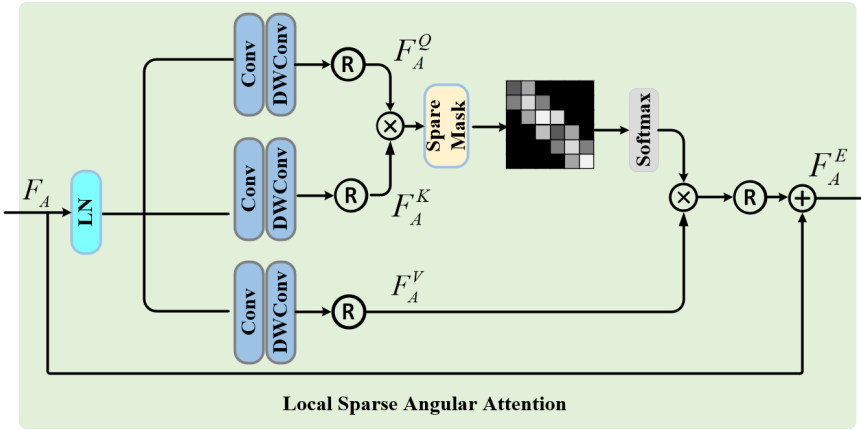


Figure 2. The framework of our local sparse angular attention. It is designed to improve the recovery of high-frequency details by utilizing a 3×3 local attention window. This approach focuses on learning the correlations between adjacent sub-views, thereby mitigating the impact of distant and occluded sub-views on the restoration of high-frequency details.

Next, the light field features are projected into three distinct feature spaces using 3×3 convolutions and depthwise separable convolutions [33]. The extracted features are then reshaped into sequential representation $T_A \in \mathbb{R}^{(UV)\times(HW)\times C}$ suitable for Transformer processing. Based on the sequential representation T_A and the positional encoding P_A, the key features F_A^K, value features F_A^V, and query features F_A^Q are computed. The process is as follows:

$$\begin{aligned} F_A^K &= \mathrm{DWConv}(\mathrm{Conv}(\mathrm{LN}(T_A + P_A))), \\ F_A^V &= \mathrm{DWConv}(\mathrm{Conv}(\mathrm{LN}(T_A + P_A))), \\ F_A^Q &= \mathrm{DWConv}(\mathrm{Conv}(\mathrm{LN}(T_A + P_A))), \end{aligned} \tag{3}$$

$\mathrm{Conv}(\cdot)$ represents the 3×3 convolution, and $\mathrm{DWConv}(\cdot)$ represents the depthwise separable convolution.

Subsequently, the interaction range for each SAI is defined by a 3×3 local window, which includes nine neighboring SAIs. For the i-th SAI at the sub-view position (u_i, v_i), the set of interactive SAIs are defined as follows:

$$\mathcal{N}_i = \{j \mid\mid u_j - u_i \mid \leq 1, \mid v_j - v_i \mid \leq 1\}. \tag{4}$$

Thus, each sub-aperture image can observe its eight neighboring sub-views. Subsequently, a mask matrix $M \in \{0,1\}^{U \times V}$ is defined, as given by the following equation:

$$M_{i,j} = \begin{cases} 1, j \in \mathcal{N}_i \\ 0, j \notin \mathcal{N}_i \end{cases}. \tag{5}$$

Finally, the angle-enhanced feature F_A^E is obtained by modeling the correlations of light field angular information using a local sparse self-attention mechanism, as expressed in the following equation:

$$F_A^E = \mathrm{softmax}(\frac{F_A^Q F_A^{K^T} M}{\sqrt{C}}) \cdot F_A^V \tag{6}$$

3.3. Dual-Domain High-Frequency Detail Recovery Sub-Network

3.3.1. High-Frequency Information Recovery in the Frequency Domain

First, the high-frequency information of the sub-aperture image is recovered from the frequency domain. The SAI is first transformed to the frequency domain using the discrete cosine transform (DCT), and then a coefficient matrix is applied to mask its low-frequency components. The masked coefficient matrix is then applied to the frequency domain matrix F_f to extract the high-frequency information of the SAI. Finally, an inverse DCT is performed on the masked frequency domain matrix to obtain the high-frequency image.

Specifically, as illustrated in Figure 3, given an input sub-aperture image $F_s \in \mathbb{R}^{C \times H \times W}$, its representation in the frequency domain $F_f \in \mathbb{R}^{C \times H \times W}$ is obtained through a 2D DCT transform, as expressed by the following formula:

$$F_f(u,v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} F_S(x,y) C(u,v,x,y), \tag{7}$$

$$C(u,v,x,y) = \cos\left(\frac{\pi(2x+1)u}{2H}\right) \cos\left(\frac{\pi(2y+1)v}{2W}\right), \tag{8}$$

here, (x,y) represents the spatial coordinates of SAI, while (u,v) denotes the coordinates of the coefficient matrix in the high–low frequency matrix array. The low-frequency parts are concentrated in the upper-left corner of the DCT coefficient matrix, where u and v are small. These components correspond to the background and global information of the image. In contrast, the high-frequency components are concentrated in the lower-right corner of the DCT coefficient matrix, representing the edge and texture details of the image.

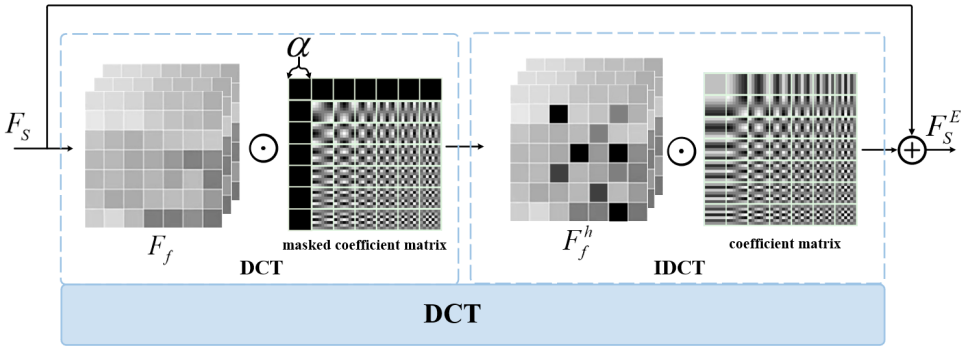


Figure 3. The framework of our mask-based discrete cosine transform employs a learnable parameter α to determine the size of the masked region in the DCT coefficient matrix. As α increases, the proportion of high-frequency details preserved in the image becomes higher.

