This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

LF-BVN: Blind-View Network for Self-Supervised Light Field Denoising

Longzhao Guo, Shuo Zhang*, Chen Gao, Qian Tian, Youfang Lin
Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence,
School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China.
{longzhaoguo, zhangshuo, gaochen, tianqian, yflin}@bjtu.edu.cn

Abstract

Recent advances in learning-based Light Field (LF) image denoising have achieved impressive results. However, these methods rely heavily on large-scale noisy-clean image pairs and often fail to generalize to unseen or complex noise. In this work, we observe that the inherent multi-view consistency of LF images makes it highly unlikely for noise to be coherent across views, offering a more reliable supervisory signal for self-supervised denoising. Building on this insight, we extend the blind-spot principle to the LF domain and propose a novel LF Blind-View denoising Network (LF-BVN). We first introduce a geometric invariance mask that leverages angular redundancy for efficient full-view supervision. To enforce cross-view photometric consistency, we further introduce latent representation volumes and enforce consistency between them. Additionally, we exploit focus stacks to extract latent depth cues from noisy observations, providing further guidance. Extensive experiments show that LF-BVN achieves competitive denoising performance while maintaining strong cross-view consistency without requiring clean data or external supervision. Our code will be publicly accessible at https://github.com/shuozh/LF-BVN.

1. Introduction

Light Field (LF) images record abundant scene information by documenting the intensity and direction of light rays as they propagate in space. However, LF images captured using camera arrays [34], Lytro Illum [21], or Raytrix systems [25] are often affected by noise due to hardware limitations and suboptimal lighting conditions. Such noise significantly degrades the application of LF images in downstream tasks, such as view synthesis [10], super-resolution [40], and depth estimation [7, 39].

To mitigate these issues, several denoising approaches specifically designed for LF images have been proposed.

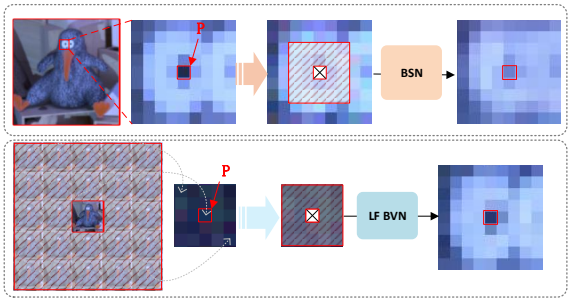


Figure 1. Comparison of blind-spot (upper) and blind-view (bottom) strategy. In BSN, the surrounding pixels in spatial domain are used to reconstruct the blind spot, which often leads to errors in reconstructing abrupt signal (such as the toy's eyes). In BVN, the contextual pixels in other views are used, enabling accurate reconstruction of such high-frequency details.

Early traditional LF denoising methods [1, 26] aimed to reduce noise by exploiting similar regions across multiple views. However, the results often exhibit significant artifacts and noticeable blurring. With the advancement of deep learning techniques, data-driven methods [4, 5, 11, 30] typically utilize noisy-clean image pairs as training data and have demonstrated superior performance in enhancing image quality. However, their performance hinges on precisely aligned noisy-clean pairs, which are exceptionally difficult to acquire and severely limit their applicability to real-world scenarios with complex noise patterns.

Motivated by the success of monocular self-supervised denoising methods [19, 24, 32], several approaches [31, 42] have extended the Noise2Noise [18] framework to enable self-supervised denoising in light field microscopy (LFM). VCDNet [31] models the degradation in LFM as a PSF-based convolution process and learns to reverse it using a volumetric convolutional neural network. V2V3D [42] partitions the LF views into two non-overlapping subsets and employs two independent networks for denoising different views. Although these methods perform well on LFM, these methods fail to effectively process LF images, due to two main weaknesses: First, these methods are overly reliant on the point spread function, which is absent in conventional

*Corresponding author: zhangshuo@bjtu.edu.cn

LF images due to fundamental differences in imaging principles. Second, processing different views by multiple independent networks makes it difficult to maintain view consistency and often results in visible artifacts.

Recently, blind-spot based denoising methods [3, 15, 17] rely on surrounding pixels to predict the blind pixel by assuming the spatial signal is continuous in natural images. The strong consistency across different views in LF images enables us to extend the blind-spot to a blind-view, which focuses on leveraging points in other views for denoising (Fig. 1). However, unlike blind-spot training with random masking, blind-view requires fixed masks to maintain geometric invariance. To denoise all views, multiple models with dedicated masks are needed, making it hard to ensure output consistency.

In this paper, we propose LF Blind-View Network (LFBVN), a novel self-supervised framework for LF denoising. First, since the geometric relationship is preserved when the views in LF are spatially and angularly rotated, we specially design a Geometric Invariance Mask (GIM) for the LFs. The proposed GIM masks only one-quarter of the views and can be implemented using a weight-sharing mechanism so that the network effectively utilizes information from all available views. Then we develop the blind-view denoising framework using a multi-branch structure. In order to preserve the view consistency between denoised views, we introduce a latent representation volume and design a reconstruction consistency loss to keep the representation consistent across all branches. Furthermore, by designing the depth estimation module, we extract and incorporate latent depth information from noisy LF to provide additional guidance for the denoising process. The proposed method demonstrates state-of-the-art performance on the synthetic, real-world and microscopy LF datasets, and exhibits exceptional generalization capability.

2. Related Work

Single Image Denoising: Recently, many supervised single image denoising approaches [2, 37, 38] have been proposed. Unfortunately, these methods require large amounts of noisy-clean image pairs for training and perform poorly when noise patterns are unknown. To alleviate these limitations, self-supervised denoising methods have received increasing attention. Noise2Noise [18] uses noisy-noisy pairs for training, which significantly reduces the difficulty of data collection. Similarly, R2R [22] corrupts noisy images with known noise levels to obtain training pairs. Neighbor2Neighbor [13] generates training pairs by sub-sampling the single noisy image. Furthermore, several methods have designed specific blind-spot network structures to process single noisy images directly. Noise2Self [3] and Noise2Void [15] were proposed to train networks with only one noisy observation per scene. Specifically, they leverage the independence of noise and the spatial correlation of clean signals, predicting the real signals using carefully designed blind-spot networks. Probabilistic Noise2Void [16], Laine19 [17], and Dilated Blind-Spot Network [35] further introduced explicit noise modeling and probabilistic inference for better performance. Blind2Unblind [32] proposed a re-visible loss to train the denoising network, effectively making blind spots visible during training. More recently, several methods have been proposed to address spatially correlated noise. Quan et al. [24] introduced a zero-shot blind-spot image denoising method, which achieves the removal of spatially correlated noise via implicit neural representation. TBSN [19] proposed a transformer-based self-attention mechanism for blind-spot denoising, coupled with U-Net distillation to accelerate inference.

