GaussianMatch: Adaptive Learning Continuous Surfaces for Light Field Depth Estimation

Zexin Sun[1,2], Rongshan Chen[1,2], Yu Wang[1,2], Zhenglong Cui[2], Da Yang[2], Siyang Li[2], Xuefei Huang[2], Hao Sheng[1,2,3†]

[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R.China

[2]Data Science and Intelligent Computing Laboratory, Hangzhou International Innovation Institute, Beihang University, Hangzhou, Zhejiang 311115, P.R.China

[3]Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR 999078, P.R.China. (Emails: zexinsun, rongshan, wy4ward, zhenglong.cui, da.yang, lisiyang, xuefei.huang, shenghao@buaa.edu.cn).

Abstract

Light field (LF) depth estimation plays a vital role in computational imaging by reconstructing 3D structures from multiple viewpoints. However, images are merely discrete expressions of scenes due to the resolution constraints of cameras, leading to depth discontinuities and outliers-particularly in textureless or occluded regions, degrading reconstruction coherence. To address the challenges mentioned above, we propose GaussianMatch, a probabilistic depth estimation framework that models per-pixel depth as a learnable Gaussian distribution in continuous space. This scheme effectively alleviates the discretization problem of LF images by adaptively reconstructing continuous surfaces, while enabling uncertainty-aware optimization.Furthermore, the framework naturally fuses information among adjacent pixels and adapts each Gaussian's variance according to scene complexity, achieving robustness in both texture-rich and ambiguous regions. We further design GaussianNet, which regresses per-pixel Gaussian parameters and generates the final depth map via Gaussian accumulation. Extensive experiments on multiple LF benchmarks demonstrate that GaussianNet achieves state-of-the-art accuracy, with significant improvements in handling depth discontinuities and occlusions.

CCS Concepts

• ***Computing methodologies*** → ***Matching;*** *Epipolar geometry; Computational photography; Image representations;*

1. Introduction

Light field (LF) technology captures both the directional and intensity information of light rays from multiple viewpoints, enabling richer geometric understanding compared to conventional 2D images [LH96]. This multi-view nature inherently encodes parallax cues [CSY*23b], which form the basis for depth estimation. Represented as sub-aperture images (SAI)(Fig. 1.(a)), macro-pixel images (MacPI), or epipolar plane images (EPI), LF data provides a robust foundation for reconstructing 3D scene structures [SCY*22, CSZ*24]. As a core component in computational imaging systems, LF depth estimation enables advanced applications such as digital refocusing, super-resolution imaging, image reconstruction, and virtual reality rendering [ZLL*24, YZC24, CYT23, WZF*25].

With the rapid advancement of deep neural networks, various deep learning-based techniques have been proposed [CZL21, HP16, HHX*21, SJY*18, TLOC20], significantly improving the accuracy and performance of LF depth estimation. Among these, cost volume-based methods are currently the mainstream approach. These methods typically follow a four-step pipeline: feature extraction, cost volume construction, cost volume aggregation, and depth regression, which are based on the assumption that the intensity of corresponding pixels across different viewpoints remains consistent. By aggregating information from surrounding viewpoints toward the central view, such methods generate accurate depth maps.

Although cost volume-based methods have demonstrated strong performance in depth estimation, they still face a fundamental limitation: due to the resolution constraints of cameras, LF images can only capture discrete samples of scene rays. However, real-world 3D surfaces are continuous and, in theory, emit an infinite number of rays. This discretization can lead to deviations from the true depth distribution, resulting in local depth discontinuities and numerous outliers. Therefore, investigating approaches to reconstruct continuous surface representations from discretely sampled LF images is of significant importance for achieving more accurate and realistic scene understanding.

To address the outlier issue caused by image discretization, PlaneNet [CSY*24] introduces a plane-prior-based depth predic-

† Corresponding author.

DOI: 10.2312/pg.20251283

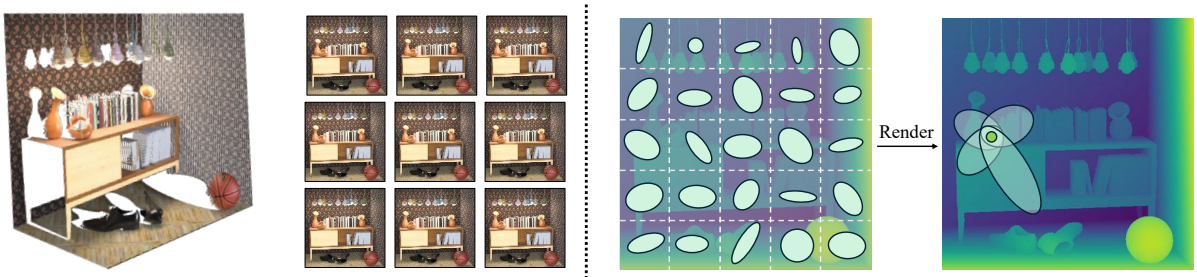


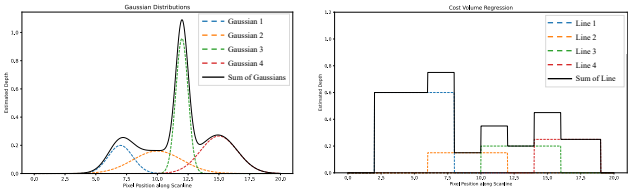
Figure 1: *(a) An illustration of LF visualization, which includes the true scene and the sub-aperture images (SAIs). (b) An illustration of GaussianMatch, which models per-pixel depth as a Gaussian distribution.*

tion approach. This strategy effectively suppresses isolated errors through geometric priors but relies on manually designed plane segmentation strategies, which struggle to adapt to diverse geometric structures in complex scenes. Another class of methods inspired by PatchMatch [BRR11] generates pixel-wise depth hypotheses through random initialization. While this avoids manual prior design, the stochastic nature leads to unstable convergence and high computational cost. Moreover, in textureless areas and near occlusions, the reliability of disparity metrics declines sharply, resulting in residual noise.

To systematically overcome these challenges, we propose GaussianMatch, a probabilistic depth estimation framework that operates in continuous space. Unlike traditional approaches that rely on random hypothesis generation, GaussianMatch models each pixel's depth as a Gaussian distribution, as shown in Fig. 1.(b), characterized by its mean and variance. Leveraging this representation, GaussianMatch can adaptively reconstruct continuous surfaces, mitigating discretization artifacts and enabling uncertainty-aware optimization in both well-structured and ill-posed regions. Furthermore, because each Gaussian spreads over neighboring pixels in the image plane, the model naturally fuses information among adjacent pixels and across multiple subviews, allowing ambiguous or occluded regions to benefit from a holistic fusion of probabilistic evidence. Building on GaussianMatch, we design GaussianNet for LF depth estimation. Experimental results demonstrate that our method achieves state-of-the-art (SOTA) performance, particularly in occluded and repetitive regions.