Subsequently, an adaptive occlusion mask is generated, where the learnable parameter α determines the number of rows and columns to be masked. The masked coefficient matrix $\mathcal{M} \in \{0,1\}^{U \times V}$ is defined as follows:

$$\mathcal{M}(u,v) = \begin{cases} 0, & u \leq \alpha, v \leq \alpha \\ 1, & \text{otherwise} \end{cases}. \tag{9}$$

Evidently, a larger value of α tends to preserve more high-frequency detail information in the image. By combining the masking mechanism with the coefficient matrix, the low-frequency coefficients can be effectively set to zero, thereby preserving the high-frequency information of the image. Subsequently, applying the masking mechanism to the DCT coefficient matrix yields the high-frequency feature representation F_f^h of the frequency-domain representation of the image, which can be expressed as

$$F_f^h(u,v) = F_f(u,v) \cdot \mathcal{M}(u,v). \tag{10}$$

Then, the inverse DCT transform is applied to reconstruct the high-frequency spatial domain feature $F_S^E(x,y)$, with the specific transformation process as follows:

$$F_S^E(x,y) = \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} F_f^h(u,v) C(u,v,x,y) + F_S. \tag{11}$$

However, the aforementioned approach can only extract high-frequency details from sub-aperture images but fails to effectively restore them. To address this, we integrate the DCT operation with a CNN to fully learn the texture details of the image. As shown in Figure 4, we design a discrete cosine transform module (DCTM), which employs three DCT modules at different scales to extract high-frequency information comprehensively. Specifically, convolutional blocks are used to reconstruct the multi-scale high-frequency features. The extracted multi-scale features are concatenated and fused using a convolutional block. A residual connection incorporates high-frequency features into the original image features, producing the high-frequency-enhanced feature F_S^{EE}. This model enhances the clarity and accuracy of details while maintaining global consistency.

Specifically, given the input feature map $F_s \in \mathbb{R}^{C \times U \times V \times H \times W}$, multi-scale downsampling operations are first performed at different scales, followed by high-frequency feature extraction using DCT. The process is described as follows:

$$\begin{aligned} F_s^{(i)} &= \mathrm{Down}_i(F_s), \\ F_{dct}^{(i)} &= \mathrm{DCT}(F_s^{(i)}), \end{aligned} \tag{12}$$

here, i denotes the i-th scale transformation. $\mathrm{Down}_i(\cdot)$ denotes the downsampling operation at different scales, and $\mathrm{DCT}(\cdot)$ represents the discrete cosine transform based on the masking operation, which extracts the high-frequency information of the SAI in LF. Subsequently, convolutional operations are applied to the extracted multi-scale high-frequency features to generate enhanced high-frequency features. These features are then upsampled to restore the original resolution. The specific process is as follows:

$$\begin{aligned} F_{conv}^{(i)} &= \mathrm{Conv}(F_{dct}^{(i)}), \\ F_{up}^{(i)} &= \mathrm{Up}_i(F_{conv}^{(i)}), \end{aligned} \tag{13}$$

$\mathrm{Up}(\cdot)$ denotes the upsampling operation. Subsequently, the high-frequency features restored at different scales are fused through channel concatenation and 3×3 convolutional operation. Finally, residual learning is utilized to improve the model's ability to recover high-frequency image details, ultimately yielding the high-frequency fused feature F_S^{EE}. The process is as follows:

$$F_S^{EE} = \mathrm{Conv}(\mathrm{cat}(F_{up}^{(0)}, F_{up}^{(1)}, F_{up}^{(2)})) + F_S. \tag{14}$$

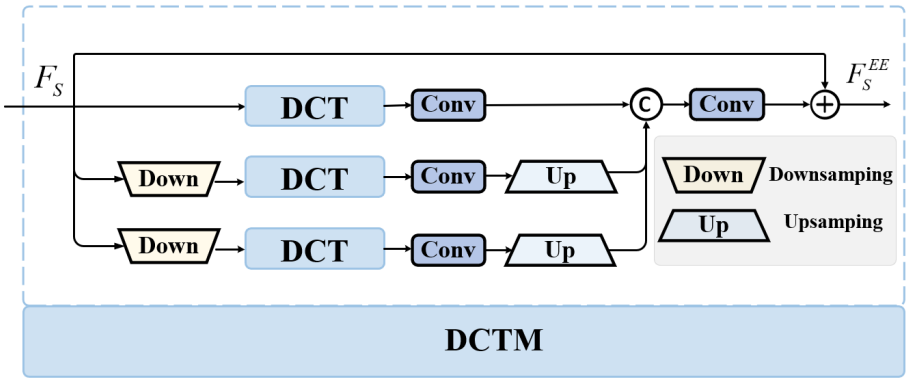


Figure 4. The architecture of our multi-scale DCT high-frequency feature restoration module. It extracts high-frequency features at different scales and integrates them with a CNN network to restore the corresponding high-frequency features for each scale. Through deep fusion, it effectively enhances the high-frequency components.

3.3.2. High-Frequency Information Recovery in the Spatial Domain

We further reconstruct the high-frequency image details of the SAIs from the spatial domain perspective. Transformer-based image reconstruction [34–36] has demonstrated remarkable performance, primarily due to the capability of the self-attention mechanism to capture global contextual information, thereby achieving higher-quality high-resolution image. However, the self-attention mechanism limits capability in capturing local details. Therefore, as illustrated in Figure 5, we employ the multi-scale detail enhancement module (MDEM) [37] to extract the detailed information in the SAI and utilize the cross-attention mechanism to improve the reconstruction of high-frequency details in the SAIs. By jointly utilizing local and global feature interactions, this method enhances the accuracy of image reconstruction.

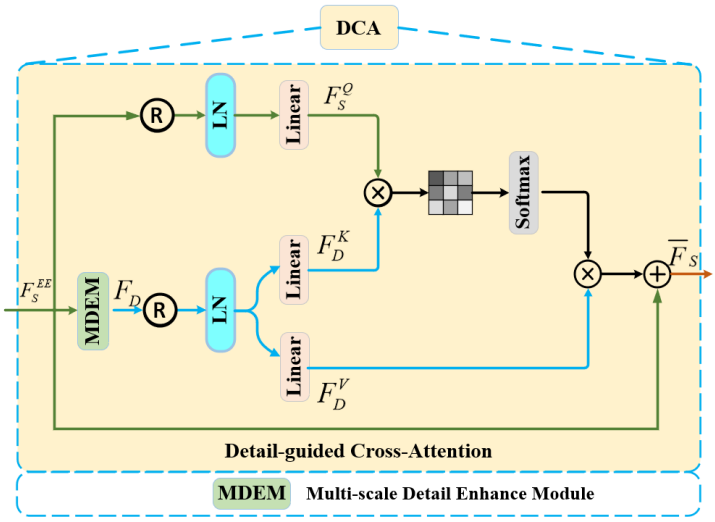


Figure 5. The architecture of our detail-guided cross-attention module. It utilizes the detailed texture features extracted by the MDEM module as a prior to perform cross-attention with the high-frequency features. This process guides the restoration of high-frequency details in the spatial domain while suppressing the influence of noise.