LF Image Denoising: LF image denoising has also witnessed considerable progress. Early traditional methods predominantly relied on hand-crafted priors. For instance, Sepas-Moghaddam et al. [26] proposed converting noisy LF images into snake-like epipolar-plane image (EPI) sequences and applying V-BM4D [20] to achieve collaborative filtering. Dansereau et al. [9] derived that the spectral region of support in LF images is a 4D hyperfan, and utilized a 4D hyperfan filter to segregate the LF signal from noise (Hyperfan4D). LFBM5D [1] extended BM3D [8] to LF images by stacking disparity-compensated 4D patches to suppress noise. Based on Hyperfan4D, Premaratne et al. [23] proposed a novel 4D linear and shift-invariant hyperfan filter, which exploits the spectral sparsity of LF images. With the advent of deep learning, data-driven approaches have emerged. APA [6] introduced the first learning-based LF denoising method, consisting of two CNN modules. Wang et al. [30] proposed restoring noisy LF images with a multi-stream progressive network, which stacks LF images into three different configurations to provide complementary information. Chang et al. [4] proposed a symmetrical refocusing strategy to construct a focal stack for every view and designed a simple learning-based framework for LF denoising. Additionally, V2V3D [42] employed a Noise2Noise-based self-supervised network specifically for LF microscopy denoising.

3. Blind View Theory

The extension of blind-spot methodology to LF introduces unique challenges. In the following, we analyze the fundamental differences between monocular and LF blind-view denoising approaches, and present our strategy for designing view masks.

3.1. Blind-Spot Denoising Revisit

The underlying principle of monocular blind-spot denoising is to infer the value of a target pixel from its surrounding

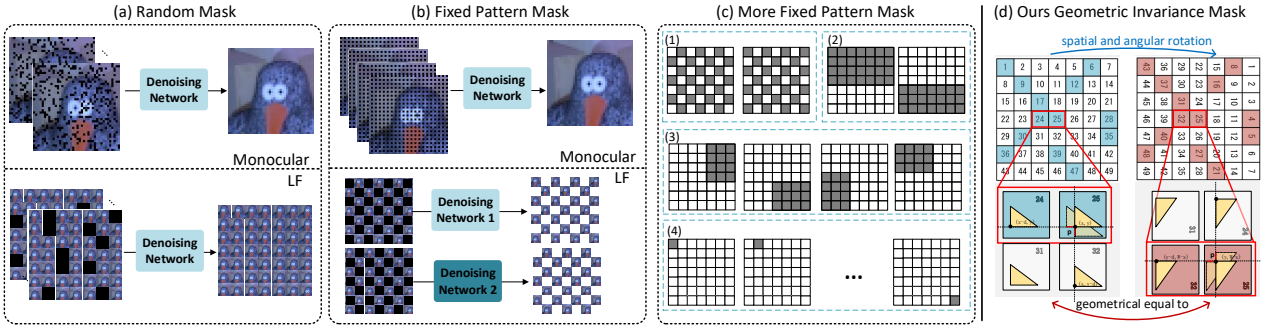


Figure 2. The mask design in blind-spot and blind-view network. The pixel or view is randomly masked (a) or fixed masked (b). However, due to the different position relationship between views, only with the fixed mask the blind view networks can learn the geometric correspondence in LF images. For other kinds of fixed patterns in (c), two or more models have to be trained for each pattern. We finally design our GIM as in (d), where the geometric maintain consistent as the LF rotates spatially and angularly. Employing our view mask only one model need to be trained and the photometric consistency between views can better maintained.

spatial context. Let $\tilde{\mathbf{y}} = \mathbf{y} + \mathbf{n}$ be a noisy image, where $\mathbf{y}$ denotes the ground-truth clean image and $\mathbf{n}$ denotes noise. The blind-spot-based denoising model is typically trained by minimizing the following empirical risk:

$$\arg\min_{\theta} ||f_\theta(\{\tilde{\mathbf{y}}(\bar{\mathbf{s}})\}_{\bar{s}\in \mathsf{N}(\mathbf{s})}) - \tilde{\mathbf{y}}(\mathbf{s})||_2^2, \tag{1}$$

where θ is a denoising network, $\mathbf{s}$ represent the spatial coordinates, $\mathsf{N}(\mathbf{s})$ denotes the index set of neighboring pixels of $\mathbf{y}(\mathbf{s})$, excluding $\mathbf{y}(\mathbf{s})$. This formulation requires the image to satisfy the following two conditions: (1) The noise $\mathbf{n}$ is zero-mean and pixel-wise independent. (2) The truth pixel value $\mathbf{y}(\mathbf{s})$ is strongly correlated with its neighboring pixels. Although these methods are capable of performing denoising, their reliance on local spatial context makes them fundamentally limited [15]. For example, noise patterns on flat regions can resemble textures and real discontinuities may be misinterpreted as noise.

3.2. Blind-Spot to Blind-View

In contrast, LF images capture multiple observations of the same scene from different angles, introducing strong geometric constraints that inherently disentangle true scene content from noise. This property effectively addresses the limitations of single-image blind-spot denoising. According to the two-plane parameterization model [14], a LF can be formulated as a 4D tensor $\mathcal{L}(\mathbf{a}, \mathbf{x}) \in \mathbb{R}^{U\times V\times H\times W}$, where $\mathbf{a} = (u, v)$ denotes the angular coordinates and $\mathbf{x} = (h, w)$ denotes the spatial coordinates. Typically, the LF image is organized as a $U \times V$ array of views that captures the same scene in one shot. The specific view at $\mathbf{a}_i$ can be denoted as $\mathbf{I}_i \in \mathbb{R}^{H\times W}$.