Our contributions can be summarized as follows:

- **GaussianMatch:** A continuous, probabilistic depth framework that models per-pixel depth as Gaussian distributions, effectively alleviating the discretization problem of LF images by adaptively learning continuous surfaces in continuous depth space.
- **GaussianNet:** A network architecture that embeds GaussianMatch in LF depth estimation, effectively projecting features into probabilistic depth space, enabling mutual information exchange among pixels.
- **Extensive evaluation:** Demonstration of SOTA performance on multiple LF datasets, with better results in occluded and repetitive regions.



(a) Gaussian Distribution (b) Cost Volume Regression

Figure 2: *An illustration of Gaussian Distribution and Cost Volume Regression. Obviously, Gaussian distribution can be adaptively used to express continuous surfaces.*

2. Related Work

2.1. Deep Learning-Based Methods in LF Depth Estimation

Recent advancements in deep learning have significantly improved LF depth estimation performance. These methods can be divided into three categories: epipolar-plane image (EPI)-based, cost volume-based and focal-stack-based [CSY*23b, WSC*24b].

EPI-based methods leverage geometric structures to enhance the accuracy of depth estimation. Shin et al. [SJY*18] introduced EPINet, a multi-stream convolutional neural network designed for fast and effective depth estimation by fully utilizing EPI information. Later, Leistner et al. [LSM*19] proposed a wide-baseline framework that exploits EPI-shift properties, enabling robust depth estimation even under challenging conditions. Zhou et al. [ZSL*23] further advanced EPI-based approaches by introducing the concepts of stitched-EPI and half-stitched-EPI, which innovatively reorganize EPI structures to improve depth estimation performance.

Cost volume-based method aggregate all views information to achieve accurate depth estimation. Tsai et al. [TLOC20] developed LFattNet, which integrates a view selection module to prioritize regions with minimal occlusion and richer textures. To enhance efficiency, Wang et al. [WWW*22] proposed a generic mechanism to disentangle spatial-angularly coupled information in LF processing. Subsequently, OACC-Net [WWL*22] introduced a dynamic approach to modulate pixel information across different LF views. To further improve accuracy, Chao et al. [CWW*23] proposed a method for learning depth distributions and constructing a sub-

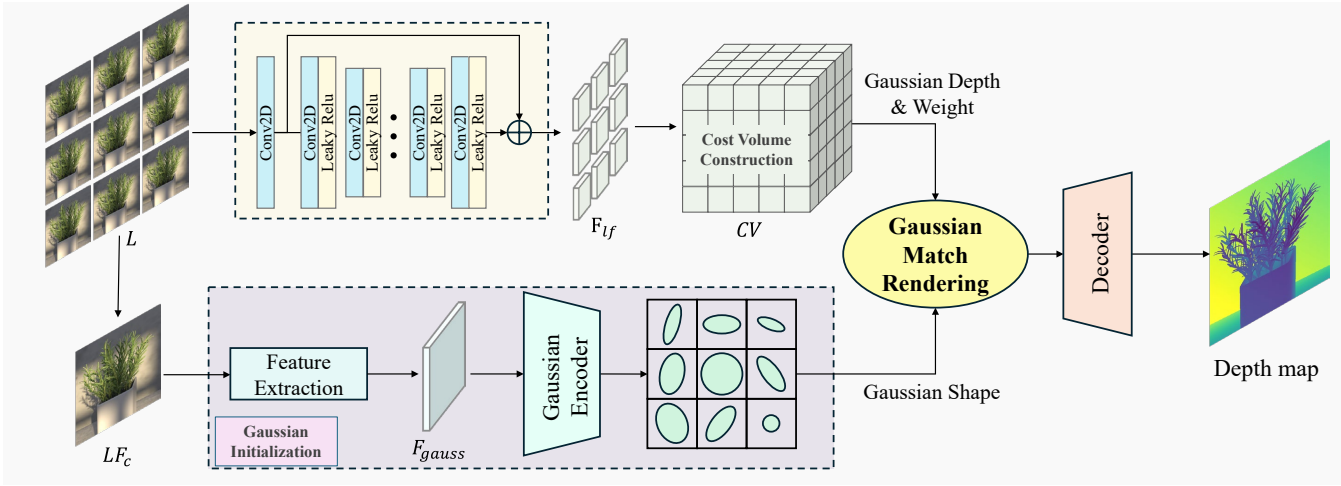


Figure 3: *An overview of our GaussianNet. Here, a 3 × 3 LF is used as an example. GaussianNet consists of two pathways: one pathway constructs a cost volume over multi-view features and interprets it as a Gaussian Depth representation, while the other extracts Gaussian Shape features. These two representations are fused via a Gaussian rendering module, producing a refined depth map for the central view.*

pixel cost volume. Additionally, Chen et al. [CSY*23a] developed an innovative post-processing procedure using Fourier transformation to correct erroneous pixels in the initial disparity map. Wang et al. [WSC*24a] enhanced flexibility by introducing an adaptive EPI-matching cost construction method. More recently, Chen et al. [CSY*25a] introduced a simple yet general learning framework for continuous-depth scene embedding, further advancing depth estimation methodologies.

Focal-stack-based methods exploit variations in focus across multiple refocused images to derive reliable depth cues. Zhou et al. [ZZY*19] introduced FocalStackNet, a dual-path convolutional architecture designed to capture both depth-semantic features and fine-grained structural information from focal stacks. Piao et al. [PZZJ21] presented an adaptive perception framework based on focal stacks and corresponding RGB images. To handle occlusion problems, Yang et al. [YCS*23] proposed the occlusion and noise-aware stereo framework based on refocusing.

2.2. Gaussian Splatting

Fast radiance-field optimization has shifted from implicit MLPs (NeRF++ [ZRSK20]) to explicit voxel grids (DVGO [SSC22]) and now to 3D Gaussian Splatting (3DGS) [KKLD23]. 3DGS [KKLD23] is a groundbreaking technology in computer graphics that enables explicit scene representation through millions of learnable 3D Gaussians. It represents the latest advancement in view synthesis, offering real-time, high-quality rendering. By implementing splatting-based rasterization-utilizing deep sorting and alpha blending of projected 2D Gaussians to compute pixel colors-3DGS bypasses the complex sampling strategies of ray marching, ensuring real-time performance. Due to its exceptional rendering capabilities, 3DGS has been widely adopted across various fields.