The specific approach is divided into the following steps: First, the high-frequency-enhanced feature $F_S^{EE} \in \mathbb{R}^{C\times U\times V\times H\times W}$ is processed through the multi-scale detail enhancement module (MDEM) to initially obtain the detail-enhanced feature $F_D \in \mathbb{R}^{C\times U\times V\times H\times W}$. Subsequently, a reshape operation is applied to transform it into spatially related sequence tokens, denoted as $T_s \in \mathbb{R}^{HW\times UV\times C}$. The process is as follows:

$$T_S = \text{reshape}(\text{MDEM}(F_S^{EE})), \tag{15}$$

reshape$(\cdot)$ indicates the feature shape reconstruction, and MDEM$(\cdot)$ refers to the multi-scale detail enhancement module in the spatial domain.

Subsequently, by combining layer normalization and linear mapping operations, the detail-enhanced features are mapped into two distinct feature spaces, denoted as $F_D^K, F_D^V \in \mathbb{R}^{HW\times UV\times C}$, to enhance the representation of high-frequency features. Similarly, F_S^{EE} can also be mapped into the corresponding feature space to obtain $F_D^Q \in \mathbb{R}^{HW\times UV\times C}$. The corresponding formulas are as follows:

$$\begin{aligned} F_s^K &= \text{Linear}(\text{LN}(T_s)), \\ F_s^V &= \text{Linear}(\text{LN}(T_s)), \\ F_S^Q &= \text{Linear}(\text{LN}(\text{reshape}(F_S^{EE}))), \end{aligned} \tag{16}$$

LN$(\cdot)$ denotes layer normalization; Linear$(\cdot)$ represents the linear mapping process.

Then, we employ a cross-attention operation to transfer the high-frequency feature F_S^{EE}, extracted from the frequency domain, to the detail-enhanced feature F_D in the spatial domain, thereby achieving further enhancement of high-frequency features. This results in the spatially enhanced feature $\overline{F_S}$, which incorporates high-frequency enhancements from both the spatial and frequency domains. The process is as follows:

$$\overline{F_S} = \frac{\text{softmax}(F_S^Q \cdot F_D^{K^T})}{\lambda} \cdot F_D^V, \tag{17}$$

here, λ represents a learnable scaling factor.

3.4. Spatial–Angular Feature Fusion

Mamba-Based Fusion Module

After capturing the high-frequency features in the spatial and angular domains, we employ a Mamba-based [24] spatial–angular fusion module to globally integrate spatial and angular information. By employing linear recurrent operations to model long-range dependencies, it significantly reduces computational burden while preserving image reconstruction capabilities. The convolutional kernels of CNNs have local receptive fields, resulting in low computational cost when fusing spatial–angular information in light fields. However, they have limited capability in capturing long-range spatial–angular dependencies. In contrast, the global attention mechanism of Transformers can capture long-range dependencies but introduces substantial memory and computational overhead.

Specifically, as illustrated in Figure 6, we first reshape the light field high-frequency feature $\overline{F_s} \in \mathbb{R}^{C\times U\times V\times H\times W}$ into a sequence of tokens $T \in \mathbb{R}^{UVHW\times C}$, which preserves spatial and angular information of LF. Here, the number of tokens is $UVHW$, and the dimensionality of each token is C. Subsequently, a fully connected layer is used to encode the sequence information, yielding the preliminary spatial–angular interaction sequence feature $T^{'} \in \mathbb{R}^{UVHW\times C}$. Then, a forward 1D convolution operation is applied to obtain the input feature x_1^f for the first time step, and a forward state-space model (SSM) [38] is employed to derive the cumulative hidden state feature $h_t^{forward}$ for each time step t relative to the previous $t-1$ time steps. Similarly, a backward 1D convolution operation and a backward state-space model are used to obtain the cumulative hidden state feature $h_t^{backward}$ for each time step t relative to the subsequent time step. The attention features obtained from the backward process fuse with the forward attention features through addition, resulting in y. Finally, the fused features are decoded and reshaped into the four-dimensional representation of the light field image, resulting in $F_{fusion} \in \mathbb{R}^{C\times U\times V\times H\times W}$.

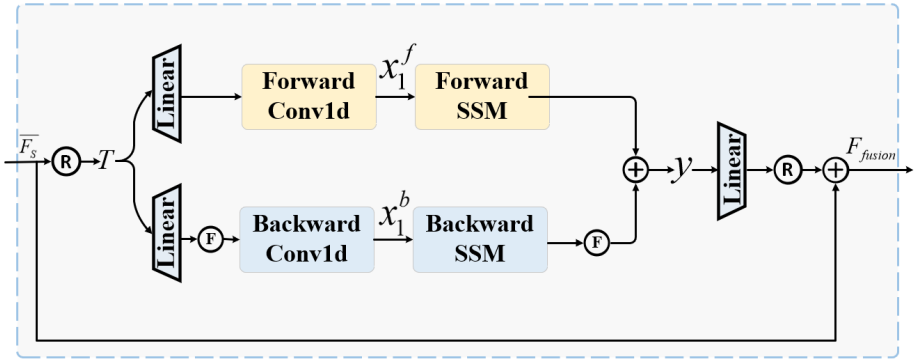


Figure 6. The architecture of Mamba-based spatial–angular feature fusion module. It leverages the complementary strengths of the forward and backward SSMs to rapidly and efficiently capture complex dependencies in the spatial and angular dimensions. By integrating a feature fusion block, it achieves robust and high-quality feature restoration.

Taking the forward state-space model as an example, the state update and observation equation are expressed as follows:

$$\begin{aligned} h_t^{forvard} &= Ah_{t-1}^{forvard} + Bx_t^f, \\ y_t &= Ch_t^{forvard} + Dx_t^f, \end{aligned} \tag{18}$$

here, $x_t^f \in \mathbb{R}^C$ represents the input feature at time step t, $y_t \in \mathbb{R}^C$ denotes the output feature at the corresponding time step, $A \in \mathbb{R}^{C\times C}$ is the state transition matrix, modeling the transition process from time step $t-1$ to t, $B \in \mathbb{R}^{C\times C}$ is the mapping matrix that

projects the input feature x_t into the state space, $C \in \mathbb{R}^{C\times 1}$ is the mapping matrix that projects the hidden state $h_t^{forward}$ into the output space, and $D \in \mathbb{R}^{C\times C}$ represents a residual connection process, representing the contribution of the initial input state x_1^f to the final output state y, which we set to 0 in our method. The hidden state $h_t^{forward}$ captures the continuously accumulated global contextual information from the previous $t-1$ time steps. Through the hidden state $h_t^{forward}$, the propagation paths of high-frequency components in the spatial–angular dimensions can be captured, ensuring the consistency of high-frequency details across different sub-views and spatial locations on a global scale. In summary, to address the causal modeling limitations of Mamba, a bidirectional Mamba architecture tailored to the structural characteristics of light field images is introduced. By performing forward and backward state propagation along the angular dimensions, the proposed model effectively captures complementary information across views and spatial correlations among SAIs, thereby enhancing angular consistency and high-frequency detail reconstruction in LFSR tasks.