For a 3D point in the scene, its projections onto different views $\{\mathbf{I}_j(\mathbf{x}_j) \| j = 1, 2, ..., U \cdot V\}$ ideally exhibit photometric consistency under noise-free conditions: $\mathbf{I}_1(\mathbf{x}_1) = \mathbf{I}_2(\mathbf{x}_2) = ... = \mathbf{I}_{U\cdot V}(\mathbf{x}_{U\cdot V})$. However, in real-world scenarios, global consistency is often disrupted by factors such as occlusions or non-Lambertian reflections. Nevertheless, following the principles of blind-spot denoising, as long as the noisy projection $\tilde{\mathbf{I}}_i(\mathbf{x}_i)$ remains strongly correlated with a valid subset of other projections $\{\tilde{\mathbf{I}}_k(\mathbf{x}_k)\}_{k\neq i}$, this correlation provides sufficient informative context to predict the clean content of the target pixel. This process can be formulated as follows:

$$\arg\min_{\theta} ||f_\theta(\{\tilde{\mathbf{I}}_k(\mathbf{x}_k)\}_{k\neq i}) - \tilde{\mathbf{I}}_i(\mathbf{x}_i)||_2^2. \tag{2}$$

Compared with Equ. 1, Equ. 2 focuses on leveraging multi-view projections in LF images, where view consistency provides a more reliable supervisory signal. To better explore the blind-view formulation, we apply masking across entire views in the angular domain, without introducing any spatial-domain masking. The blind-view network θ is then trained by minimizing the following loss:

$$\arg\min_{\theta} ||f_\theta(\mathcal{M} \odot \mathcal{L}) - (1 - \mathcal{M}) \odot \mathcal{L}||_2^2, \tag{3}$$

where $\mathcal{M} \in \mathbb{R}^{U\times V\times H\times W}$ and $\odot$ is hadamard product. We set $\mathcal{M}(\mathbf{a}_i, :) = 0$ to indicate that the view $\mathbf{I}_i$ is masked. We present a concrete example in Fig. 1 to highlight the key differences between the monocular blind-spot network (BSN) and the blind-view network (BVN). When the pixel P represents a discontinuous signal, the BSN struggles to recover the original signal using its spatial context. In contrast, by leveraging the corresponding pixels of the target point in other views, which follow a distribution centered around the clean signal value, the BVN provides a statistically reliable basis for estimation.

3.3. Blind View Mask Analysis

Similar to BSN, selecting an appropriate masking scheme is a crucial challenge for BVN. For BSN, a simple approach is to use a random selection strategy [15, 24], as illustrated in Fig. 2(a). Since BSN relies on the statistical properties

inherent within a single image—properties that remain consistent regardless of the blind spot's location—some methods [32] design fixed, simple patterns, as shown in Fig. 2(b). This statistical foundation enables these methods to tailor their masking strategy to address specific challenges. However, BVN for LF images fundamentally depends on the geometric relationships between views, and these relationships vary significantly across view pairs. Several additional constraints must be considered to ensure optimal performance. In this section, we elaborate on the key principles of designing the view masking strategy.

(1) Fixed mask pattern: The projection of a 3D point across different views is inherently governed by disparity constraints. The projections of a 3D point on view $\mathbf{I}_i(\mathbf{s}_i)$ and view $\mathbf{I}_j(\mathbf{s}_j)$ are supposed to satisfy the constraint:

$$\mathbf{D}_{j\to i}(\mathbf{x}_i) = \mathbf{x}_j - \mathbf{x}_i = (\mathbf{a}_j - \mathbf{a}_i) \cdot \frac{B \cdot f}{\mathbf{Z}(\mathbf{x}_i)} \tag{4}$$

where B represents the baseline length between adjacent views, f represents the focal length of the camera, $\mathbf{Z}(\mathbf{x}_i)$ represents the depth of $\mathbf{x}_i$. Since the scene geometry ($\mathbf{Z}$) is fixed, the disparity between views is solely determined by their relative angular positions. Therefore, a fixed angular masking pattern is essential to ensure geometric consistency across views and facilitate effective learning of view relationships.

(2) Uniform blind-view coverage: Monocular blind-spot denoising requires masked pixels to be uniformly distributed in the spatial domain to ensure sufficient contextual information is available for each prediction [32]. Similarly, when applying view masking to LF images, the blind views should be evenly distributed across the angular domain so that each masked view has nearby visible views for reference. This strategy maximizes the use of complementary angular information and helps to mitigate information loss caused by disparity and occlusion. As illustrated in Fig. 2(c), the uneven distribution of masked views (e.g. in patterns (2) and (3)) forces some predictions to rely on distant views with large disparity, significantly increasing the difficulty of accurate reconstruction.

(3) Trade-off in masking degree: In monocular blind-spot denoising, increasing the ratio of masked pixels inevitably leads to greater information loss [32], and a similar issue arises in blind-view denoising for LF images. However, as discussed above, each unique view mask requires a separately trained network. This constraint necessitates selecting an optimal number of simultaneously masked views to balance two competing objectives: (1) the accuracy of learned geometric correspondences, and (2) the computational efficiency of training multiple networks to achieve full LF coverage. As shown by pattern (4) in Fig. 2(c), although only one view is masked, denoising the entire LF would require training $U \cdot V$ independent networks, resulting in significant computational overhead and parameter redundancy. Conversely, using fewer visible views—as in pattern (1)—reduces the number of required networks (e.g., to two), but at the cost of halving the visible context available for reconstruction. Furthermore, assigning different networks to denoise different views within the same LF can introduce inconsistencies in photometric appearance, which may degrade performance in downstream tasks such as depth estimation.

4. Method

In this section, we first propose the geometric invariance mask that is designed tailored to the LF structure. Then the denoising framework and training strategy are analyzed.

4.1. Geometric Invariance Mask

As discussed in Sec. 3.3, it is essential to design a fixed masking scheme that both (i) allows the model to effectively exploit the geometric consistency of LF images, and (ii) avoids the need for training separate networks to denoise different views within one LF.