Building on this foundation, Liu et al. [LKG*23] proposed a method for continuous per-pixel depth modeling, enabling accurate depth prediction and distribution reasoning. Huang et al. [HYC*24] introduced a perspective-accurate 2D splatting process that employs ray-splat intersection and rasterization to improve rendering precision. Yu et al. [YCH*24] developed a 3D smoothing filter that constrains the size of 3D Gaussian primitives based on the maximal sampling frequency derived from input views. Additionally, Szymanowicz et al. [SRV24] leveraged 2D operators to map input images into 3D Gaussian representations, assigning one Gaussian per pixel. This continuous and differentiable representation facilitates the use of gradient-based optimization techniques [HXC*25].

3. Methodology

3.1. Preliminary: 3D Gaussian Splatting

Gaussian Splatting represents complex real-world scenes using a collection of 3D Gaussian primitives. Each primitive encodes essential attributes, such as spatial position, orientation, scale, opacity, and radiance, enabling smooth blending and continuous appearance transitions across overlapping regions.

A 3D Gaussian primitive is defined as:

$$\mathcal{G}(\mathbf{x};\mu,\Sigma) = \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^3$ denotes a point in 3D space, $\mu \in \mathbb{R}^3$ is the mean of the Gaussian, and $\Sigma \in \mathbb{R}^{3\times3}$ is the covariance matrix that characterizes the scale and orientation of the distribution.

During rendering, each 3D Gaussian defined in the camera coordinate system is projected onto the 2D image plane. Let i denote a pixel in the image. The contribution of a projected Gaussian $\mathcal{G}_i(\mathbf{x};\mu,\Sigma)$ to the pixel's intensity is modeled as:

$$\alpha_i = \sigma_i \cdot \mathcal{G}_i(\mathbf{x};\mu,\Sigma), \tag{2}$$

where σ_i is an opacity factor.

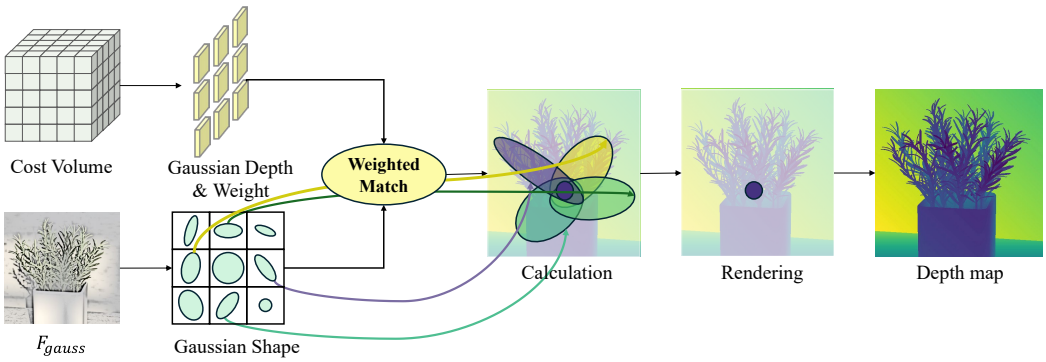


Figure 4: *Pipeline of GaussianMatch rendering. The final depth map is computed by aggregating Gaussian components using Gaussian Shape, Gaussian depths, Gaussian weights.*

To synthesize the final image, the contributions of multiple Gaussians are composited. An over-compositing approach is typically employed, in which the pixel color $C(p)$ is computed as:

$$C(p) = \sum_{i=1}^{N} c_i \cdot \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

where c_i denotes the color of the i-th Gaussian. The product term accounts for occlusion and transparency by attenuating the influence of each Gaussian based on the visibility of previously accumulated contributions. In 3DGS, multiple Gaussians are superimposed on the image plane and α-synthesized to obtain a color image; but in our LF depth estimation framework, we dispense with color modeling and instead employ Gaussian superposition solely to construct a depth-probability field.

3.2. Adaptive Learning of Continuous Scene Surfaces

Motivation and Overview In the real world, object surfaces are inherently continuous. However, due to the limited resolution of cameras, captured images represent discrete samples of the underlying scene. This "continuous-to-discrete" discrepancy becomes even more pronounced in LF data, where the inherent trade-off between angular and spatial resolution poses additional challenges, particularly for depth estimation, which requires reconstructing continuous geometric signals from sparse disparities and discrete pixel samples. Therefore, it is necessary to try to learn to reconstruct continuous surfaces.

Although existing LF depth estimation methods have made strides in mitigating discretization errors, they still struggle in complex regions due to their reliance on hand-crafted priors and limited use of global multi-view consistency. To address these issues and more effectively reconstruct continuous scene signals, we draw inspiration from 3DGS as applied to novel view synthesis. In 3DGS, each point in space is represented by an anisotropic Gaussian, and these Gaussians are composited with occlusion-aware blending to achieve a balance between surface smoothness and high-frequency detail. Motivated by this paradigm, we introduce a novel "GaussianMatch" tailored specifically for LF depth estimation. At its core, GaussianMatch constructs a pixel-depth Gaussian mixture field that adaptively learns a continuous surface representation, thereby bridging the gap between discrete LF samples and the true continuous geometry. As illustrated in Fig. 2, the Gaussian distribution enables adaptive learning of continuous surfaces, unlike prior approaches that rely on discretized disparity labels.

GaussianMatch Differing from 3DGS, which typically assumes known camera intrinsics and extrinsics and targets full-scene volumetric reconstruction, LF depth estimation generally does not rely on explicit camera calibration and only requires a depth map for the central view. Therefore, we adapt the core concept of anisotropic Gaussian to LF depth estimation. Specifically, we investigate how to adaptive model continuous geometric signals jointly in local pixel space and depth space, while only leveraging the relative angular shifts inherent in the LF sampling pattern. By doing so, GaussianMatch achieves a more faithful reconstruction of continuous surfaces from discrete LF observations.

Within the depth dimension, we represent the uncertainty of each pixel's depth estimate as a mixture of Gaussians. This formulation extends beyond the traditional "single-pixel, uncorrelated fitting" strategies, enabling a mutual information exchange among multiple pixels. As a result, depth estimates in regions of ambiguous texture or along occlusion boundaries benefit from a more holistic fusion of probabilistic evidence drawn from all relevant views.