Finally, the features $F_{fusion} \in \mathbb{R}^{C\times U\times V\times H\times W}$ are processed using a pixel shuffle operation [39] to perform upsampling to generate a high-resolution LF image $\mathcal{L}_{out} \in \mathbb{R}^{U\times V\times \alpha H\times \alpha W}$.

4. Experiments

In this section, we first introduce the datasets, training strategy, and evaluation metrics adopted in our experiments. Then, we conduct both quantitative and qualitative comparisons with state-of-the-art methods. Finally, we perform ablation studies to investigate the contribution of each component within the proposed framework.

4.1. Datasets and Training Strategy and Evaluation Metric

Datasets. We employ three types of light field datasets: HCInew [40], HCIold [41], and STFgantry [42]. The HCInew dataset includes 24 distinct scenes, covering both indoor and outdoor environments, with 20 scenes used for training and 4 scenes used for testing, and the size of the sub-aperture images is 512×512. The HCIold dataset contains 12 high-quality scenes, with 10 scenes for training and 2 scenes for testing, and the size of the sub-aperture images is 768×768. The STFgantry dataset consists of 11 scenes captured using a high-precision robotic arm, providing extremely high spatial and angular resolutions, with 9 scenes used for training and 2 scenes for testing, and the size of the sub-aperture images is 512×512. During data acquisition, we retain the central 25 SAIs for each LF and perform slicing on each LF image. Patch sizes of 64 and 128 are used for $2\times$ and $4\times$ super-resolutions, respectively, during both training and inference.

Training strategy. To train our network, we adopt the following procedures and configurations. First, each training sample consists of sub-aperture image patches corresponding to different sub-views. The network is optimized using the L1 loss to ensure consistency between the reconstructed high-frequency details and the ground truth. Next, we use the Adam optimizer with an initial learning rate of 1×10^{-4} and apply a learning rate decay strategy during training. The batch size is set to 2, and the network is trained for a total of 100 epochs.

Evaluation metric. We adopt PSNR and SSIM as objective evaluation metrics to assess the reconstruction performance.

PSNR is computed from the Mean Squared Error (MSE) between the reconstructed and ground-truth high-resolution (HR) images, reflecting the accuracy of high-frequency detail recovery. For the image super-resolution task, given a reconstructed image I_{pre} and an HR image I_{gt}, both of size $h \times w$, the MSE is computed as follows:

$$MSE = \frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w}\Big[I_{pred}(i,j) - I_{gt}(i,j)\Big]^2, \tag{19}$$

here, h and w represent the height and width of the reconstructed image and the high-resolution image, respectively, while I_{pre} and I_{gt} represent the pixel values of the reconstructed image and the HR image, respectively. Thus, the PSNR can be expressed as

$$PSNR = 10 \cdot \log_{10}(\frac{MAX_I^2}{MSE}), \tag{20}$$

here, MAX_I denotes the maximum possible pixel value of the image, which is 255 for 8-bit representations. PSNR is typically measured in decibels (dBs), where a higher value indicates lower reconstruction error and thus better image quality.

SSIM assesses image similarity by jointly evaluating luminance, contrast, and structural consistency between the reconstructed and reference images. Given an HR image x and a reconstructed image y, the SSIM is calculated as

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{21}$$

Here, μ_x and μ_y denote the mean values of images x and y, respectively, σ_x^2 and σ_y^2 represent their corresponding variances, and σ_{xy} indicates the covariance between them. The constants c_1 and c_2 are introduced to stabilize the division and prevent numerical instability. SSIM values range from 0 and 1, where a value of 1 indicates perfect structural similarity, and 0 denotes no structural correlation.

4.2. Quantitative Evaluation

To validate the effectiveness of the proposed light field spatial super-resolution algorithm, we conduct quantitative comparisons with traditional methods and deep-learning-based methods [12,15,17,29,43–53]. All competing methods are trained on three light field datasets, and the input low-resolution images are super-resolved by factors of 2 and 4. As shown in Tables 1 and 2, our network, which effectively extracts high-frequency features and fuses spatial–angular information, achieves superior performance in both PSNR and SSIM compared to other methods. All methods, including ours, operate on the luminance (Y) channel in the YCbCr color space for super-resolution. Finally, the results of all methods are converted back to the RGB color space, and PSNR and SSIM are calculated between the reconstructed images and the HR ground truth (GT) images.

Table 1. PSNR and SSIM values achieved by different methods on 5 × 5 LFs for 2× SR. Best and second best results are marked in **red** and cyan, respectively.

Methods	Scale	#Params.	HClnew	HClold	STFgantry
Bilinear	×2	–	30.718/0.9192	36.243/0.9709	29.577/0.9310
Bicubic	×2	–	31.887/0.9356	37.686/0.9785	31.063/0.9498
VDSR [54]	×2	0.665 M	34.371/0.9561	40.606/0.9867	35.541/0.9789
EDSR [55]	×2	38.62 M	34.828/0.9592	41.014/0.9874	36.296/0.9818
RCAN [43]	×2	15.31 M	35.022/0.9603	41.125/0.9875	36.670/0.9831
resLF [44]	×2	7.982 M	36.685/0.9739	43.422/0.9932	38.354/0.9904
LFSSR [15]	×2	0.888 M	36.802/0.9749	43.811/0.9938	37.944/0.9898
LF-ATO [45]	×2	1.216 M	37.244/0.9767	44.205/0.9942	39.636/0.9929

Table 1. *Cont.*

Methods	Scale	#Params.	HClnew	HClold	STFgantry
LF_InterNet [46]	×2	5.040 M	37.170/0.9763	44.573/0.9946	38.435/0.9909
LF-DFnet [47]	×2	3.940 M	37.418/0.9773	44.198/0.9941	39.427/0.9926
MEG-Net [48]	×2	1.693 M	37.424/0.9777	44.097/0.9942	38.767/0.9915
LF-IINet [49]	×2	4.837 M	37.768/0.9790	44.852/0.9948	39.894/0.9936
DPT [17]	×2	3.731 M	37.355/0.9771	44.302/0.9943	39.429/0.9926
LFT [29]	×2	1.114 M	37.838/0.9791	44.522/0.9945	40.510/0.9941
DistgSSR [12]	×2	3.532 M	37.959/0.9796	44.943/0.9949	40.404/0.9942
LFSSR_SAV [50]	×2	1.217 M	37.425/0.9776	44.216/0.9942	38.689/0.9914
EPIT [51]	×2	1.421 M	38.228/0.9810	45.075/0.9949	42.166/0.9957
HLFSR-SSR [52]	×2	13.72 M	**38.317**/**0.9807**	44.978/0.9950	40.849/0.9947
LF-DET [53]	×2	1.588 M	38.314/**0.9807**	44.986/0.9950	41.762/0.9955
ours	×2	1.352 M	38.309/**0.9807**	**45.361**/**0.9964**	**42.238**/**0.9962**

Table 2. PSNR and SSIM values achieved by different methods on 5 × 5 LFs for 4× SR. Best and second best results are marked in **red** and cyan, respectively.