Geometric Invariance in LFs: As demonstrated in [28], the estimated disparity map remains invariant under arbitrary rotations of the LF data. This requires applying consistent spatial rotations to each individual view while simultaneously reordering the views in the angular domain to maintain their geometric relationships. From Equ. 4, we understand that disparity depends on both the spatial coordinates and the angular coordinates of the pixels. If we only rotate each view in the spatial domain while keeping the angular coordinates unchanged, the original geometric relationships between views are disrupted. In other words, the relationship between the views changes, requiring a new network to learn this new relationship. Thus, the geometric relationships between views remain consistent only when the entire LF is rotated concurrently in both the spatial and angular domains. As shown in Fig. 2 (d), let $\mathbf{I}_i$ denote the view located at angular position $\mathbf{a}_i$, and the point $\mathbf{p}_i$ denote the projection of a 3D point P in this view. We assume that $\mathbf{p}_{25} = \mathbf{I}_{25}(x, y)$ exhibits disparity d to its corresponding point in adjacent views, i.e. $\mathbf{p}_{24} = \mathbf{I}_{24}(x-d, y)$, $\mathbf{p}_{32} = \mathbf{I}_{32}(x, y-d)$. If only the two views in the red boxes are considered, the expectation of disparity the network ϕ learned is: $\mathbb{E}[f_\phi(\mathbf{I}_{24}, \mathbf{I}_{25})]_\mathbf{p} = (d, 0)$. Define the spatial-angular rotation operator as $\mathsf{R}^r_{\mathbf{ax}}$ and spatial rotation as $\mathsf{R}^r_{\mathbf{x}}$, where r is the rotation angle. After spatial-angular rotation, the input of ϕ has changed and the expectation of disparity at point $\mathbf{p}$ is: $\mathbb{E}[f_\phi[\mathsf{R}^{90}_{\mathbf{ax}}(\mathcal{L})]] = \mathbb{E}[f_\phi[\mathsf{R}^{90}_{\mathbf{x}}(\mathbf{I}_{32}), \mathsf{R}^{90}_{\mathbf{x}}(\mathbf{I}_{25})]]_\mathbf{p} = (y, W-x) - (y-d, W-x) = (d, 0)$. Therefore, by leveraging this property, a single network can process different views to obtain the same geometric representation of the LF.

GIM Design: Based on the above analysis, we design a GIM that adheres to the three principles outlined in Sec. 3.3. This design not only satisfies the core requirements but also

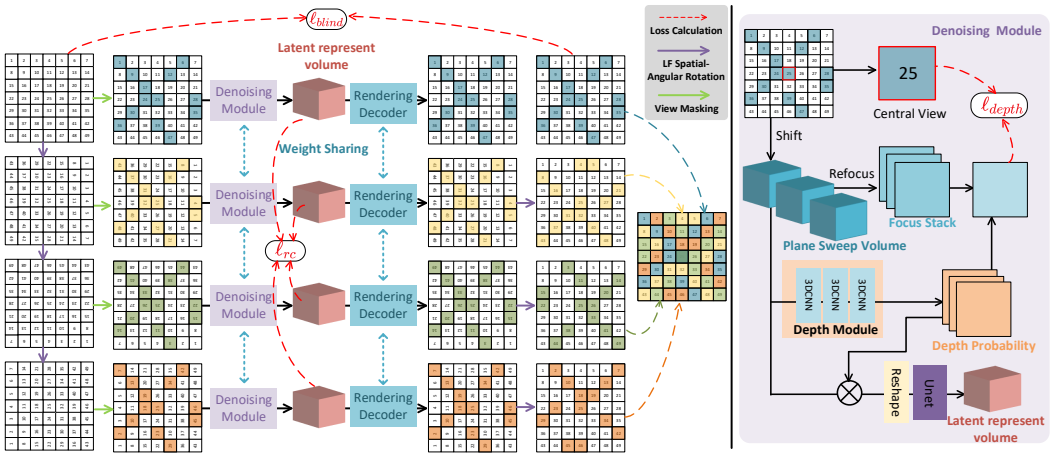


Figure 3. The overview of our proposed LF-BVN framework (left) and the detailed denoising module (right). By rotating the LF and applying the same view mask (GIM), four branches process different inputs but share network parameters to jointly generate a complete denoised LF. The denoising module reconstructs clean latent representation volumes from noisy data and the ℓ_{rc} enforces the consistency between all denoised views. Then the decoder renders the denoised views, respectively. In the denoising module, we first construct PSV and use a UNet to obtain the latent representation volume. A refocusing-based depth estimation module applies the ℓ_{depth} constraint and provides additional depth guidance.

capitalizes on the rotational invariance property of LF, enabling a single network to achieve efficient denoising across all views. The design rationale and implementation details are provided in the supplementary material. The designed GIM is visualized in Fig. 2(d). Specifically, we mask the blue views at first. After rotation, the LF changes while the view mask remains unchanged, so the set of masked views shifts to those marked in red. Despite variations in the input views, the relationship between visible views remains unchanged. Thus a weight sharing network can effectively process all transformed inputs.

4.2. Blind-View Denoising Framework

The whole framework of LF-BVN f_θ employs a four-branch architecture with shared parameters as in Fig. 3. A fixed GIM is applied across all branches, resulting in different sets of visible views due to rotation $(0°, 90°, 180°, 270°)$, while preserving consistent geometric relationships. In each branch, the masked LF is fed into a shared Denoising Module f_{θ_1}, which reconstructs a latent representation volume that encodes the geometry of the denoised scene. The rendering decoder f_{θ_2} (composed of 3 residual blocks) is then applied to recover context of the masked views. Finally, the corresponding inverse rotation is applied to each branch's output and fuses the complete denoised LF.

Since different views are denoised separately by the trained network, it is important to keep consistency across all views. Therefore, we propose using a view-invariant LF representation—termed the latent representation volume [29]—to model the LF, and introduce a consistency constraint on the latent volumes derived from different branches. Furthermore, to fully leverage the depth information inherent in LF, we design a novel noise-robust depth estimation module that provides additional depth guidance for constructing the latent representation volume.