In GaussianMatch, each center-view pixel at image coordinates (x, y) with a candidate depth d is represented by an anisotropic 3D Gaussian primitive in the (u, v, z) image-depth space, where

$$\mathbf{x} = (u, v, z)^\top \in \mathbb{R}^3, \quad \mu = (x, y, d)^\top \in \mathbb{R}^3, \quad \Sigma \in \mathbb{R}^{3\times 3}.$$

By predicting the above parameters, GaussianMatch can adaptively reconstruct continuous surfaces and deeply mitigate discretization artifacts.The Gaussian field function follows the Eq. 1 . When multiple Gaussian fields overlap, we compute each Gaussian's contribution to the rendered depth map in the central view. For a pixel (x, y) in the image plane, the contribution of the i-th Gaussian is

$$D_i(x, y) = w_i \cdot \mathcal{G}_i(\mathbf{x}; \mu, \Sigma), \tag{4}$$

where w_i is the weight of the i-th Gaussian field, determining its influence on the final depth. Because each Gaussian spreads over neighboring pixels in the image plane, the model naturally fuses information among adjacent pixels and across multiple subviews,

Table 1: *Mean square error (multiplied with 100) achieved by different methods on HCInew. The best results are in bold and the second best results are underlined, where Avg. means the average value on corresponding scenes*

Method	Backgammon	Dots	Pyramids	Stripes	Boxes	Cotton	Dino	Sideboard	Avg.MSE
EPINet [SJY*18]	3.629	1.635	0.008	0.950	6.240	0.191	0.167	0.827	1.706
EPI-Shift [LSM*19]	12.788	13.154	0.037	1.686	9.790	0.475	0.392	1.261	4.948
LFattNet [TLOC20]	3.648	1.425	**0.004**	0.892	3.996	0.209	0.093	0.531	1.350
FastLFnet [HHX*21]	3.986	3.407	0.018	0.984	4.395	0.322	0.189	0.747	1.756
OAVC [HXWH21]	3.835	16.582	0.040	1.316	6.988	0.598	0.267	1.047	3.834
OACC-Net [WWL*22]	3.938	1.418	**0.004**	0.845	2.892	0.162	**0.083**	0.542	1.236
DistgDisp [WWW*22]	4.712	**1.367**	**0.004**	0.917	3.325	0.184	0.099	0.713	1.415
FCVNet [WTZ23]	4.104	1.884	0.005	0.939	5.569	0.262	0.091	0.593	1.681
OALFGAN [YZC24]	5.381	1.740	0.013	2.388	8.010	0.756	0.736	3.393	2.802
MRAENet [LYZ*24]	4.830	1.462	0.005	0.933	3.193	0.164	0.091	0.635	1.414
PDE-Net-e [WL24]	4.028	3.413	0.017	0.998	3.909	0.270	0.128	0.557	1.665
GaussianNet(ours)	**3.617**	1.437	**0.004**	**0.803**	**2.754**	**0.155**	0.091	**0.516**	**1.172**

allowing ambiguous or occluded regions to benefit from a holistic fusion of probabilistic evidence.

GaussianMatch obtains the final depth map by summing all Gaussian contributions at each pixel. Given N splats, the depth at pixel (x,y) is

$$D(x,y) = \sum_{i=1}^{N} D_i(x,y) = \sum_{i=1}^{N} w_i \cdot \mathcal{G}_i(\mathbf{x};\mu,\Sigma). \tag{5}$$

This formulation accumulates the contributions from all Gaussians to that pixel, forming the final depth output. This design enables the network to adaptively modulate the contribution weights w_i at each pixel based on local cues such as occlusion and texture uncertainty.

3.3. Network Architecture

System Overview. Based on the proposed GaussianMatch, we develop GaussianNet. An overview of the architecture is illustrated in Fig. 3. The input to the network is a 4D LF image $\mathcal{L} \in \mathbb{R}^{U\times V\times H\times W}$, consisting of sub-aperture images (SAIs) arranged in a $U \times V$ angular grid, each of spatial resolution $H \times W$. The goal is to predict a high-quality depth map $D \in \mathbb{R}^{H\times W}$ for the central view LF_c. More specifically, GaussianNet consists of two pathways: one pathway constructs a cost volume over multi-view features and interprets it as a Gaussian Depth representation, while the other extracts Gaussian Shape features. These two representations are fused via a Gaussian rendering module in an occlusion-aware, spatially consistent manner, producing a refined depth map for the central view. This design enables the network to adptive learn continuous surfaces represetation, effectively bridging discrete LF samples and continuous scene geometry.

Feature Extraction. The feature extraction block consists of 9 stacked 3×3 convolutional blocks with 128 channels, each followed by ReLU activation and residual connections. This module is designed to efficiently extract both local spatial structures and angular-view consistency, which are critical for accurate depth estimation in LF. The final output is a feature tensor F_{lf}.

Cost Volume Construction. We use homography warping (*shift-and-concat*) to align the output $\hat{F}$ for various depths and view combinations (u,v). We use 9 uniformly sampled depth hypotheses within the range [-4, 4], following standard practice. This results in a cost volume that encapsulates matching information across views and depth. The complete cost volume $CV \in \mathbb{R}^{D\times H\times W\times C}$ is generated by stacking the matching costs across multiple depth levels.

Gaussian Initialization. The Gaussian initialization module extracts parametric Gaussian distributions (x,y) from the LF center view. Formally, given the centre view $LF_c \in \mathbb{R}^{H\times W}$, a feature tensor $F_{gauss} \in \mathbb{R}^{H\times W\times C}$ is first extracted via a convolutional encoder ϕ:

$$F_{gauss} = \phi(LF_c). \tag{6}$$

Subsequently, the extracted feature F_{gauss} is sent into Gaussian Encoder to be mapped onto a learnable Gaussian field and outputs the initial parameters for Gaussian splat to each pixel, including Mean Position(μ), Covariance Matrix(Σ), which control the shape of Gaussian field:

$$\mu_{x,y}, \Sigma_{x,y} = \text{Split}(\text{MLP}(\text{Conv}(F_{gauss}))). \tag{7}$$

To ensure positive status of Σ,we apply the sigmoid function for activation:

$$\Sigma_x, \Sigma_y = Sigmoid(\hat{\Sigma}_x, \hat{\Sigma}_y). \tag{8}$$

GaussianMatch Rendering. Given the multiscale cost volume CV, each center-view pixel (x,y) is treated as a single Gaussian. Specifically, the network predicts Gaussian depth $d_{x,y} \in \mathbb{R}$ and confidence weight $w_{x,y} > 0$ for each Gaussian associated with pixel (x,y). Consequently,

$$d_{x,y}, \alpha_{x,y} = \text{Conv}(\text{CV}), \text{MLP}(\text{Conv}(\text{CV})), \tag{9}$$

$$w_{x,y} = \text{softplus}(\alpha_{x,y}), \tag{10}$$

$$d = \{d_{x,y}\}_{x=1,\dots,H;y=1,\dots,W} \in \mathbb{R}^{H\times W},$$

$$w = \{w_{x,y}\}_{x=1,\dots,H;y=1,\dots,W} \in \mathbb{R}^{H\times W}.$$

Table 2: *Quantitative comparison results with SOTA methods on HCInew of BP(0.07), BP(0.03), BP(0.01). The best results are in bold and the second best results are underlined, respectively.*