Methods	Scale	#Params.	HClnew	HClold	STFgantry
Bilinear	×4	–	27.085/0.8397	31.688/0.9256	25.203/0.8261
Bicubic	×4	–	27.715/0.8517	32.576/0.9344	26.087/0.8452
VDSR [54]	×4	0.665 M	29.308/0.8823	34.810/0.9515	28.506/0.9009
EDSR [55]	×4	38.89 M	29.591/0.8869	35.176/0.9536	28.703/0.9072
RCAN [43]	×4	15.36 M	29.694/0.8886	35.359/0.9548	29.021/0.9131
resLF [44]	×4	8.646 M	30.723/0.9107	36.705/0.9682	30.191/0.9372
LFSSR [15]	×4	1.774 M	30.928/0.9145	36.907/0.9696	30.570/0.9426
LF-ATO [45]	×4	1.364 M	30.880/0.9135	36.999/0.9699	30.607/0.9430
LF_InterNet [46]	×4	5.483 M	30.961/0.9161	37.150/0.9716	30.365/0.9409
LF-DFnet [47]	×4	3.990 M	31.234/0.9196	37.321/0.9718	31.147/0.9494
MEG-Net [48]	×4	1.775 M	31.103/0.9177	37.287/0.9716	30.771/0.9453
LF-IINet [49]	×4	4.886 M	31.331/0.9208	37.620/0.9734	31.261/0.9502
DPT [17]	×4	3.778 M	31.196/0.9188	37.412/0.9721	31.150/0.9488
LFT [29]	×4	1.163 M	31.462/0.9218	37.630/0.9735	31.860/0.9548
DistgSSR [12]	×4	3.582 M	31.380/0.9217	37.563/0.9732	31.649/0.9535
LFSSR_SAV [50]	×4	1.543 M	31.450/0.9217	37.497/0.9721	31.362/0.9505
EPIT [51]	×4	1.470 M	31.511/0.9231	37.677/0.9737	32.179/0.9571
HLFSR-SSR [52]	×4	13.87 M	**31.571**/**0.9238**	37.776/0.9742	31.641/0.9537
LF-DET [53]	×4	1.687 M	31.558/0.9235	37.843/0.9744	32.139/0.9573
Ours	×4	1.402 M	31.545/0.9229	**38.024**/**0.9812**	**32.341**/**0.9607**

4.3. Qualitative Evaluation

Figures 7–9 represent the comparative results of our method and other methods for 2× and 4× super-resolutions, respectively. Both our method and all compared methods perform spatial super-resolution on all SAIs of the LF image simultaneously. For clarity and ease of comparison, we only display the super-resolution results of the central SAI. As shown in Figures 7–9, the first row, from left to right, shows the Bilinear, LFSSR, LF-InterNet, LFT, and DistgSSR. The second row, from left to right, presents the results of EPIT, HLFSR-SSR, LF-DET, our proposed method, and the HR LF image (GT). Our method effectively captures the inter-view correlations among sub-aperture images while comprehensively recovering high-frequency details from both the frequency and spatial domains. As illustrated in Figure 7, we perform super-resolution on scenes containing mesh-structured objects, such as a storage basket and a trash can, and observe that our method produces sharper object boundaries. This improvement can be attributed to the proposed LASS, which does not rely on long-range occluded scene points. Instead, it recovers high-frequency details by leveraging angular correlations within a limited range of views, thereby mitigating the influence of occluded regions. Furthermore, as shown in Figures 8 and 9, our method leverages Mamba-based global spatial–angular modeling to recover fine surface details even in smooth regions of the scene.

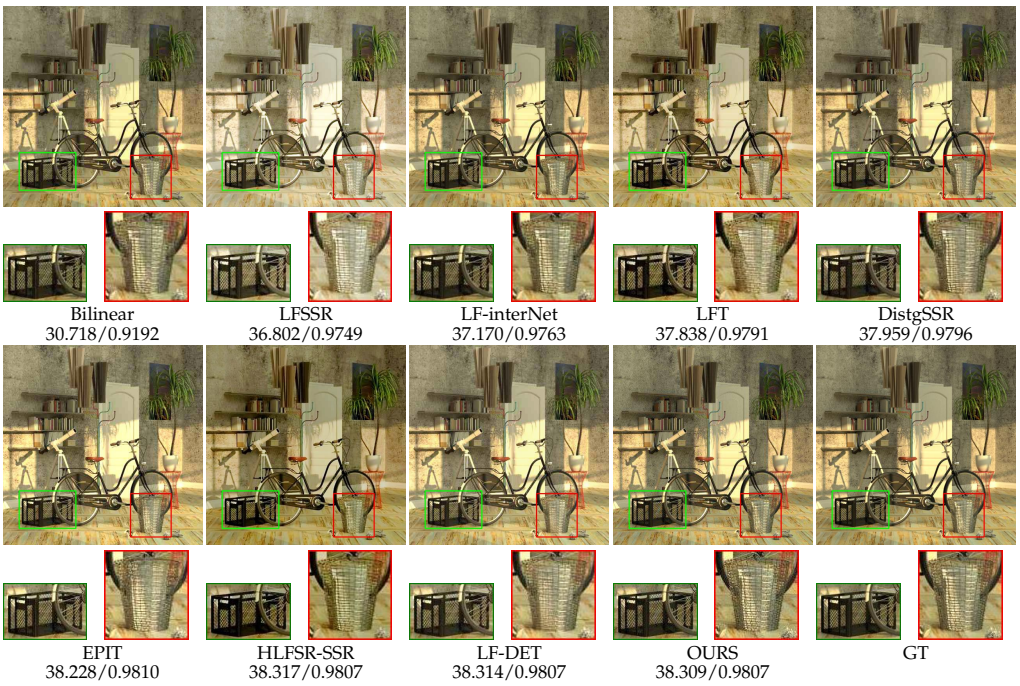


Figure 7. Visual results achieved by different methods for 2× SR in bicycle/HCInew [40] dataset.