Depth Probability from Refocus: Based on the blind-spot denoising analysis in Sec. 3.3, we recognize that the key to blind-view denoising lies in identifying the corresponding projection points of a 3D scene point across different views, where their geometric relationship follows the disparity formulation in Equ. 4. So accurate depth guidance is pivotal for enhancing network performance. However, existing self-supervised depth estimation methods [14, 36, 41] commonly rely on photometric consistency loss, which becomes unreliable under noisy conditions. To overcome this limitation, we introduce a depth estimation module that coarsely predicts the probability distribution of each pixel over a predefined set of depth planes, along with a novel depth supervision loss. Specifically, we first construct the plane sweep volume ($PSV \in \mathbb{R}^{D\times U\times V\times H\times W\times C}$) by shifting the masked LF to the central view according to a predefined set of depth planes $\{d_1, d_2, ..., d_D\}$. Then we employ LF refocusing to generate a focus stack:

$$FS = \frac{1}{|M|}\sum_{u=1}^{U}\sum_{v=1}^{V} PSV_{uv}, \tag{5}$$

where $FS \in \mathbb{R}^{D\times H\times W\times C}$ is the focal stack and $|M|$ represents the number of visible views. As the central view is masked, the refocused image lacks any information from the original central view. This allows us to employ the noisy center view as a supervision signal without the risk of the network learning an identity mapping. Therefore, we design a depth estimation module σ that predicts pixel-wise depth

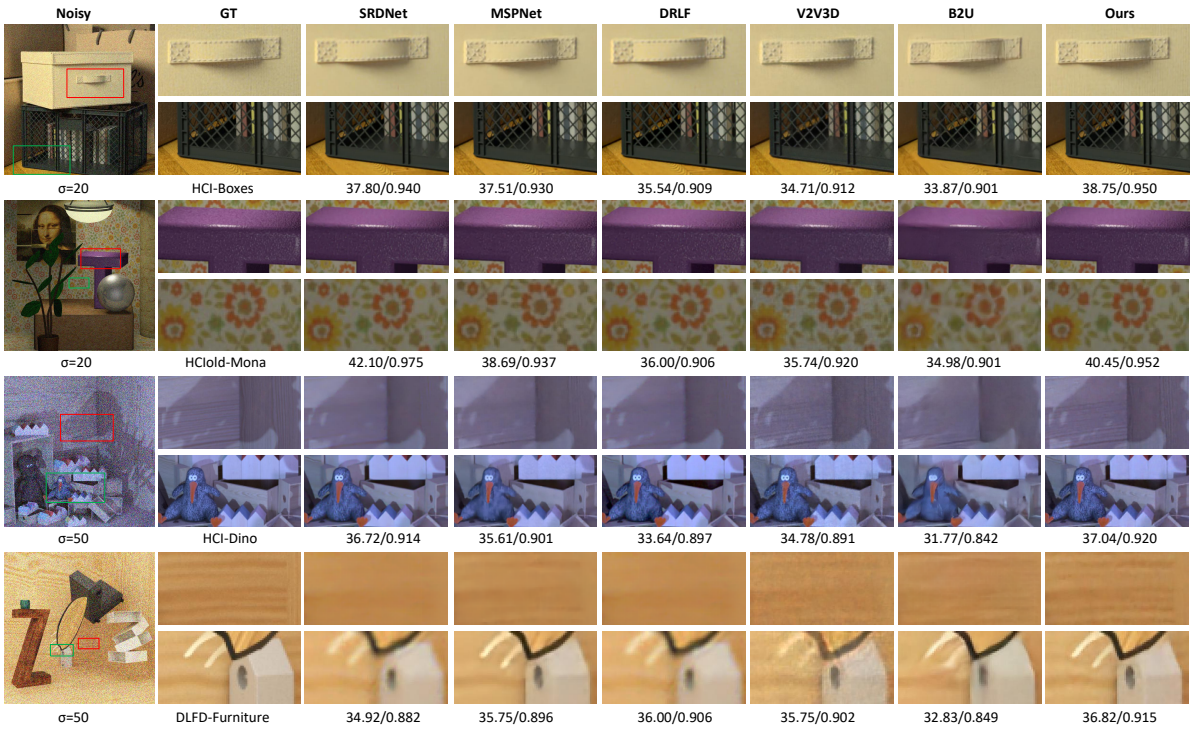


Figure 4. Qualitative denoising results on synthetic datasets in sRGB space.

probabilities $DP \in \mathbb{R}^{D\times H\times W}$ and constrains it using the following loss:

$$\ell_{depth} = ||\sum_{d}^{D}[f_\sigma(PSV) \odot FS]_d - \tilde{\mathbf{I}}_c||_1, \quad (6)$$

where $\tilde{\mathbf{I}}_c$ is the original central view. Then a weighted PSV can be obtained by applying DP to the PSV. We visualize a depth map generated from the probability volume in the supplementary material.

Latent Representation Volume: As analyzed before, all four branches of our framework are expected to learn a consistent latent representation of the same scene. To explicitly enforce coherence across all denoised views, we introduce a reconstruction consistency loss ℓ_{rc}:

$$\ell_{rc} = \sum_{r}||\mathsf{R}_{\mathbf{x}}^{-r}(f_{\theta_1}(\mathsf{R}_{\mathbf{ax}}^{r}(\mathcal{L}) \odot \mathcal{M})) - f_{\theta_1}(\mathcal{L} \odot \mathcal{M})||_1, \quad (7)$$

where $r \in \{90, 180, 270\}$. When the rotated LF is processed with our GIM, the set of visible views changes, while the underlying 3D scene structure it represents remains unchanged. Therefore, the latent representation volume (a tensor with the shape of $4C \times H \times W$) produced by θ_1, after applying the inverse rotation, should be consistent with the representation obtained from the original orientation.

Training Loss: Our network is trained by minimizing a hybrid loss function defined as follows:

$$\ell_{total} = \ell_{blind} + \alpha\ell_{rc} + \beta\ell_{depth}, \quad (8)$$

where ℓ_{blind} measures the difference between the predicted masked views and their original values as in Equ. 3.