Method	Backgammon	Dots	Pyramids	Stripes	Boxes	Cotton	Dino	Sideboard	Avg.BP(0.07)
EPINet [SJY*18]	3.287	4.030	0.147	**2.413**	12.248	0.464	1.263	4.783	3.579
EPI-Shift [LSM*19]	22.886	43.918	1.242	22.719	25.951	2.176	5.964	11.795	17.081
FastLFnet [HHX*21]	5.138	21.169	0.620	9.442	18.699	0.714	2.407	7.032	8.153
OAVC [HXWH21]	**3.121**	69.113	0.831	2.903	16.144	2.550	3.936	12.421	13.877
PlaneNet [CSY*24]	5.307	6.497	0.245	3.282	12.543	0.696	1.146	4.359	4.259
DistgDisp [WWW*22]	5.824	1.826	**0.108**	3.913	13.309	0.489	1.414	4.051	3.867
OALFGAN [YZC24]	10.486	2.036	0.303	9.039	19.870	1.942	4.707	7.076	6.932
MRAENet [LYZ*24]	6.445	2.413	0.153	5.714	13.233	0.496	1.402	4.277	4.267
PDE-Net-e [WL24]	5.492	32.357	0.487	6.548	14.970	1.733	1.463	6.010	8.633
GaussianNet (ours)	4.850	**1.614**	0.253	3.730	**10.589**	**0.449**	**1.012**	**2.836**	**3.221**

Method	Backgammon	Dots	Pyramids	Stripes	Boxes	Cotton	Dino	Sideboard	Avg.BP(0.03)
EPINet [SJY*18]	6.289	12.736	0.913	**3.115**	19.759	2.310	3.452	12.080	7.582
EPI-Shift [LSM*19]	40.530	53.184	7.315	47.702	44.146	10.683	22.146	36.639	32.793
FastLFnet [HHX*21]	11.409	41.109	2.193	32.594	37.452	6.785	13.268	21.622	20.804
OAVC [HXWH21]	**5.117**	75.383	9.027	19.877	33.675	20.785	19.028	37.833	27.591
PlaneNet [CSY*24]	7.815	21.292	0.748	7.582	22.525	1.481	3.047	12.567	9.632
DistgDisp [WWW*22]	10.538	4.464	0.539	6.885	21.130	1.478	4.018	9.575	7.328
OALFGAN [YZC24]	17.373	**2.879**	1.825	13.075	30.873	5.307	10.608	14.782	12.090
MRAENet [LYZ*24]	11.910	5.805	0.804	9.595	21.313	1.370	4.048	10.184	8.129
PDE-Net-e [WL24]	12.498	43.470	2.424	15.623	27.265	9.605	5.959	22.523	17.421
GaussianNet (ours)	6.492	3.281	**0.527**	5.219	**19.043**	**1.261**	**2.637**	**8.727**	**5.898**

Method	Backgammon	Dots	Pyramids	Stripes	Boxes	Cotton	Dino	Sideboard	Avg.BP(0.01)
EPINet [SJY*18]	20.899	41.052	11.876	15.674	49.040	28.066	22.401	41.880	28.861
EPI-Shift [LSM*19]	70.581	74.550	40.476	78.945	74.362	46.858	64.162	73.418	65.419
FastLFnet [HHX*21]	39.839	68.146	22.188	63.400	71.820	49.339	56.244	61.963	54.117
OAVC [HXWH21]	49.051	92.332	33.656	28.136	71.906	61.352	61.815	73.849	59.012
PlaneNet [CSY*24]	15.911	50.394	3.074	18.054	47.171	10.713	15.194	37.910	24.803
DistgDisp [WWW*22]	26.168	25.366	4.953	19.254	41.620	7.594	20.460	28.283	21.712
OALFGAN [YZC24]	37.997	25.604	24.163	41.452	62.006	44.421	41.964	42.649	40.032
MRAENet [LYZ*24]	33.161	29.586	6.662	34.030	43.192	8.863	23.918	33.063	26.559
PDE-Net-e [WL24]	37.987	60.688	18.488	40.880	61.509	46.774	33.087	63.695	45.389
GaussianNet (ours)	**15.243**	**18.216**	**2.587**	**14.101**	**40.357**	**6.124**	**14.297**	**25.439**	**17.046**

Using the Gaussian shapes from Gaussian Initialization and the predicted (d, w) pairs, we synthesize the final depth map through weighted summation over all Gaussian components. Each component contributes a depth value at pixel $\mathbf{x}$ according to its predicted d, modulated by its Gaussian field defined by μ and Σ, and scaled by w. This process ensures smoothness and differentiability, and is formulated as Eq. 5 and visualized in Fig. 4.

4. Experiment

4.1. Datasets and Implementation Details

Datasets. To validate the effectiveness of our method, we perform experiments on multiple datasets: The HCI 4D LF Dataset (HCInew) [HJKG17] features a spatial resolution of 512×512 and an angular resolution of 9×9, includes 24 synthetic scenes with highly accurate ground-truth depth maps. The Dense LF Dataset (DLFD) [SJG19] comprises 39 scenes, each annotated with accurate depth labels for every viewpoint. The (New) Standford LF Archieve [RLSW16] is a real-scene dataset captured with the Lytro Illum camera [BJK16]. Our model is trained exclusively on the HCInew dataset, with other datasets reserved solely for testing.

Training. During training, each SAI is randomly cropped into 48×48 patches, converted to grayscale, and subjected to extensive on-the-fly augmentation, including random flips, rotations, brightness and contrast adjustments, noise injection, refocusing, downsampling, and translations [CSY*25b]. GaussianNet is trained in a fully supervised manner using L_1 reconstruction loss and optimized with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [KB14]. We use a batch size of 16 and an initial learning rate of 1×10^{-3}. Training proceeds for 3×10^5 iterations-requiring approximately seven days on a single NVIDIA Tesla V100 GPU. To ensure a fair comparison, we use the same depth range as recent SOTA methods.