4.4. Ablation Study

In this subsection, we evaluate variants of our network to demonstrate the effectiveness of the local sparse angular attention module (LSAA), multi-scale detail enhancement module (MDEM), DCT high-frequency feature restoration module (DCTM), Mamba-based fusion module (MF), and Mamba-based fusion direction. We conducted ablation experiments on the HCInew dataset to evaluate the effectiveness of each module in our network for the 2× super-resolution task.

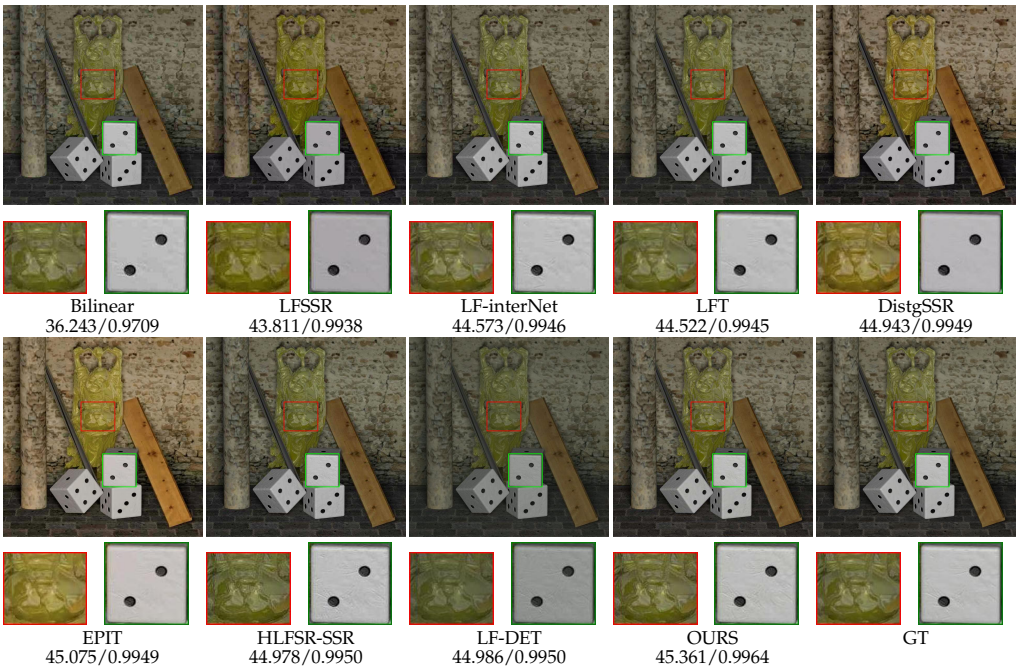


Figure 8. Visual results achieved by different methods for 2× SR in buddha/HCIold [41] dataset.

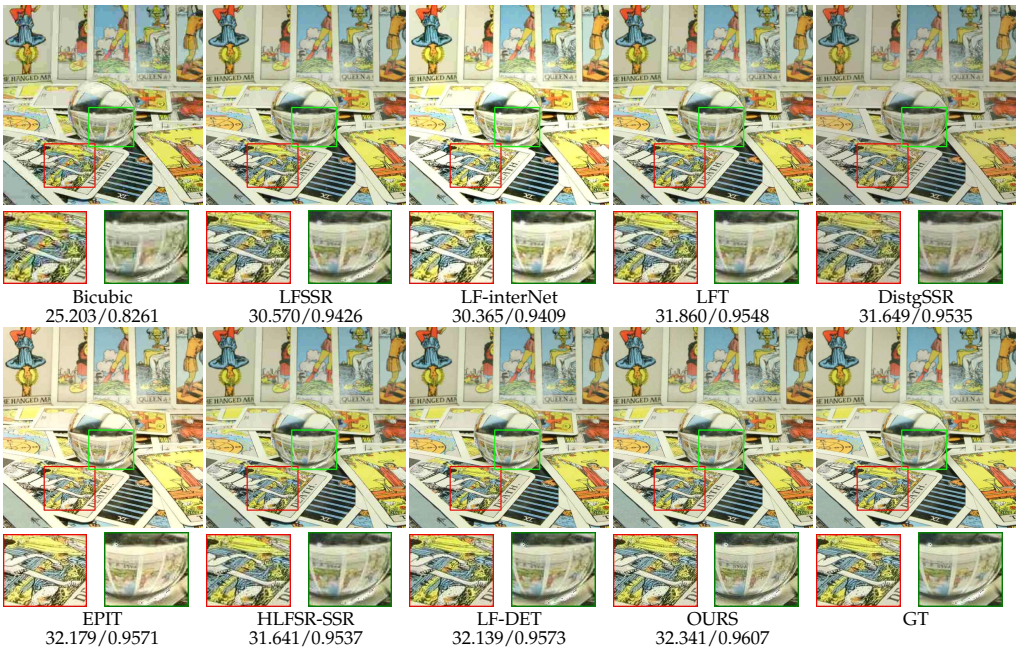


Figure 9. Visual results achieved by different methods for 4× SR in Tarot Cards S/STFgantry [42] dataset.

Effectiveness of local sparse angular attention module. We replace the LSAA with a self-attention mechanism. Due to the presence of disparity in LF images, SAIs farther from the reference view are more likely to be affected by occlusions, introducing interference that hinders the recovery of high-frequency details from other views. As a result, LSAA achieves better overall performance than global self-attention (SA). As shown in Table 3, the computational cost increased by 64%, while the PSNR and SSIM metrics decreased by 0.795 and 0.006, respectively.

Table 3. The PSNR and SSIM values obtained by several variants of our network tested on the HCInew dataset for 2× SR. "LSAA→SA" refers to replacing our LSAA with self-attention. "w/o MDEM" represents removing MDEM. "w/o DCTM" represents removing DCT module. "w/o MF" means removing our MF. "MF→SA" refers to replacing our MF with self-attention. "↑" indicates that higher values represent better performance. "↓" indicates that lower values represent better performance. The best performances are marked in **bold**.

Method	Gflops	Param ↓	PSNR ↑	SSIM ↑
LSAA→SA	71.943	1.352M	37.514	0.9743
w/o MDEM	**35.734**	**1.038M**	36.952	0.9688
w/o DCTM	41.390	1.349M	37.216	0.9721
w/o MF	39.566	1.202M	36.871	0.9639
MF→SA	97.293	1.416M	36.295	0.9577
Ours	43.868	1.352M	**38.309**	**0.9807**

Effectiveness of multi-scale detail enhancement module. After removing the MDEM from our detail-guided cross-attention module, the network fails to capture high-frequency detail regions, which adversely affects the recovery of high-frequency details. As shown in Table 3, we observe the PSNR and SSIM metrics decrease by 1.357 and 0.012, respectively.