5. Experiments

5.1. Implementation Details

We use PyTorch deep learning framework and Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) to optimize the network. The batch size is set to 20. The learning rate is fixed at $1e-4$. The model is trained on an NVIDIA A5000 GPU and stopped after $1e5$ iterations. During training, we select the central 7×7 views and randomly crop them into 48×48 patches as input. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are employed as evaluation metrics. We set the weighting coefficient $\alpha = 0.3$ and $\beta = 0.1$.

5.2. Comparison with SOTA methods

To validate the efficacy of our method, we evaluate against three benchmark categories: (1) monocular blind-spot denoising methods, B2U [32], ZSBlind [24], (2) supervised LF methods, APA [5], DRLF [11], SRDNet [4], MSPNet [30], and (3) self-supervised LF methods, V2V3D [42]. For fair comparison, all methods are trained under same experimental settings.

Synthetic LF Images: We employ 16 LF scenes from the HCI dataset [33] for training and evaluate on the other 8 scenes. To evaluate the generalization capability of these methods across different datasets, we conduct zero-shot testing on the HCIold dataset [33] (5 LF scenes) and DLFD dataset [27] (30 LF scenes).

Table 1. Quantitative denoising results on synthetic datasets in sRGB space. The highest PSNR(dB)/SSIM is highlight in **bold**.

	Method	HCI	HCIold	DLFD
$\sigma = 20$	APA	33.55/0.885	33.09/0.862	31.47/0.830
	DRLF	37.71/0.927	36.44/0.927	35.19/0.899
	MSPNet	37.87/0.928	36.58/0.931	35.56/0.910
	SRDNet	**38.17**/0.925	36.78/0.937	35.44/0.901
	B2U	33.85/0.887	33.61/0.872	32.98/0.856
	ZSBlind	30.93/0.802	30.39/0.788	29.56/0.778
	V2V3D	36.21/0.909	35.97/0.896	33.89/0.891
	Ours	37.86/**0.960**	**37.72/0.953**	**36.04/0.937**
$\sigma = 50$	APA	31.06/0.833	29.97/0.812	27.31/0.772
	DRLF	33.73/0.877	32.69/0.855	30.54/0.844
	MSPNet	33.27/0.881	33.15/0.853	31.94/0.866
	SRDNet	**34.73**/0.901	33.96/0.898	32.27/0.884
	B2U	28.76/0.824	28.37/0.833	27.88/0.801
	ZSBlind	27.83/0.745	27.39/0.707	27.32/0.715
	V2V3D	33.37/0.901	34.21/0.897	33.04/0.883
	Ours	34.51/0.911	**34.63/0.907**	**33.19/0.883**
$\sigma \in [5, 50]$	APA	33.15/0.880	33.07/0.870	31.31/0.817
	DRLF	35.51/0.925	35.57/0.921	34.47/0.902
	MSPNet	34.73/0.908	34.52/0.910	33.75/ 0.887
	SRDNet	**35.97**/0.933	35.59/0.927	33.67/0.889
	B2U	32.32/0.869	32.37/0.872	32.11/0.849
	ZSBlind	28.59/0.756	28.29/0.758	27.92/0.743
	V2V3D	34.51/0.911	34.63/0.907	34.11/0.900
	Ours	35.91/**0.939**	**35.93/0.937**	**34.45/0.911**

(1) Gaussian noises: To ensure comprehensive evaluation, we train three separate models for each method using datasets corrupted with different noise levels: (1) Gaussian noise with $\sigma = 20$, (2) Gaussian noise with $\sigma = 50$, (3) Gaussian noise with $\sigma \in [5, 50]$. As shown by the quantitative comparison results in Table 1, whether the noise is fixed or variable, our method outperforms other self-supervised methods and three supervised methods. Although SRDNet achieves a higher SSIM on the HCI dataset, its supervised strategy leads to a strong dependency on the specific distribution of the training data. Since the HCIold and DLFD datasets contain more weak-texture regions, the method's performance degrades significantly on them. Fig. 4 presents qualitative comparisons of these methods. Compared to other methods including supervised methods, our method can recover subtle textures, such as those on desktops and walls, even under high-noise conditions. This is because we construct the representation of LF and constrain the consistency of denoised views. Although V2V3D also constructs the representation volume, it relies on a simple average to fuse multiple volumes obtained from different models, which leads to visible artifacts in textured areas.

(2) Unseen Noise Types: To assess the generalization capability to noise types they have never seen, models trained on dataset with Gaussian noise ($\sigma = 20$) were evaluated on datasets corrupted by: (4) Poisson noise with $\lambda = 30$, (5) Uniform noise with intensity values uniformly distributed in $[-50, 50]$. Table 2 presents the quantitative results. It is evident that supervised methods generally fail when handling noise types not present in the training data, whereas our approach maintains robust performance under such challenging conditions. The superiority originates from our unsupervised learning architecture's ability to capture the essential geometry of LF through its training strategy. By avoiding supervised learning's tendency to overfit specific noise distributions[38], our method learns robust scene representations that facilitate accurate image restoration.

Table 2. Denoising performance on out-of-distribution noise.

	Method	HCI	HCIold	DLFD
Poisson $\lambda = 30$	APA	28.76/0.593	28.80/0.627	26.48/0.591
	DRLF	31.29/0.789	31.22/0.786	30.19/0.758
	MSPNet	30.84/0.775	30.23/0.841	29.66/0.744
	SRDNet	30.64/0.763	30.20/0.755	29.26/0.726
	B2U	33.02/0.861	33.12/0.868	32.76/0.859
	V2V3D	34.17/0.889	34.26/0.887	33.85/0.867
	Ours	**36.09/0.932**	**36.43/0.944**	**34.51/0.917**
Uniform $[-50, 50]$	APA	29.83/ 0.706	29.55/0.696	26.50/0.623
	DRLF	31.32/0.695	31.20/0.716	30.36/0.719
	MSPNet	33.74/0.873	33.21/ 0.878	31.31/0.867
	SRDNet	29.96/0.670	30.07/0.669	27.83/0.659
	B2U	32.55/0.879	32.03/0.882	32.39/0.876
	V2V3D	33.33/0.885	33.22/0.894	32.78/ 0.898
	Ours	**36.46/0.943**	**36.16/0.928**	**35.13/0.917**

Table 3. MAE and MSE metrics for the depth maps using the denoised results. Both training and inference phases utilized Gaussian noise with $\sigma = 20$.