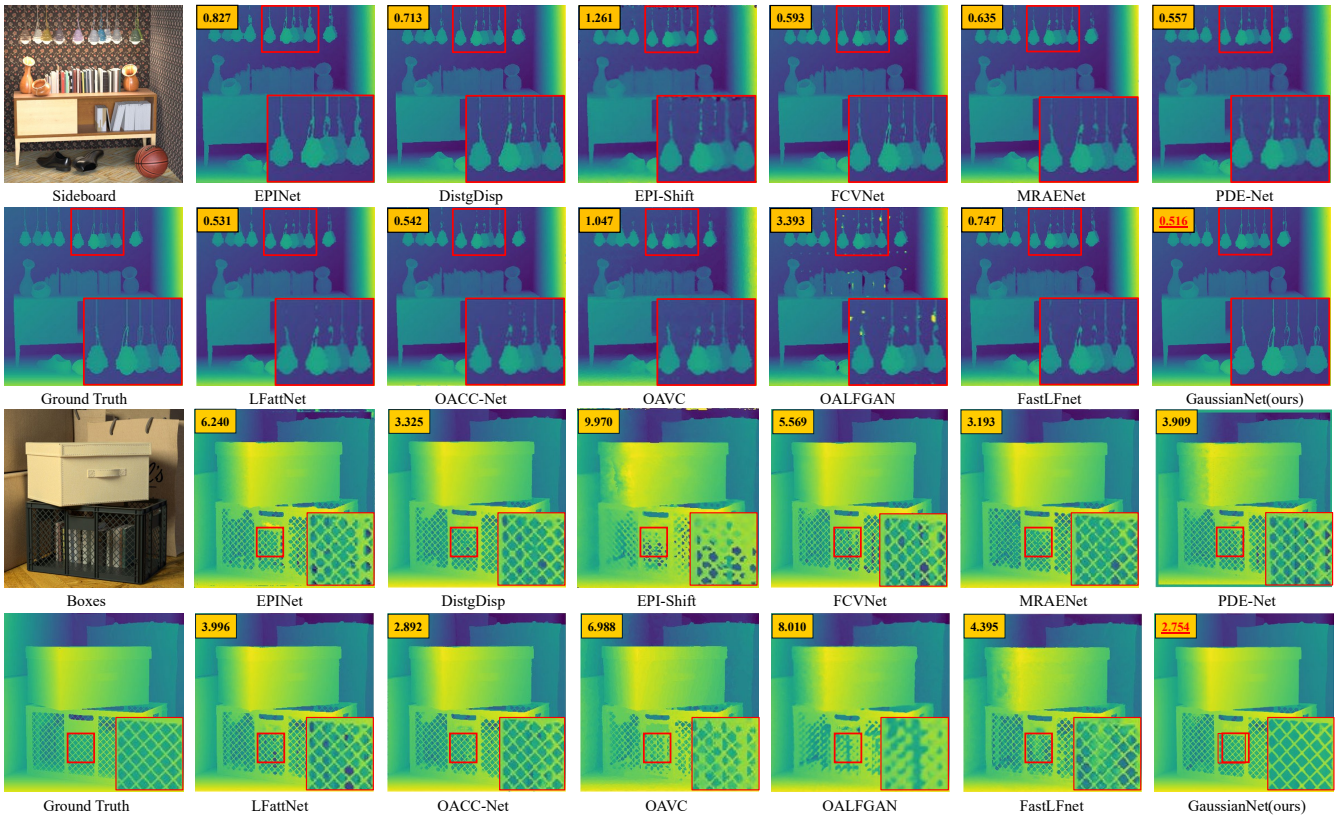


Figure 5: *Qualitative results on HCInew with the corresponding MSE×100 are shown. The best results are highlighted in red. The top-left image is the center view, and the bottom-left image is the ground truth.*

Evaluation. We employ quantitative performance using two complementary metrics: mean squared error (MSE) and bad-pixel ratio (BP). MSE measures the average squared deviation between the predicted depth $Depth$ and the ground-truth depth $Depth^{gt}$, normalized by the number of valid pixels m and scaled by 100 to yield a percentage; BP (ϵ) captures the percentage of pixels whose absolute error exceeds a chosen threshold ϵ, thereby highlighting the prevalence of large outliers. Formally:

$$\mathrm{MSE} = 100 \times \frac{1}{m} \sum_{i=1}^{m} \left(Depth_i - Depth_i^{gt}\right)^2, \tag{11}$$

$$\mathrm{BP}(\epsilon) = 100 \times \frac{1}{m} \sum_{i=1}^{m} \left(|Depth_i - Depth_i^{gt}| > \epsilon\right), \tag{12}$$

where m and i denotes total number of pixels and pixel coordinates.

4.2. Comparison with SOTA Methods

Quantitative Results Table 1 and Table 2 present a comprehensive comparison between GaussianNet and recent SOTA methods on the HCInew dataset. In Table 1, our model achieves an average MSE of 1.172. Table 2 further shows that GaussianNet attains a BadPix(0.07) of 3.221, BadPix(0.03) of 5.898, and BadPix(0.01) of 17.246. Our approach achieves the best performance across all metrics, outperforming existing methods in both average MSE and BadPix. These results demonstrate that GaussianNet not only sets new benchmarks in accuracy but also enables the network to effectively learn and adapt to continuous surfaces, thereby better approximating the depth distribution of real scenes. This significantly enhances the performance of LF depth estimation.

Table 3: *Ablation study on HCInew with key metrics: MSE, BP, and inference time.*

Method	MSE	BP(0.07)	BP(0.03)	Time(s)
Component Removal				
w/o Cost Volume	2.174	5.831	10.724	**6.93**
w/o Gaussian Branch	1.342	3.354	8.430	10.58
Design Alternatives				
Direct Regression	1.409	3.637	8.971	8.60
Full GaussianNet	**1.172**	**3.221**	**5.898**	11.08

Qualitative results Visual results on HCInew and DLFD are shown in Fig. 5 and Fig. 6 along with their corresponding detail regions. It is evident that our results outperform the others. For example, in the 'Sideboard' and 'Boxes', our method excels in preserving intricate details and reconstructing regions with pronounced depth discontinuity. Additionally, in the 'Toy_freinds', our

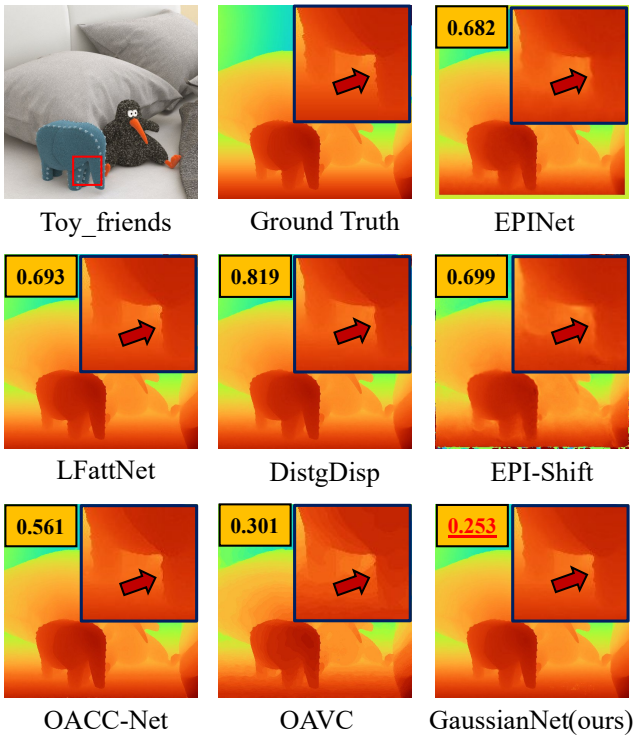


Figure 6: *Visual comparisons on DLFD with the corresponding MSE×100. The best results are highlighted in red. Our results show superior performance compared to other SOTA.*

approach demonstrates superior performance in capturing fine edge details and handling depth transition more effectively.