Effectiveness of DCT high-frequency restoration module. When we remove DCTM from our network, the network fails to effectively extract high-frequency information from the frequency domain. As shown in Table 3, PSNR and SSIM metrics decrease by 1.093 and 0.009, respectively.

Effectiveness of Mamba-based fusion module. When we remove MF, the network fails to efficiently fuse the spatial and angular information of the LF image, resulting in a lack of global consistency in the high-frequency information extracted from two perspectives. As shown in Table 3, PSNR and SSIM metrics decrease by 1.438 and 0.017, respectively.

Effectiveness of Mamba-based fusion direction. Based on the state-space model, we fuse spatial and angular information in different directions. As shown in Table 4, we notice that the sequential fusion of spatial information followed by angular information achieves the highest PSNR and SSIM metrics, reaching 38.309 and 0.9807, respectively.

Table 4. The PSNR and SSIM values obtained by several variants of Mamba-based fusion on the HCInew dataset for 2× SR. "$UV \times HW$" indicates that spatial–angular information is fused in the order of spatial dimensions (HW) followed by angular dimensions (UV). "$UH \times VW$" denotes the fusion in the order of horizontal EPI first, followed by vertical EPI. "$VW \times UH$" denotes the fusion in the order of vertical EPI first, followed by horizontal EPI. "$HW \times UV$" denotes the fusion in the order of spatial first, followed by angular. "↑" indicates that higher values represent better performance. The best performances are marked in **bold**.

Direction	$\boldsymbol{UV \times HW}$	$\boldsymbol{UH \times VW}$	$\boldsymbol{VW \times UH}$	$\boldsymbol{HW \times UV}$
PSNR ↑	**38.309**	37.902	37.884	38.288
SSIM ↑	**0.9807**	0.9747	0.9726	0.9792

Effectiveness of the local window range. We adjust the local window range by employing 2×2 and 4×4 local windows, respectively. As shown in Table 5, the results illustrate that a 2×2 local window is too small to effectively perform global modeling of angular information, leading to decreases of 0.305 and 0.004 in PSNR and SSIM, respectively. In contrast, the 4×4 local window covers an excessively large angular range, where distant views are more susceptible to occlusions or depth variations. This introduces interference in the

recovery of high-frequency image details, resulting in PSNR and SSIM decreases of 0.373 and 0.007, respectively.

Table 5. The PSNR and SSIM values obtained by several variants of local window range in LSAA on the HCInew dataset for 2× SR. "2 × 2" denotes the 2 × 2 local windows. "3 × 3" denotes the 3 × 3 local windows, which are adopted in our model. "4 × 4" denotes the 4 × 4 local windows. "↑" indicates that higher values represent better performance. The best performances are marked in **bold**.

Local Window Range	2 × 2	3 × 3	4 × 4
PSNR ↑	38.004	**38.309**	37.936
SSIM ↑	0.9762	**0.9807**	0.9733

5. Conclusions

In this paper, we propose an innovative light field super-resolution network, which consists of a local sparse angular attention module (LSAA), a dual-domain high-frequency detail recovery sub-network, and a Mamba-based fusion module (MF). The LSAA recovers high-frequency information from neighboring sub-views by modeling inter-view relationships. The dual-domain high-frequency detail recovery sub-network thoroughly explores high-frequency features from both the spatial and frequency domains, thereby enhancing the recovery of texture details. The MF efficiently integrates spatial and angular information using a bidirectional state-space model to preserve the global consistency of high-frequency features across both spatial and angular perspectives. Extensive experiments demonstrate that our method achieves excellent performance in the light field super-resolution (LFSR) task. Although the experiments have achieved promising performance, the number of scenes in the employed datasets is limited. As a result, the model cannot fully exploit the effectiveness of the proposed LSAA module and may fail to address real-world challenges, such as extreme occlusions or motion blur. In future work, we will collect more real-world scene datasets for network training to further enhance the generalization capability of our model.

Author Contributions: Conceptualization, T.Y.; methodology, Z.Z.; visualization, C.W. (Cihang Wei); investigation, Z.Z.; resources, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, J.L., H.H. and C.W. (Chenglong Wang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (Grant No. 61902151).

Data Availability Statement: The dataset is available at https://github.com/lightfield-analysis/resources, accessed on 1 March 2024.

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

References

1. Wu, G.; Masia, B.; Jarabo, A.; Zhang, Y.; Wang, L.; Dai, Q.; Chai, T.; Liu, Y. Light field image processing: An overview. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 926–954. [CrossRef]
2. Kim, C.; Zimmer, H.; Pritch, Y.; Sorkine-Hornung, A.; Gross, M.H. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph. TOG* **2013**, *32*, 73:1–73:4. [CrossRef]
3. Zhang, Q.; Li, H.; Wang, X.; Wang, Q. 3D scene reconstruction with an un-calibrated light field camera. *Int. J. Comput. Vis.* **2021**, *129*, 3006–3026. [CrossRef]
4. Zhou, Y.; Guo, H.; Fu, R.; Liang, G.; Wang, C.; Wu, X. 3D reconstruction based on light field information. In Proceedings of the 2015 IEEE International Conference on Information and Automation (ICIA), Lijiang, China, 8–10 August 2015; IEEE: New York, NY, USA, 2015; pp. 976–981.