Method	HCI	HCIold	DLFD
DRLF	0.196/0.203	0.216/0.225	0.302/0.320
MSPNet	0.162/0.109	0.177/0.115	0.265/0.231
SRDNet	0.158/0.097	0.162/0.094	0.234/0.217
V2V3D	0.187/0.124	0.189/0.133	0.208/0.176
Ours	**0.122/0.056**	**0.135/0.061**	**0.172/0.113**

(3) Depth Estimation from Denoised LFs: To evaluate the view consistency of the denoised results, we use the OAVC [12] to estimate the depth of the denoised LFs. The results shown in Table 3 demonstrate that the depth maps estimated from our denoised method demonstrate higher accuracy compared to those from competing methods. This indicates that our denoising approach better preserves the geometric and photometric consistency across the LF views. Visual results are provided in the supplementary materials.

Real-world LF Images: We show two LF images captured by a Lytro Illum camera and their denoised results in Fig. 5. All models are trained on HCI dataset with Gaussian noise ($\sigma = 20$). Additional results are provided in the supplementary material. As evidenced by the results, SRDNet and DRLF fail to effectively handle defective pixels. While MSP demonstrates some capability in addressing these bad pixels, it incurs severe texture over-smoothing. V2V3D exhibits noticeable artifacts in its output. In contrast, our approach successfully processes realistic noise while recovering high-fidelity textures. We show more results in the supplementary material.

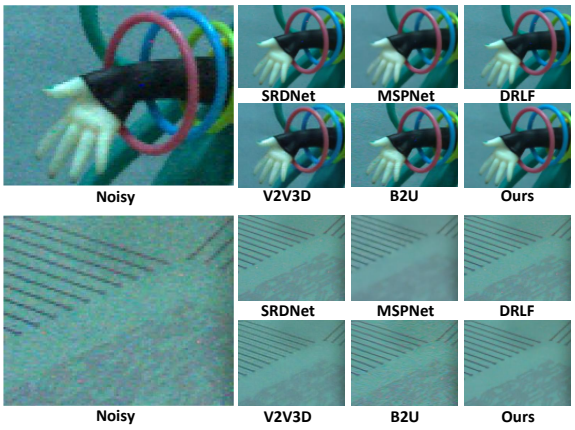


Figure 5. Qualitative denoising results on real-world LF images.

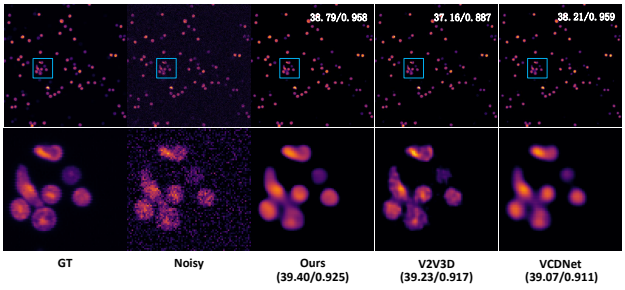


Figure 6. Qualitative and quantitative denoising results on synthetic LF microscopy dataset. The data in parentheses are the average metrics on the test set.

Microscopy LF Images: Following V2V3D, we trained our method on their provided synthetic LFM dataset, followed by the addition of Gaussian noise with $\sigma = 10$. Since the objective of LFM denoising is to obtain denoised explicit representation volumes, we render the LFM images using the PSFs instead of the rendering decoder. Qualitative and quantitative results are presented in Fig. 6. This demonstrates that our approach can not only denoise but also keep more details.

5.3. Efficiency Comparison

We compare the parameter and computational time in Table 4. Compared to V2V, our method reduces the number of required models and optimizes the rendering scheme, significantly decreasing both parameter counts and inference time. Although supervised methods such as MSPNet and SRDNet achieve a lower parameter count, their inefficient architectural choices—such as processing views sequentially (SRDNet) or requiring multiple iterations (MSPNet)-significantly increase inference latency.

5.4. Ablation Study

We conduct ablation studies on the HCI dataset with added Gaussian noise ($\sigma = 20$).

Masking Ratio: As analyzed in Sec. 3.3, a lower mask ratio yields better results but requires more branches. With a $90°$ rotation angle, the optimal minimum masking ratio is 25%, requiring 4 branches. As shown in Table 5, we also evaluate 50% and 75% masking ratios. Increasing the mask ratio while keeping the branch count fixed (i.e., averaging multiple masked outputs) degrades performance due to reduced visible context. Furthermore, using larger mask ratios with fewer branches reduces computational costs but further drops performance.

Table 4. Number of parameters and run time comparison.

Metric	B2U	MSPNet	SRDNet	V2V3D	Ours
Params (M)	1.10	1.22	0.62	53.86	12.30
Run Time (s)	1.46	2.27	3.12	1.05	0.55
FLOPs (T)	3.21	5.20	9.53	8.75	5.00

Table 5. Ablation study on Mask ratio and Branch number

Branch Num	4 with 90°			2 with 180°	
Mask Ratio	25%	50%	75%	50%	75%
PSNR ↑	**37.86**	36.91	36.10	35.97	35.19
SSIM ↑	**0.960**	0.942	0.926	0.925	0.910
FLOPs(T) ↓	5.0	5.0	5.0	2.5	2.5

Table 6. Ablation study about the proposed GIM, Latent Representation Volume, Depth Estimation Loss and Reconstruction Consistency Loss. The base model is trained using random view masks.

GIM	✗	✓	✓	✓	✓
LRV	✗	✗	✓	✓	✓
ℓ_{depth}	✗	✗	✗	✓	✓
ℓ_{lrc}	✗	✗	✗	✗	✓
PSNR	30.56	36.47	37.27	37.55	37.86
SSIM	0.812	0.915	0.930	0.954	0.960

Network Components: As in Table 6, we evaluate individual modules of LF-BVN. The base model trained with random view masks fails to learn cross-view relationships and cannot recover LF images. The proposed GIM successfully adapts the blind-view strategy to LFs. Synergistically integrating the latent representation volume (LRV) significantly improves the denoising performance. Finally, adding the supervision of the depth estimation loss and reconstruction consistency loss achieves the best overall performance.