We also test the performance of our GaussianNet on Stanford archive [RLSW16]. Since ground truth depths are unavailable, we used the model trained on HCInew for inference and compare the visual performance with DistgDisp [WWW*22] , LFattNet [TLOC20], OACC-Net [WWL*22] and EPINet [SJY*18]. Fig. 7 demonstrates our superior performance on 'Lego Truck', our method gets great ability in keep continuous.

4.3. Ablation Study

Compared to a direct regression baseline, which forgoes probabilistic depth modeling and underperforms, the full GaussianNet delivers the best overall trade-off between accuracy and efficiency. These results demonstrate that both cost-volume construction and Gaussian rendering are critical: the former for robust multi-view matching, and the latter for continuous, fully differentiable depth estimation without discrete depth layers.

These results demonstrate that both cost-volume construction and Gaussian rendering are critical: the former for robust multi-view matching, and the latter for continuous, fully differentiable depth estimation. It produces depth maps with sharp object boundaries, smooth slanted planes, and minimal quantization artifacts. The learned variance map systematically offsets disparities that fall between discrete depth hypotheses, effectively improving accuracy.

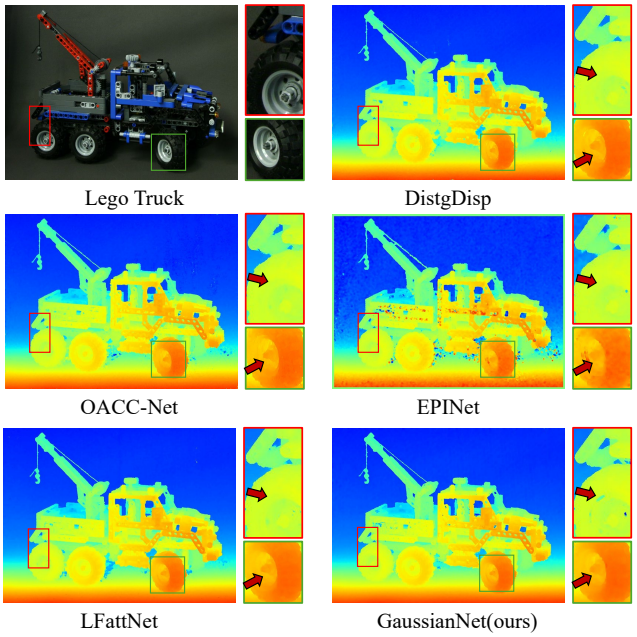


Figure 7: *Visual comparisons on real scene Lego Truck. Our results show superior performance to keep continuous in complex regions.*

5. Conclusion

In this paper, we present GaussianMatch, a probabilistic framework for LF depth estimation. By modeling each pixel's depth as a learnable Gaussian distribution, we adaptively reconstruct continuous surfaces, effectively alleviating the discretization issues. The framework adaptively adjusts each Gaussian's variance to achieve robust fusion in both texture-rich and ambiguous regions, and it naturally shares information among adjacent pixels-significantly reducing depth discontinuities and outliers in textureless or occluded areas. Building on this, we design GaussianNet to regress per-pixel Gaussian parameters and generate the final depth map. Extensive experiments on multiple public LF benchmarks demonstrate that GaussianNet achieves SOTA accuracy, with notable improvements in handling depth discontinuities and occlusions. In future work, we will explore extending GaussianMatch to arbitrary viewpoint light field depth estimation to further enhance the model's generalization capabilities.

6. Acknowledgments

This study is partially supported by the National Key R&D Program of China(No.2022YFC3803600), the National Natural Science Foundation of China(No.62394332, 62372023) , and the Open Fund of the State Key Laboratory of Software Development Environment(No. SKLSDE-2023ZX-11), the Research Start-up Funds of Hangzhou International Innovation Institute of Beihang University under Grant No.2024KQ012, and the Haiyou Plan Fund. This study is supported by High-Performance Computing Center of Hangzhou International Innovation Institute, Beihang University. Thank you for the support from HAWKEYE Group.

References

[BJK16] BOK Y., JEON H.-G., KWEON I. S.: Geometric calibration of micro-lens-based light field cameras using line features. *IEEE transactions on pattern analysis and machine intelligence 39*, 2 (2016), 287–300. 6

[BRR11] BLEYER M., RHEMANN C., ROTHER C.: Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc* (2011), vol. 11, pp. 1–11. 2

[CSY*23a] CHEN R., SHENG H., YANG D., WANG S., CUI Z., CONG R.: Take your model further: A general post-refinement network for light field disparity estimation via badpix correction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2023), vol. 37, pp. 331–339. 3

[CSY*23b] 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 Transactions on Multimedia* (2023). 1, 2

[CSY*24] CHEN R., SHENG H., YANG D., CUI Z., CONG R.: Surface-continuous scene representation for light field depth estimation via planarity prior. *IEEE Transactions on Circuits and Systems for Video Technology* (2024). 1, 6

[CSY*25a] CHEN R., SHENG H., YANG D., CONG R., CUI Z., WANG S., WANG T., ZHAO M.: Towards depth-continuous scene representation with displacement field for robust light field depth estimation. *IEEE Transactions on Multimedia* (2025). 3

[CSY*25b] CHEN R., SHENG H., YANG D., WANG S., CUI Z., CONG R.: Pixel-wise matching cost function for robust light field depth estimation. *Expert Systems with Applications 262* (2025), 125560. 6

[CSZ*24] CONG R., SHENG H., ZHAO M., YANG D., WANG T., CHEN R., SHEN J.: Multimodal perception integrating point cloud and light field for ship autonomous driving. *IEEE Transactions on Intelligent Transportation Systems 25*, 9 (2024), 12477–12489. 1

[CWW*23] CHAO W., WANG X., WANG Y., WANG G., DUAN F.: Learning sub-pixel disparity distribution for light field depth estimation. *IEEE Transactions on Computational Imaging 9* (2023), 1126–1138. 2

[CYT23] CHEN R., YANG Y., TONG C.: G2ifu: Graph-based implicit function for single-view 3d reconstruction. *Engineering Applications of Artificial Intelligence 124* (2023), 106493. 1