5. Overbeck, R.S.; Erickson, D.; Evangelakos, D.; Pharr, M.; Debevec, P. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Trans. Graph. TOG* **2018**, *37*, 1–15. [CrossRef]
6. Yu, J. A light-field journey to virtual reality. *IEEE Multimed.* **2017**, *24*, 104–112. [CrossRef]
7. Lam, E.Y. Computational photography with plenoptic camera and light field capture: Tutorial. *J. Opt. Soc. Am. A* **2015**, *32*, 2021–2032. [CrossRef]
8. Levoy, M. Light fields and computational imaging. *Computer* **2006**, *39*, 46–55. [CrossRef]
9. Choy, C.; Gwak, J.; Savarese, S. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3075–3084.
10. Deeba, F.; Kun, S.; Dharejo, F.A.; Zhou, Y. Wavelet-based enhanced medical image super resolution. *IEEE Access* **2020**, *8*, 37035–37044. [CrossRef]
11. Beaini, D.; Passaro, S.; Létourneau, V.; Hamilton, W.; Corso, G.; Liò, P. Directional graph networks. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, 18–24 July 2021; pp. 748–758.
12. 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*, 425–443. [CrossRef]
13. 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]
14. Wang, X.; Ma, J.; Yi, P.; Tian, X.; Jiang, J.; Zhang, X.P. Learning an epipolar shift compensation for light field image super-resolution. *Inf. Fusion* **2022**, *79*, 188–199. [CrossRef]
15. Yeun, 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 TIP* **2018**, *28*, 2319–2330.
16. Yan, T.; Jiao, J.; Liu, W.; Lau, R.W. Stereoscopic image generation from light field with disparity scaling and super-resolution. *IEEE Trans. Image Process.* **2019**, *29*, 1827–1842. [CrossRef]
17. Wang, S.; Zhou, T.; Lu, Y.; Di, H. Detail-preserving transformer for light field image super-resolution. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI), Virtual, 22 February–1 March 2022; Volume 36, pp. 2522–2530.
18. Xu, R.; Kang, X.; Li, C.; Chen, H.; Ming, A. DCT-FANet: DCT based frequency attention network for single image super-resolution. *Displays* **2022**, *74*, 102220. [CrossRef]
19. Dai, T.; Wang, J.; Guo, H.; Li, J.; Wang, J.; Zhu, Z. FreqFormer: Frequency-aware transformer for lightweight image super-resolution. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI), Jeju, Republc of Korea, 3–9 August 2024; pp. 731–739.
20. 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]
21. Ghassab, V.K.; Bouguila, N. Light field super-resolution using edge-preserved graph-based regularization. *IEEE Trans. Multimed.* **2019**, *22*, 1447–1457. [CrossRef]
22. Gao, C.; Lin, Y.; Chang, S.; Zhang, S. Spatial-angular multi-scale mechanism for light field spatial super-resolution. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 1961–1970.
23. Liu, G.; Yue, H.; Wu, J.; Yang, J. Efficient light field angular super-resolution with sub-aperture feature learning and macro-pixel upsampling. *IEEE Trans. Multimed.* **2022**, *25*, 6588–6600. [CrossRef]
24. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
25. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. *NeurIPS* **2024**, *37*, 103031–103063.
26. Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. Videomamba: State space model for efficient video understanding. In Proceedings of the 18th European Conference (ECCV), Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 237–255.
27. Pendry, J.B. Negative Refraction Makes a Perfect Lens. *Phys. Rev. Lett.* **2000**, *85*, 3966–3969. [CrossRef]
28. Liu, J.; Sun, W.; Wu, F.; Shan, H.; Xie, X. Macroscopic Fourier Ptychographic Imaging Based on Deep Learning. *Photonics* **2025**, *12*, 170. [CrossRef]
29. Liang, Z.; Wang, Y.; Wang, L.; Yang, J.; Zhou, S. Light field image super-resolution with transformers. *IEEE Signal Process. Lett.* **2022**, *29*, 563–567. [CrossRef]

30. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arxiv* **2024**, arXiv:2401.09417. [CrossRef]
31. Hsu, R.; Kodama, K.; Harashima, H. View interpolation using epipolar plane images. In Proceedings of the 1st International Conference on Image Processing (ICIP), Austin, TX, USA, 13–16 November 1994; IEEE: New York, NY, USA, 1994; Volume 2, pp. 745–749.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *NeurIPS* **2017**, *30*, 11.
33. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
34. Park, S.C.; Park, M.K.; Kang, M.G. Super-resolution image reconstruction: A technical overview. *IEEE Signal Process. Mag.* **2003**, *20*, 21–36. [CrossRef]
35. Defrise, M.; Gullberg, G.T. Image reconstruction. *Phys. Med. Biol.* **2006**, *51*, R139. [CrossRef]
36. Demoment, G. Image reconstruction and restoration: Overview of common estimation structures and problems. *IEEE Trans. Autom. Sci. Eng.* **1989**, *37*, 2024–2036. [CrossRef]
37. Gao, S.; Zhang, P.; Yan, T.; Lu, H. Multi-scale and detail-enhanced segment anything model for salient object detection. In Proceedings of the ACM Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 9894–9903.
38. Hamilton, J.D. State-space models. *Handb. Econom.* **1994**, *4*, 3039–3080.
39. 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
40. Honauer, K.; Johannsen, O.; Kondermann, D.; Goldluecke, B. A dataset and evaluation methodology for depth estimation on 4D light fields. In Proceedings of the 13th Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, 20–24 November 2016; Springer: Berlin/Heidelberg, Germany, 2017; pp. 19–34.
41. Wanner, S.; Meister, S.; Goldluecke, B. Datasets and benchmarks for densely sampled 4D light fields. In Proceedings of the 18th International Workshop on Vision, Modeling and Visualization (VMV), Lugano, Switzerland, 11–13 September 2013; Volume 13, pp. 225–226.
42. 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]
43. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
44. Zhang, S.; Lin, Y.; Sheng, H. Residual networks for light field image super-resolution. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11046–11055.
45. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 2260–2269.
46. Wang, Y.; Wang, L.; Yang, J.; An, W.; Yu, J.; Guo, Y. Spatial-angular interaction for light field image super-resolution. In Proceedings of the 16th European Conference (ECCV), Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 290–308.
47. Wang, Y.; Yang, J.; Wang, L.; Ying, X.; Wu, T.; An, W.; Guo, Y. Light field image super-resolution using deformable convolution. *IEEE Trans. Image Process.* **2020**, *30*, 1057–1071. [CrossRef]
48. Sarma, M.; Bond, C.; Nara, S.; Raza, H. MEGNet: A MEG-Based Deep Learning Model for Cognitive and Motor Imagery Classification. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 5–8 December 2023; IEEE: New York, NY, USA, 2023; pp. 2571–2578.
49. Liu, G.; Yue, H.; Wu, J.; Yang, J. Intra-inter view interaction network for light field image super-resolution. *IEEE Trans. Multimed.* **2021**, *25*, 256–266. [CrossRef]
50. Cheng, Z.; Liu, Y.; Xiong, Z. Spatial-angular versatile convolution for light field reconstruction. *IEEE Trans. Comput. Imaging* **2022**, *8*, 1131–1144. [CrossRef]
51. Liang, Z.; Wang, Y.; Wang, L.; Yang, J.; Zhou, S.; Guo, Y. Learning non-local spatial-angular correlation for light field image super-resolution. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 12376–12386.
52. Van Duong, V.; Huu, T.N.; Yim, J.; Jeon, B. Light field image super-resolution network via joint spatial-angular and epipolar information. *IEEE Trans. Comput. Imaging* **2023**, *9*, 350–366. [CrossRef]
53. Cong, R.; Sheng, H.; Yang, D.; Cui, Z.; Chen, R. Exploiting spatial and angular correlations with deep efficient transformers for light field image super-resolution. *IEEE Trans. Multimed.* **2023**, *26*, 1421–1435. [CrossRef]

54. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
55. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.

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.