6. Conclusion

We present the first self-supervised denoising framework applicable to both conventional LF images and LF microscopy. We propose a Geometric Invariance Mask to achieve efficient LF blind-view denoising. Through the refocusing-based depth estimation module and reconstruction consistency loss, the consistency between denoised views is further enhanced. Experimental results demonstrate superiority over state-of-the-art self-supervised methods and exceptional generalization across modalities.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (No.62372032)

References

[1] Martin Alain and Aljosa Smolic. Light field denoising by sparse 5d transform domain collaborative filtering. In *IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2017. 1, 2

[2] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3155–3164, 2019. 2

[3] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533, 2019. 2

[4] Song Chang, Youfang Lin, Wenqi Wang, Da An, and Shuo Zhang. Learning light field denoising with symmetrical refocusing strategy. *IEEE Transactions on Computational Imaging*, 10:1786–1798, 2024. 1, 2, 6

[5] Jie Chen, Junhui Hou, and Lap-Pui Chau. Light field denoising via anisotropic parallax analysis in a cnn framework. *IEEE Signal Processing Letters*, 25(9):1403–1407, 2018. 1, 6

[6] Jie Chen, Junhui Hou, and Lap-Pui Chau. Light field denoising via anisotropic parallax analysis in a cnn framework. *IEEE Signal Processing Letters*, 25(9):1403–1407, 2018. 2

[7] Jiaxin Chen, Shuo Zhang, and Youfang Lin. Attention-based multi-level fusion network for light field depth estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1009–1017, 2021. 1

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007. 2

[9] Donald G Dansereau, Daniel L Bongiorno, Oscar Pizarro, and Stefan B Williams. Light field image denoising using a linear 4d frequency-hyperfan all-in-focus filter. In *Computational Imaging XI*, pages 176–189. SPIE, 2013. 2

[10] Mantang Guo, Jing Jin, Hui Liu, and Junhui Hou. Learning dynamic interpolation for extremely sparse light fields with wide baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2450–2459, 2021. 1

[11] Mantang Guo, Junhui Hou, Jing Jin, Jie Chen, and Lap-Pui Chau. Deep spatial-angular regularization for light field imaging, denoising, and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6094–6110, 2022. 1, 6

[12] Kang Han, Wei Xiang, Eric Wang, and Tao Huang. A novel occlusion-aware vote cost for light field depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8022–8035, 2021. 7

[13] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14781–14790, 2021. 2

[14] Jing Jin and Junhui Hou. Occlusion-aware unsupervised learning of depth from 4-d light fields. *IEEE Transactions on Image Processing*, 31:2216–2228, 2022. 3, 5

[15] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2129–2137, 2019. 2, 3

[16] Alexander Krull, Tomáš Vičar, Mangal Prakash, Manan Lalit, and Florian Jug. Probabilistic noise2void: Unsupervised content-aware denoising. *Frontiers in Computer Science*, 2:5, 2020. 2

[17] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[18] J Lehtinen. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018. 1, 2

[19] Junyi Li, Zhilu Zhang, and Wangmeng Zuo. Rethinking transformer-based blind-spot network for self-supervised image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 1, 2

[20] Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *IEEE Transactions on Image Processing*, 21(9): 3952–3966, 2012. 2

[21] Ren Ng. Lytro redefines photography with light field cameras. *Accessed: Oct*, 22, 2018. 1

[22] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2043–2052, 2021. 2

[23] Sanduni U Premaratne, Namalka Liyanage, Chamira US Edussooriya, and Chamith Wijenayake. Real-time light field denoising using a novel linear 4-d hyperfan filter. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 67 (8):2693–2706, 2020. 2

[24] Yuhui Quan, Tianxiang Zheng, Zhiyuan Ma, and Hui Ji. Zero-shot blind-spot image denoising via implicit neural sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7502–7512, 2025. 1, 2, 3, 6

[25] Raytrix. Raytrix: Light filed technology. 2020. 1

[26] Alireza Sepas-Moghaddam, Paulo Lobato Correia, and Fernando Pereira. Light field denoising: exploiting the redundancy of an epipolar sequence representation. In *The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pages 1–4. IEEE, 2016. 1, 2

[27] Jinglei Shi, Xiaoran Jiang, and Christine Guillemot. A framework for learning depth from a flexible subset of dense and sparse light field views. *IEEE Transactions on Image Processing*, 28(12):5867–5880, 2019. 6

[28] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4748–4757, 2018. 4

[29] Thomas Tanay and Matteo Maggioni. Global latent neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19723–19733, 2024. 5

[30] Xianglang Wang, Youfang Lin, and Shuo Zhang. Multi-stream progressive restoration for low-light light field enhancement and denoising. *IEEE Transactions on Computational Imaging*, 9:70–82, 2023. 1, 2, 6

[31] Zhaoqiang Wang, Lanxin Zhu, Hao Zhang, Guo Li, Chengqiang Yi, Yi Li, Yicong Yang, Yichen Ding, Mei Zhen, Shangbang Gao, Tzung Hsiai, and Peng Fei. Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning. *Nature Methods*, 18, 2021. 1

[32] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2027–2036, 2022. 1, 2, 4, 6

[33] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–48, 2012. 6

[34] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. In *ACM siggraph 2005 papers*, pages 765–776. 2005. 1

[35] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368. Springer, 2020. 2

[36] Wenbin Yan, Xiaogang Zhang, and Hua Chen. Occlusion-aware unsupervised light field depth estimation based on multi-scale gans. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6318–6333, 2024. 5

[37] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 492–511, 2020. 2

[38] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2, 7

[39] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016. 1

[40] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11046–11055, 2019. 1

[41] Shansi Zhang, Nan Meng, and Edmund Y Lam. Unsupervised light field depth estimation via multi-view feature matching with occlusion prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2261–2273, 2023. 5

[42] Jiayin Zhao, Zhenqi Fu, Tao Yu, and Hui Qiao. V2v3d: View-to-view denoised 3d reconstruction for light field microscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26451–26461, 2025. 1, 2, 6