[CZL21] CHEN J., ZHANG S., LIN Y.: Attention-based multi-level fusion network for light field depth estimation. In *Proceedings of the AAAI conference on artificial intelligence* (2021), vol. 35, pp. 1009–1017. 1

[HHX*21] HUANG Z., HU X., XUE Z., XU W., YUE T.: Fast light-field disparity estimation with multi-disparity-scale cost aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 6320–6329. 1, 5, 6

[HJKG17] HONAUER K., JOHANNSEN O., KONDERMANN D., GOLDLUECKE B.: A dataset and evaluation methodology for depth estimation on 4d light fields. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13* (2017), Springer, pp. 19–34. 6

[HP16] HEBER S., POCK T.: Convolutional networks for shape from light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 3746–3754. 1

[HXC*25] HU J., XIA B., CHEN B., YANG W., ZHANG L.: Gaussiansr: High fidelity 2d gaussian splatting for arbitrary-scale image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2025), vol. 39, pp. 3554–3562. 3

[HXWH21] HAN K., XIANG W., WANG E., HUANG T.: A novel occlusion-aware vote cost for light field depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). 5, 6

[HYC*24] HUANG B., YU Z., CHEN A., GEIGER A., GAO S.: 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers* (2024), pp. 1–11. 3

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 6

[KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph. 42*, 4 (2023), 139–1. 3

[LH96] LEVOY M., HANRAHAN P.: Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques* (1996), pp. 31–42. 1

[LKG*23] LIU C., KUMAR S., GU S., TIMOFTE R., VAN GOOL L.: Single image depth prediction made better: A multivariate gaussian take. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 17346–17356. 3

[LSM*19] LEISTNER T., SCHILLING H., MACKOWIAK R., GUMHOLD S., ROTHER C.: Learning to think outside the box: Wide-baseline light field depth estimation with epi-shift. In *2019 international conference on 3D vision (3DV)* (2019), IEEE, pp. 249–257. 2, 5, 6

[LYZ*24] LI J., YANG W., ZHANG C., LI H., LI X., WANG L., WANG Y., WANG X.: High precision light field image depth estimation via multi-region attention enhanced network. *IET Computer Vision 18*, 8 (2024), 1390–1406. 5, 6

[PZZJ21] PIAO Y., ZHANG Y., ZHANG M., JI X.: Dynamic fusion network for light field depth estimation. *arXiv preprint arXiv:2104.05969* (2021). 3

[RLSW16] RAJ A. S., LOWNEY M., SHAH R., WETZSTEIN G.: Stanford lytro light field archive. http://lightfields.stanford.edu/LF2016.html, 2016. 6, 8

[SCY*22] SHENG H., CONG R., YANG D., CHEN R., WANG S., CUI Z.: Urbanlf: A comprehensive light field dataset for semantic segmentation of urban scenes. *IEEE Transactions on Circuits and Systems for Video Technology* (2022). 1

[SJG19] SHI J., JIANG X., GUILLEMOT C.: A framework for learning depth from a flexible subset of dense and sparse light field views. *IEEE Transactions on Image Processing 28*, 12 (2019), 5867–5880. 6

[SJY*18] SHIN C., JEON H.-G., YOON Y., KWEON I. S., KIM S. J.: Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 4748–4757. 1, 2, 5, 6, 8

[SRV24] SZYMANOWICZ S., RUPPRECHT C., VEDALDI A.: Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2024), pp. 10208–10217. 3

[SSC22] SUN C., SUN M., CHEN H.-T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 5459–5469. 3

[TLOC20] TSAI Y.-J., LIU Y.-L., OUHYOUNG M., CHUANG Y.-Y.: Attention-based view selection networks for light-field disparity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 12095–12103. 1, 2, 5, 8

[WL24] WANG Q., LI Y.: Pde-net: Pyramid depth estimation network for light fields. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing* (2024), pp. 670–676. 5, 6

[WSC*24a] WANG T., SHENG H., CHEN R., CONG R., ZHAO M., CUI Z.: Adaptive epi-matching cost for light field disparity estimation. *IEEE Transactions on Instrumentation and Measurement* (2024). 3

[WSC*24b] WANG T., SHENG H., CHEN R., YANG D., CUI Z., WANG S., CONG R., ZHAO M.: Light field depth estimation: A comprehensive survey from principles to future. *High-Confidence Computing 4*, 1 (2024), 100187. 2

[WTZ23] WANG X., TAO C., ZHENG Z.: Occlusion-aware light field depth estimation with view attention. *Optics and Lasers in Engineering 160* (2023), 107299. 5

[WWL*22] WANG Y., WANG L., LIANG Z., YANG J., AN W., GUO Y.: Occlusion-aware cost constructor for light field depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 19809–19818. 2, 5, 8

[WWW*22] WANG Y., WANG L., WU G., YANG J., AN W., YU J., GUO Y.: Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*, 1 (2022), 425–443. 2, 5, 6, 8

[WZF*25] WU G., ZHOU Y., FANG L., LIU Y., CHAI T.: Geo-ni: Geometry-aware neural interpolation for light field rendering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025), 1–1. 1

[YCH*24] YU Z., CHEN A., HUANG B., SATTLER T., GEIGER A.: Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2024), pp. 19447–19456. 3

[YCS*23] YANG D., CUI Z., SHENG H., CHEN R., CONG R., WANG S., XIONG Z.: An occlusion and noise-aware stereo framework based on light field imaging for robust disparity estimation. *IEEE Transactions on Computers* (2023). 3

[YZC24] YAN W., ZHANG X., CHEN H.: Occlusion-aware unsupervised light field depth estimation based on multi-scale gans. *IEEE Transactions on Circuits and Systems for Video Technology 34*, 7 (2024), 6318–6333. 1, 5, 6

[ZLL*24] ZHENG X., LI Z., LIU D., ZHOU X., SHAN C.: Spatial attention-guided light field salient object detection network with implicit neural representation. *IEEE Transactions on Circuits and Systems for Video Technology* (2024). 1

[ZRSK20] ZHANG K., RIEGLER G., SNAVELY N., KOLTUN V.: Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020). 3

[ZSL*23] ZHOU P., SHI L., LIU X., JIN J., ZHANG Y., HOU J.: Light field depth estimation via stitched epipolar plane images. *IEEE Transactions on Visualization and Computer Graphics 30*, 10 (2023), 6866–6879. 2

[ZZY*19] ZHOU W., ZHOU E., YAN Y., LIN L., LUMSDAINE A.: Learning depth cues from focal stack for light field depth estimation. In *2019 IEEE International Conference on Image Processing (ICIP)* (2019), IEEE, pp. 1074–1078. 3